ARTICLE OPEN

Multi-omic characterization of genome-wide abnormal DNA methylation reveals diagnostic and prognostic markers for esophageal squamous-cell carcinoma

Yiyi Xi[1], Yuan Lin[2], Wenjia Guo[3], Xinyu Wang[4], Hengqiang Zhao[4], Chuanwang Miao[1], Weiling Liu[1], Yachen Liu[1], Tianyuan Liu[1], Yingying Luo[1], Wenyi Fan[1], Ai Lin[1], Yamei Chen[1], Yanxia Sun[1], Yulin Ma[1], Xiangjie Niu[1], Ce Zhong[1], Wen Tan[1], Meng Zhou[4], Jianzhong Su[4,5]✉, Chen Wu[1,6,7]✉ and Dongxin Lin[1,6,8]✉

This study investigates aberrant DNA methylations as potential diagnosis and prognosis markers for esophageal squamous-cell carcinoma (ESCC), which if diagnosed at advanced stages has <30% five-year survival rate. Comparing genome-wide methylation sites of 91 ESCC and matched adjacent normal tissues, we identified 35,577 differentially methylated CpG sites (DMCs) and characterized their distribution patterns. Integrating whole-genome DNA and RNA-sequencing data of the same samples, we found multiple dysregulated transcription factors and ESCC-specific genomic correlates of identified DMCs. Using featured DMCs, we developed a 12-marker diagnostic panel with high accuracy in our dataset and the TCGA ESCC dataset, and a 4-marker prognostic panel distinguishing high-risk patients. In-vitro experiments validated the functions of 4 marker host genes. Together these results provide additional evidence for the important roles of aberrant DNA methylations in ESCC development and progression. Our DMC-based diagnostic and prognostic panels have potential values for clinical care of ESCC, laying foundations for developing targeted methylation assays for future non-invasive cancer detection methods.

Signal Transduction and Targeted Therapy (2022)7:53 ; https://doi.org/10.1038/s41392-022-00873-8

INTRODUCTION

Esophageal squamous-cell carcinoma (ESCC) accounts for 80% of esophageal cancer cases worldwide[1] and has a 5-year survival rate of <30%.[2,3] About 350,000 people die of ESCC every year in China where this malignancy largely occurs.[4] Treating this disease at early stages generally results in better prognosis than at late stages, but effective biomarkers that aid early detection and/or accurate prognosis prediction are currently lacking. Aberrant DNA methylation[5,6] plays an important role in cancer initiation and progression[7–9] and have been investigated to derive diagnostic/prognostic biomarkers for several types of human cancer including ESCC.[10–13] For cancer detection, differentially methylated CpG sites (DMCs) are considered better than other genetic features due to their tissue-of-origin and cancer-type specificity, early emergence during carcinogenesis and relative stability in fixed samples and body fluid over time.[14–17] A recent clinical study has demonstrated the superiority of DMC markers when working with circulating cell-free tumor DNAs (cfDNAs).[18] In pursuit of potential DMC-based markers, it is crucial to conduct unbiased genome-wide screening in a large number of samples. Equally important is subsequent association testing with the same patient's other relevant genomic or transcriptomic features particularly the gene expression profile. Epigenetic anomalies often disturb gene regulation; systematically investigating the interactions between these two omics layers would help pinpoint biologically sound DMC markers. However, few previous studies have adequately fulfilled these two steps. Early works usually focused on a small number of aberrantly methylated genes instead of performing genome-wide search.[19,20] More recent studies either ignored the genomic and transcriptomic contexts or investigated them in a different set of patients, likely due to a lack of matched multi-omics data. For example, Wang et al. interrogated the methylome of 84 TCGA ESCC patients and developed diagnostic models, but gene expression data were not consulted.[21,22] Talukdar et al. developed a diagnostic 7-CpG panel based on methylation profiling of more than 100 ESCC samples collected from Africa, Asia and South America countries, but they weighted each CpG marker based on gene expression from TCGA ESCC patients (mostly Caucasian).[13] Chen et al. integrated DNA methylation and gene expression profiles from the same samples, but the sample size was too small ($n = 4$).[23] Furthermore, although DMC sites in cancer genomes bear ethnic specificity,[24] Caucasian

[1]Department of Etiology and Carcinogenesis, National Cancer Center/National Clinical Research Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China; [2]Beijing Advanced Innovation Center for Genomics, Biomedical Pioneering Innovation Center, Peking University, Beijing 100871, China; [3]Cancer Institute, Affiliated Cancer Hospital of Xinjiang Medical University, Urumqi 830000, China; [4]School of Biomedical Engineering, School of Ophthalmology and Optometry and Eye Hospital, Wenzhou Medical University, Wenzhou 325011, China; [5]Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325000, China; [6]Jiangsu Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China; [7]CAMS Key Laboratory of Genetics and Genomic Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China and [8]Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou 510060, China
Correspondence: Jianzhong Su (sujz@wmu.edu.cn) or Chen Wu (chenwu@cicams.ac.cn) or Dongxin Lin (lindx@cicams.ac.cn)
These authors contributed equally: Yiyi Xi, Yuan Lin, Wenjia Guo.

Received: 3 May 2021 Revised: 23 November 2021 Accepted: 30 December 2021
Published online: 25 February 2022

samples are by far the mostly used for biomarker discovery, with only a couple of exceptions.[13,23] No large-scale methylome interrogation has been carried out for Chinese ESCC patients.

In our previous study,[25] we have performed whole-genome sequencing and RNA-sequencing on 91 Chinese ESCC patients' matched tumor and adjacent normal tissue samples. In the present study, we continued to profile genome-wide DNA methylation of the same samples and correlate DMCs with a variety of gene expression alterations as well as somatic and germline variants specific to ESCC. Based on the results, we have developed a diagnostic model comprised of 12 promoter/gene-body DNA methylation CpG sites that robustly distinguishes ESCC from adjacent tissues or normal esophagus in multiple patient sets. We have also developed a prognostic model comprised of 4 promoter/gene-body CpG sites that can classify ESCC patients into high-risk and low-risk groups. The host genes of identified markers have potentially functional roles in ESCC development and progression, as implicated in the literature or by our in-vitro experiments. Overall, this study demonstrates that the ESCC genome abounds with specific DNA methylation patterns that could be effective diagnostic or prognostic biomarkers and potential mediators of tumor development and/or progression.

RESULTS

Overview of differentially methylated CpG sites in ESCC

Among 429,717 probes (out of 467,079, 92%) that had passed quality control, 35,577 (8.28%) were differentially methylated between tumor and adjacent normal samples (FDR $q < 0.05$, absolute median methylation difference $|MMD| > 0.20$; Fig. 1a), with 56.54% (20,114/35,577) of these DMCs hypo-methylated (Fig. 1b). The distribution of DMCs varied among chromosomes (Supplementary Fig. S1a, b), mostly enriched in Chromosome 8 (odds ratio (OR) = 1.32, $P = 1.00e\text{-}89$) and mostly absent in Chromosome 22 (OR = 0.67, $P = 1.00e\text{-}45$). Hyper-methylated sites were mostly enriched in Chromosomes 18 and 19 (OR = 1.32, $P = 3.60e\text{-}6$, OR = 1.11, $P = 4.25e\text{-}4$) whereas hypo-methylated sites in Chromosome 8 (OR = 1.60, $P = 2.24e\text{-}68$), respectively (Fig. 1c, d). Furthermore, DMCs were significantly enriched in intergenic and enhancer regions (Supplementary Fig. S1c, d), with more hypo- than hyper-methylated CpG sites in intergenic regions and similarly abundant hyper- and hypo-methylated sites in enhancer regions (Fig. 1e, f). Hyper-methylated CpG sites were also enriched in CpG islands (OR = 1.66, $P = 1.00e\text{-}1502$) and DNase I hypersensitivity sites (OR = 1.77, $P = 6.06e\text{-}258$), while hypo-methylated sites also enriched in open sea (OR = 1.89, $P = 1.00e\text{-}4373$) (Fig. 1e, f). We found more hyper- than hypo-methylated sites within promoter regions (Fig. 1e, f). At the chromosome level, Chromosome 8 was enriched with hypo-methylated sites mainly found in open sea, intergenic and enhancer regions; Chromosomes 18 and 19 were enriched with hyper-methylated sites mainly found in CpG islands, promoter and DNase I hypersensitivity sites (Supplementary Fig. S1e–j).

Commercial DNA methylation arrays are intentionally focused on DNA methylation CpGs at promoters and gene bodies, which often regulate the expression of host genes in a cis manner. Even so, the genome of our ESCC samples still contained more-than-expected hyper-methylation in promoter and adjacent regions, while hypo-methylation dominates genome wide, as observed in several other cancer types.[7,9,17] The methylation status of 3241 (9.11%) promoter or gene-body DMCs were significantly correlated with the expression levels of their host genes in ESCC, quantified using RNA sequencing data (|Spearman's correlation coefficient $r| > 0.30$, FDR $q < 0.05$). These DMCs were mostly overrepresented in Chromosome 7 (OR = 1.55, $P = 1.80e\text{-}14$; Fig. 1g, h) and were more likely to reside at gene-bodies than at promoter and adjacent regions (OR = 1.70, $P = 6.22e\text{-}197$; Fig. 1i, j). Because DMCs at promoters and gene bodies often affect host gene expression differently,[26] we investigated promoter-DMC involved negative expression-methylation correlations and gene-body-DMC involved positive expression-methylation correlations, using same patients' gene expression profiles obtained in our previous study[25] (Fig. 2a, b). Among protein coding genes differentially expressed between tumor and adjacent normal samples, 90 downregulated and 44 upregulated genes were associated with 224 hyper-methylated and 70 hypo-methylated CpG sites at their promoter regions, respectively; 274 downregulated and 70 upregulated genes were associated with 764 hypo-methylated and 221 hyper-methylated CpG sites in their gene-bodies, respectively (Fig. 2a, b). We then looked into the methylation-expression correlation in Chromosomes 8, 18, 19. Chromosomes 8 mainly contained genes whose expression levels were associated with hypo-methylated sites in gene bodies (Supplementary Fig. S2a, b); Chromosome 19 mainly contained genes whose expression levels were associated with promoter hyper-methylated sites (Supplementary Fig. S2c, d). No preference was observed in Chromosome 18.

Host genes potentially dysregulated by negatively correlated promoter-DMCs were enriched in the GO categories of metal ion binding, transcription factor activity and transcription regulation, whereas host genes potentially dysregulated by positively correlated gene-body DMCs were enriched in the GO categories of system development and cell part morphogenesis (Supplementary Fig. S3). In light of this result, we examined the overlap between these DMC-associated genes with known human transcription factors (TFs)[27] and found significant TF enrichment in hyper-methylation associated genes, including 32 of 90 downregulated genes ($P = 4.12e\text{-}15$) associated with promoter hyper-methylation and 39 of 70 upregulated genes ($P = 4.83e\text{-}27$) associated with gene-body hyper-methylation (Fig. 2b). The former group (32 downregulated TFs) were mostly the members of the zinc finger gene family, such as *ZNF382*[28] (Fig. 2c, f), whereas the latter group (39 upregulated TFs) included 29 potential oncogenic Homeobox genes such as *HOXB13* and *DLX1*[29] (Fig. 2d, e, g, h), suggesting genome-wide DNA methylation anomalies may have led to the dysregulation of multiple TFs involved in a variety of molecular processes contributing to ESCC initiation and progression. Twenty-three of the 32 downregulated TFs (71.88%) locate in Chromosome 19, accounting for 85.19% (23/27) of all the protein coding genes in that chromosome that were associated with hyper-methylated promoter CpGs and downregulated in tumor samples (Supplementary Fig. S2d).

Differentially methylated CpG sites are associated with ESCC-specific genetic variations

We found recurrent promoter or gene-body DMCs (each in >51% (46/91) patients) in 9 previously identified ESCC driver genes *FAT1*, *NOTCH1*, *JUB*, *MLL2*, *PIK3CA*, *TGFBR2*, *NFE2L2*, *NOTCH3* and *ZNF750*[25] (Fig. 3a; Supplementary Data S1). As many as 97% (88/91) of our patients had promoter or gene-body DMCs in these genes. Surprisingly, no DMCs were found in *TP53*. Among DMC-correlated TF genes, some have been implicated in the progression of ESCC, including zinc finger genes such as *ZNF382*, *ZNF582* and *ZNF667*[28,30,31] and Homeobox genes such as *BARX1*, *HOXA13* and *HOXC10*.[32–34] All 7 CpG sites at the promoter of *ZNF382*, a NF-κB inhibitor frequently downregulated in ESCC,[28,35] were recurrently hyper-methylated in our data: at least one of the 7 DMCs were found in 75.82% (69/91) ESCC genomes. We also found hypo-methylation in the promoter of *TP63*, which is part of a core regulatory circuitry for ESCC.[36]

The frequencies of differential methylation events were then correlated with recurrent ESCC genomic variants that we have previously identified.[25] We found that hyper-methylation events were significantly correlated with somatic mutations in the genes *RB1*, *NOTCH1*, *CDKN2A* and *PIK3CA*, 3q26.32, 7p22.3 and 14q13.3 amplifications and 2q22.1 deletions; hypo-methylation events

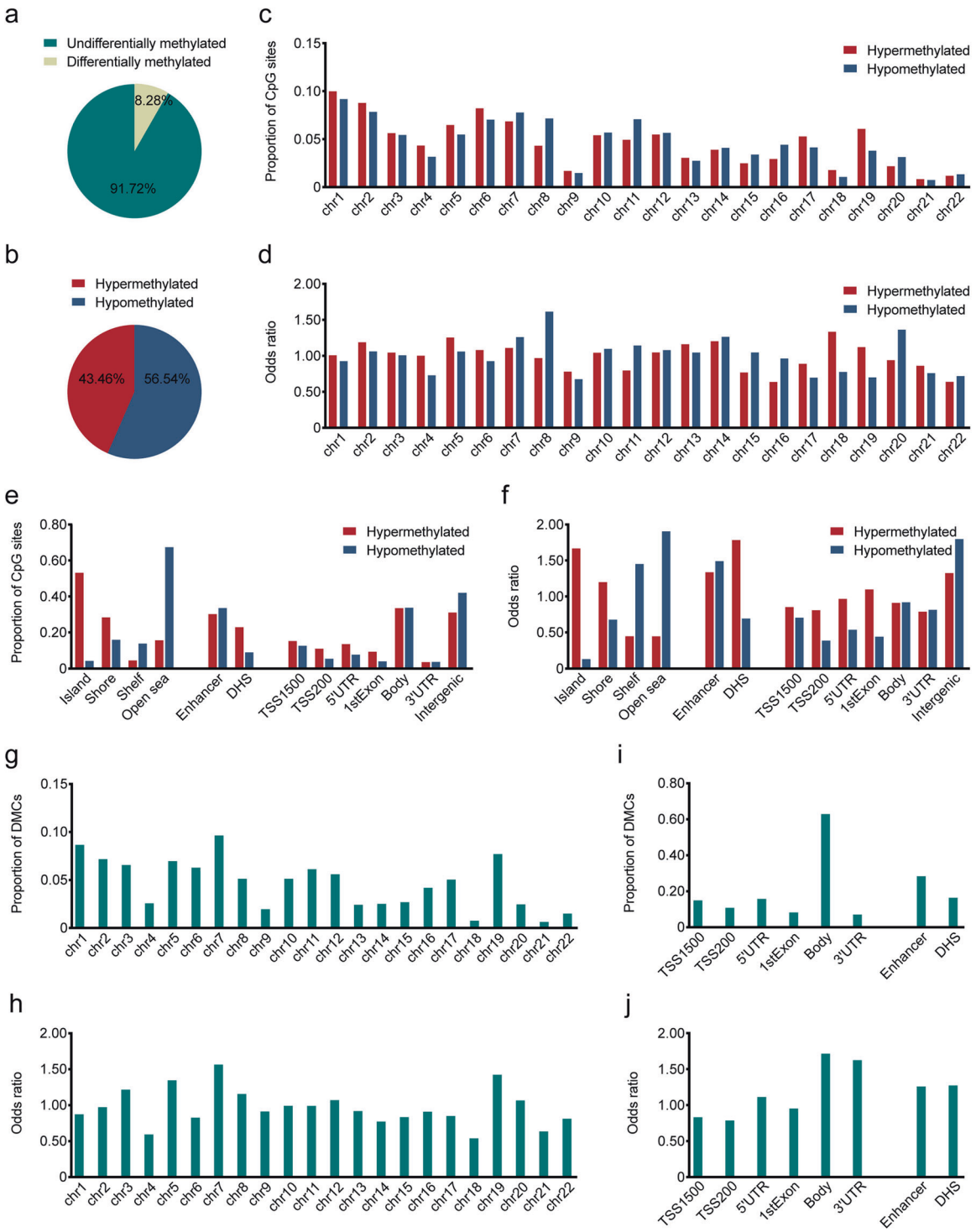
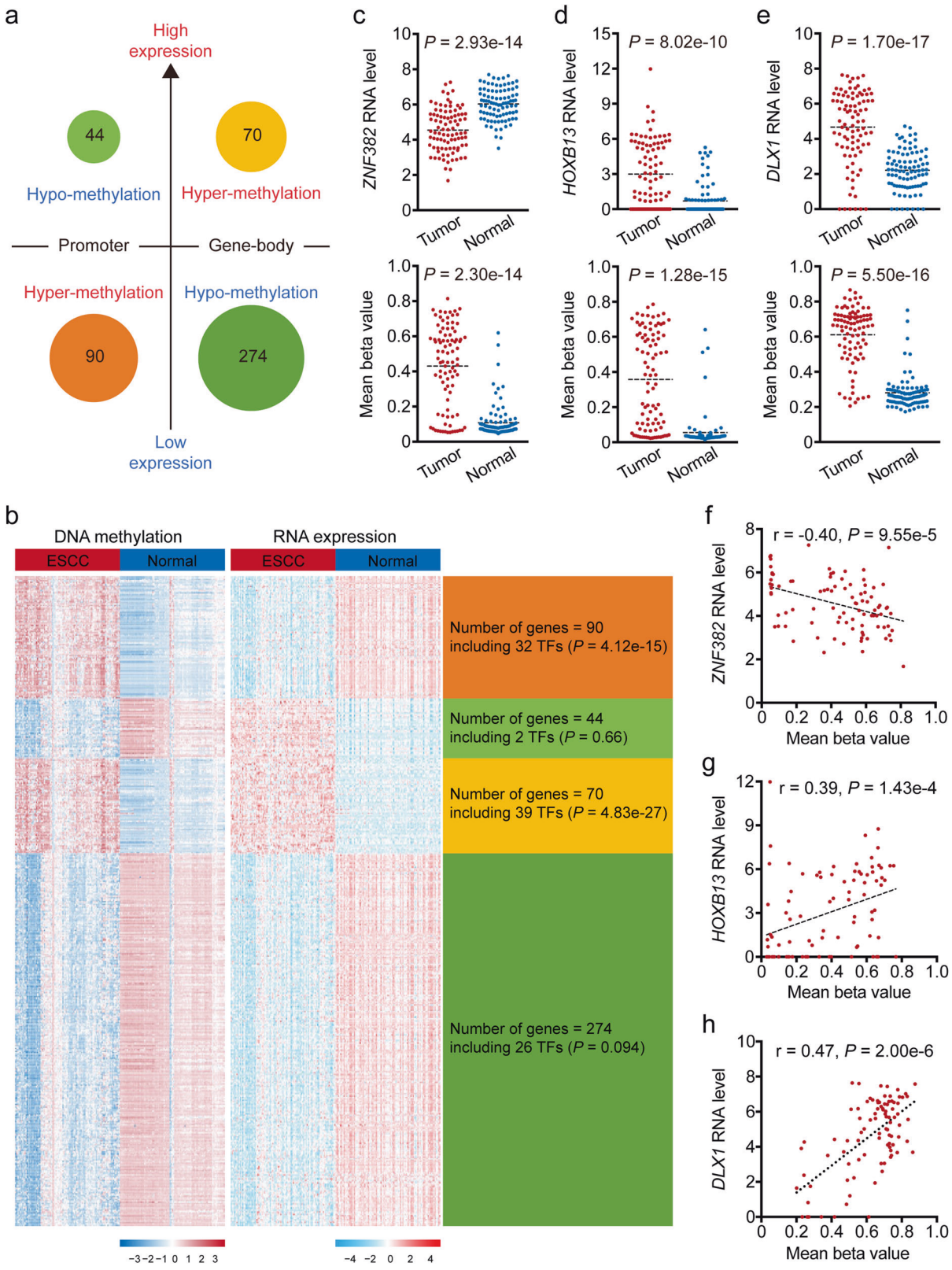


Fig. 1 Characteristics of differentially methylated probes in ESCC. a The proportion of all filtrated CpG sites that are differentially methylated or not methylated. **b** The proportion of differentially hyper-methylated (red) or hypo-methylated (blue) CpG islands. **c**, **d** The proportion (**c**) and odds ratio (**d**) of hyper-methylated (red) or hypo-methylated (blue) CpG sites in different chromosomes. **e**, **f** The category (**e**) and odds ratio (**f**) of genomic locations for hyper-methylated (red) or hypo-methylated (blue) sites. **g**, **h** The proportions (**g**) and odds ratio (**h**) of methylated CpG sites correlated with the expression levels of genes in different chromosomes. **i**, **j** The category (**i**) and odds ratio (**j**) of genomic locations for CpG sites correlated with genes expression. Odds ratio was computed against the general distribution and *P* value was computed by Hypergeometric test. Island, CpG island; shore, 0–2 kb from CpG island; shelf, 2–4 kb from CpG island; open sea, other genomic regions; TSS1500, 200–1500 bases upstream of the transcriptional start site (TSS); TSS200, 0–200 bases upstream of the TSS; 5′UTR, within the 5′ untranslated region and between the TSS and the ATG start site; body, between the ATG and stop codon regardless the presence of introns, exons, TSS or promoters; 3′UTR, between the stop codon and poly A signal



were significantly correlated with somatic mutations in the genes *CREBBP* and *NOTCH3*, 19q13.12 amplifications and 3p14.2 and 13q14.3 deletions (Fig. 3b, c and Supplementary Data S2, 3). *Cis*-meQTL analysis indicated 292 DMCs associated with 4864 nearby SNPs (within a 100-kb window centering each DMC) in ESCC tissues and 2064 DMCs associated with 29,321 SNPs in adjacent normal tissues (Supplementary Data S4, 5 and Supplementary Fig. S4). Compared with adjacent normal, 1974 DMCs lost genetic control in tumor genomes and 202 DMCs gained new correlations. Moreover, ESCC-associated SNPs (14,761, nominal $P < 0.05$)

Fig. 2 Integrative analysis of whole-genome DNA and RNA-sequencing data uncovered methylation-mediated dysregulation of multiple TFs in ESCC. a, **b** The association between promoter or gene-body methylation and host gene expression were identified. There are four clusters: genes ($n = 90$) that are hyper-methylated in promoter with low expression in ESCC; genes ($n = 44$) that are hypo-methylated in promoter with high expression; genes ($n = 70$): that are hyper-methylated in gene-body with high expression; genes ($n = 274$) that are hypo-methylated in gene-body with low expression. Number of known TFs are shown in each cluster. *P* value was computed by Hypergeometric test. **c** Differential expression (top) and promoter methylation (bottom) levels of *ZNF382* in ESCC and normal samples. **d**, **e** Differential expression (top) and gene-body methylation (bottom) levels of *HOXB13* (**d**) and *DLX1* (**e**) in ESCC and normal samples. **f** The correlation between mRNA expression and promoter DNA methylation levels of *ZNF382*. **g**, **h** The correlation between mRNA expression and gene-body DNA methylation levels of *HOXB13* (**g**) and *DLX1* (**h**). *P* of Student's *t* test for gene expression and Wilcoxon signed-rank test for methylation. Genes mRNA expression level (RSEM) was added by 1 and then log2 transformed. Dotted short line indicates mean expression level of each gene

ascertained from our previous GWAS studies[37] were enriched in the identified meQTLs (tumor, $P = 5.66e\text{-}260$; normal, $P = 1.00e\text{-}1240$), which suggests that perturbed DNA methylation may contribute to ESCC predisposition.

Differentially methylated CpG sites are effective diagnostic/prognostic markers for ESCC

Not only were the identified DMCs associated with various molecular characteristics of ESCC, together they could distinguish ESCC from normal esophageal tissues (Supplementary Fig. S5a). We hypothesized that a relatively small number of DMC markers are sufficient for ESCC diagnosis. To identify such markers, we randomly divided the patients into a training set ($n = 60$) and a validation set ($n = 31$) and started with the 1034 promoter/gene-body DMCs whose methylation levels were negatively (or positively) correlated with the expression levels of their host genes. Tumors and adjacent normal samples in the training set were adequately separated by the 1034 DMCs (Supplementary Fig. S5b). Applying random-forest and LASSO to these DMCs generated a model of 12 DMCs (Supplementary Table S1). This model achieved 98.33% sensitivity and 93.33% specificity in the training set (Fig. 4a) and 96.77% sensitivity and 100% specificity in the validation set (Fig. 4b). We also computed receiver operating characteristic (ROC) curves and the area-under-curve values (AUCs) were 99.6% and 97.1% in the training and the validation sets, respectively (Fig. 4d, e). When tested in other ESCC datasets, including TCGA ESCC data and additional GEO datasets (GSE52826 and GSE77991), this diagnostic model consistently showed high sensitivity, specificity and AUCs (Fig. 4c, f and Supplementary Fig. S6a−d), which indicates robustness and generalizability. Unsupervised hierarchical clustering based on these DMCs clearly distinguishes ESCC from normal tissue samples (Fig. 4g−i).

We looked for potential prognosis markers among DMCs based on how strong they were associated with the overall survival (OS) time of ESCC patients in our sample and the TCGA ESCC sample. For each DMC, we constructed a Cox regression model including that DMC as a single predictor and age, sex, smoking status, drinking status and tumor TNM stage as covariates. Four DMCs (cg23378365, cg06090867 and cg03244277 in the promoters of *CYFIP2*, *UBXN10*, *AREG*, respectively, and cg02370667 in the gene-body of *NECAB2*) were significantly associated with patient survival in our sample. We then constructed a prognostic model by summing the methylation levels of these 4 DMCs, each weighted by the hazard ratio (HR) in the corresponding Cox regression result (Supplementary Table S2). This model classified our patients as having high or low prognostic risk (Supplementary Table S3) where the high-risk patients had significantly shorter median OS than others (12 versus 33 months, $P_{\text{log-rank}} = 1.74e\text{-}4$; Fig. 5a), the HR being 3.22 (95% confidence interval (CI) = 1.84−5.62) adjusted for age, sex, smoking status, drinking status and tumor TNM stage. Applying this model to the TCGA ESCC sample yielded a similar result: the predicted high-risk patients had significantly shorter median OS than low-risk patients (23 versus 42 months, $P_{\text{log-rank}} = 0.032$; Fig. 5b), the HR being 4.25 (95% CI = 1.58−11.42) adjusted for age, sex and tumor TNM stage.

We further carried out survival analysis in patients with different tumor stages to evaluate the discriminating ability of our prognostic panel. Within early-stage (I and II) patients in our sample, the low-risk group had longer OS time than the high-risk group (Fig. 5c), although the statistic was marginally significant ($P_{\text{log-rank}} = 0.097$; HR = 1.62, 95% CI 0.34−7.84), probably due to relatively small sample size ($n = 24$). Our model did not perform as well in early-stage patients of the TCGA ESCC cohort (Fig. 5d). For patients with advanced disease (stage III and IV), our model was particularly strong. In our sample, the median OS time in advanced ESCC patients was 12 months for the high-risk group versus 23.5 months for the low-risk group ($P_{\text{log-rank}} = 4.70e\text{-}3$; HR = 2.69, 95% CI 1.48−4.90; Fig. 5e). Similarly, in the TCGA ESCC sample, the median OS time in advanced ESCC patients was 13 and 42 months for high and low-risk groups, respectively ($P_{\text{log-rank}} = 0.018$; HR of 24.11, 95% CI 3.50−166.18; Fig. 5f).

We examined the differential methylation status of the above 16 diagnostic/prognostic markers in the TCGA ESCC data. Ten of them showed similarly significant methylation changes in TCGA ESCC samples compared with normal samples. The rest six markers (cg05446471, cg19310604, cg21041579, cg03244277, cg06090867, cg23378365 in *HDAC11*, *HOXC10*, *SYNE3*, *AREG*, *UBXN10* and *CYFIP2*, respectively) showed similarly significant yet less intensive methylation changes ($P < 0.05$, absolute methylation difference <0.20). We also compared the methylation patterns of the 16 markers across 22 cancer types that have available 450 K array data and normal samples in the TCGA database, with esophageal cancer samples further divided into ESCC and esophageal adenocarcinoma (EAC). The result (Supplementary Fig. S7) confirmed that these markers are ESCC specific.

Functional implications of identified diagnostic and prognostic markers

Some DMC markers we identified locate in the promoters or gene bodies of protein-coding genes and may have contributed to ESCC development or progression by affecting the expression of these genes. To test our hypothesis, we first looked for markers whose methylation levels were correlated with the expression levels of host or nearby genes. Among the 12 DMC markers for ESCC diagnosis, cg10085326, cg24276395, cg05446471, cg21553182 reside at the promoters of *MMP13*, *YEATS2*, *HDAC11* and *ZNF578*, respectively. We classified patients into two groups by the median methylation levels of each site and then compared the expression levels of the corresponding genes. Patients with high methylation of each marker had significantly lower expression levels of these 4 genes in ESCC than patients with low methylation (Supplementary Fig. S8a−d). The methylation status of each marker was negatively correlated with the expression level of the corresponding host gene (all Spearman $r < -0.30$, $P < 0.05$, Supplementary Fig. S9a−d).

The other 8 diagnostic DMCs locate in the gene-body of *AFF3*, *PDE4D*, *SYNE3*, *SLC8A3*, *CPS1*, *HOXC10*, *LDB2* and *PACRG*, respectively. The high methylation status in the 8 sites corresponded to significantly higher expression levels than low methylation status, except for *AFF3* (Supplementary Fig. S8e−l). The expression levels

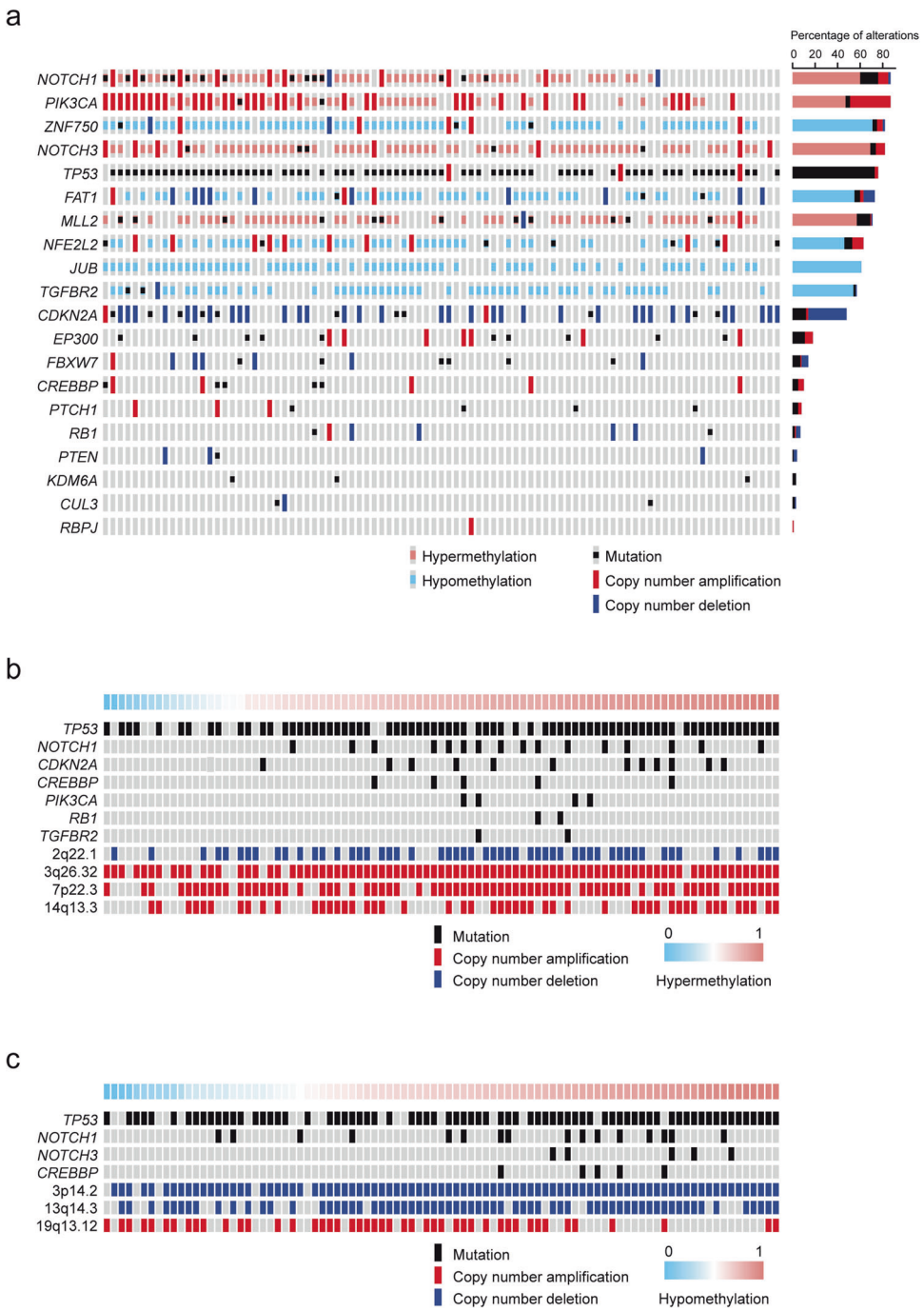


Fig. 3 Hyper- and hypo-methylation events across ESCC and integrated profiling of ESCC driver genes. a Map overview of genetic and epigenetic alterations in 20 ESCC driver genes previously identified. ach column denotes an individual patient and each row represents the status of one gene including somatic mutations (black squares), copy number amplifications (red bars), copy number deletions (blue bars), hyper- (pink bars) and hypo-methylated events (azure bars). Wild-type cases are in gray. Right, percentage of alterations for each gene in 91 ESCC patients while the *X* axis represents total percentage of alterations for each gene. **b**, **c** We tested recurrent genetic alterations in ESCC for their associations with frequency of hyper- (**b**) or hypo-methylated event (**c**). Significant associations (Wilcoxon $P < 0.05$) were shown in above and labeled by gene symbol for somatic mutations or cytoband for amplifications and deletions. Each column denotes an individual patient and each row is one genetic alteration including somatic mutations (black bars), copy number amplifications (red bars) and copy number deletions (blue bars). Wild-type cases are in gray. Top color bars represent the frequency of DNA methylation

of these genes were positively correlated with the methylation levels in their gene bodies (all Spearman $r > 0.30$, $P < 0.05$, Supplementary Fig. S9e−l).

Three of the four DMCs associated with the survival time in ESCC patients, cg23378365, cg06090867, cg03244277 locate at the promoters of *CYFIP2*, *UBXN10* and *AREG*, respectively; cg02370667 resides in the gene-body of *NECAB2*. The expression levels of these genes showed no significant difference between high and low methylation groups of each marker except for *NECAB2* (Supplementary Fig. S8m−p). The expression levels of *NECAB2* and *UBXN10* were significantly correlated with the methylation levels of corresponding CpG sites (*NECAB2*, $r = 0.42$, $P = 3.48e\text{-}5$,

a

Training set	Real ESCC	Real normal	
Predict ESCC	59	4	
Predict normal	1	56	Total
Total	60	60	120
Correct	59	56	115
Sensitivity (%)	98.33		
Specificity (%)		93.33	

b

Validation set	Real ESCC	Real normal	
Predict ESCC	30	0	
Predict normal	1	31	Total
Total	31	31	62
Correct	30	31	61
Sensitivity (%)	96.77		
Specificity (%)		100	

c

TCGA-ESCC set	Real ESCC	Real normal	
Predict ESCC	88	1	
Predict normal	7	15	Total
Total	95	16	111
Correct	88	15	103
Sensitivity (%)	92.63		
Specificity (%)		93.75	

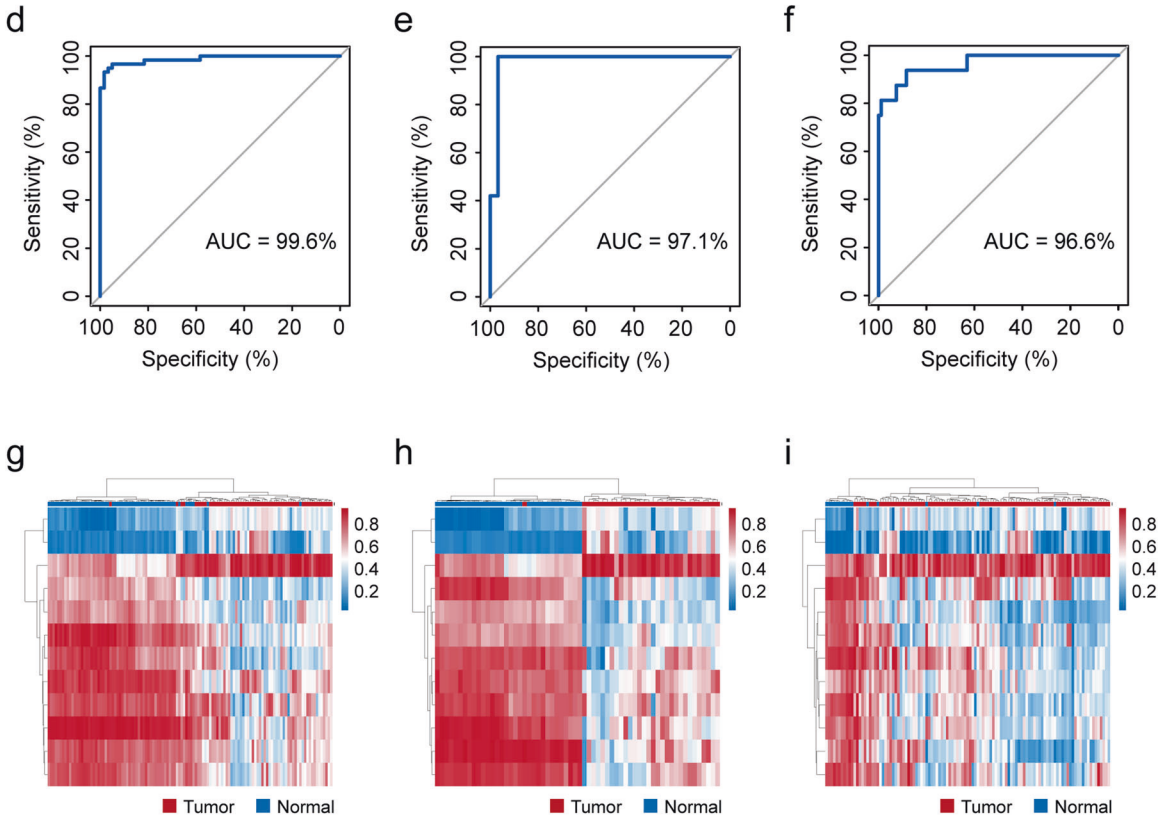


Fig. 4 Diagnosis of ESCC with a DNA methylation panel. a–c The confusion tables of binary results of diagnostic prediction model in the training (**a**), validation (**b**) and TCGA ESCC (**c**) datasets. **d–f** The receiver operating characteristic curve (ROC) of the diagnostic prediction model in the training (**d**), validation (**e**) and TCGA ESCC (**f**) datasets. **g–i** Unsupervised hierarchical clustering and heatmap of 12 methylation markers screened for constructing the diagnostic prediction model in the training (**g**), validation (**h**) and TCGA ESCC (**i**) datasets

Supplementary Fig. S9m; *UBXN10*, $r = 0.22$, $P = 0.033$, Supplementary Fig. S9o).

Next, we focused on marker host genes whose expression levels significantly increased in ESCC samples and associated with the methylation levels of corresponding DMC markers, hypothesizing that knocking down the expression of such genes would diminish the malignancy of ESCC cells in vitro. *MMP13*, *YEATS2* and *HOXC10* with DMC markers in their promoters and *NECAB2* with a gene-body DMC marker were selected for the functional experiments. These four genes were overexpressed in our ESCC samples compared to matched normal tissue samples and the expression levels were strongly correlated with the methylation levels of corresponding DMCs (|Spearman's $r| > 0.30$). We knocked down the expression of these genes in ESCC cell lines, one at a time (Supplementary Fig. S10). Silencing the expression of *YEATS2*, *HOXC10* or *NECAB2* by siRNA significantly inhibited ESCC cell proliferation, migration and invasion; silencing the expression of *MMP13* significantly suppressed ESCC cell migration and invasion but not proliferation (Fig. 6a–h).

DISCUSSION

Despite the promising future of targeted DNA methylation assays in ESCC detection, only recently have we begun to rely on large sample size and genome-wide profiling, and the interactions between DNA methylation and other omic features (e.g., aberrant gene expression and genomic alterations) are too often ignored.[23,38–44] A large-scale, systematic screening of diagnostic and prognostic DMC markers has not been conducted on Chinese samples before, even though China has the highest incidence of ESCC around the world. Here, we explored genome-wide DNA methylation anomalies of 91 Chinese ESCC patients. By comparing their tumor and paired normal samples, we identified 35,577 DMCs and characterized their genome-wide distribution patterns.

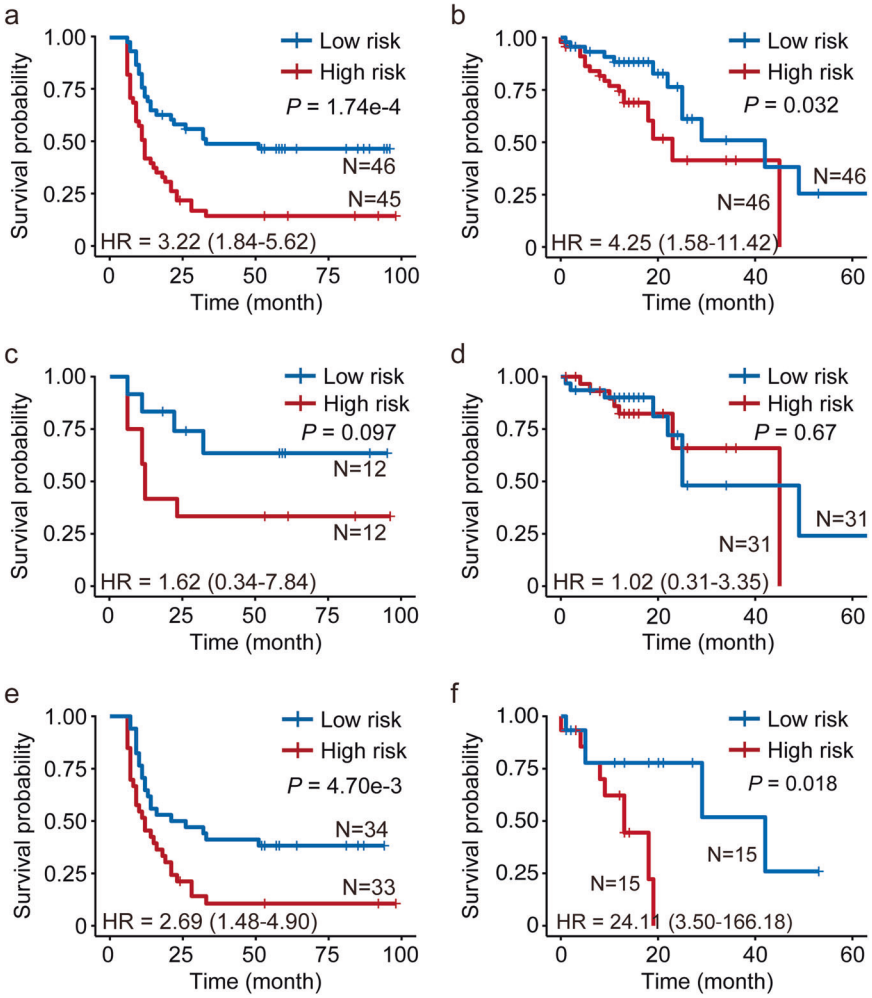


Fig. 5 The correlation of the methylation signature and survival time in patients with ESCC. a, **b** Kaplan–Meier survival curves for all our patient sample (**a**) and all TCGA ESCC patient sample (**b**). **c**, **d** Kaplan–Meier survival curves for patients with early stage ESCC in our sample (**c**) and in TCGA ESCC sample (**d**). **e**, **f** Kaplan–Meier survival curves for patients with advanced stage ESCC in our sample (**e**) and in TCGA ESCC sample (**f**). High- or low-risk group was defined by the weighted hazard ratios of the 4 methylation sites in patients. The *P* value was calculated by log rank test. HR and 95% CI was computed with Cox hazard proportion model

By further integrating genomic and gene expression data of the same samples, we associated many of these DMCs with ESCC-specific genomic or transcriptomic variations, such as somatic mutations in putative ESCC driver genes, germline SNPs associated with ESCC predisposition, as well as aberrant expression of characteristic functional gene sets and pathways.

Notably, the expression of multiple TFs, including several members of the zinc finger family and the homeobox family, were perturbed in ESCC likely due to associated DMCs. Two of these DMCs, associated with *ZNF578* and *HOXC10* respectively, were selected into our diagnostic panel and we validated the functional role of *HOXC10* in vitro. Previous pan-cancer analyses have shown that DNA methylation plays a predominant role in dysregulating TFs in general and the homeobox family in particular.[29,45] Since TFs are "master regulators" of biological processes and pathways critical for the development and differentiation of specific cell types[46] and their defining functions,[47] DNA methylation associated with TF dysregulation could happen at an early or even precancerous stage of ESCC, making them good candidates for early diagnosis markers. In our data, most of the perturbed TFs locate in Chromosome 19, suggesting that Chromosome 19 targeted marker detection could be an alternative to whole-genome screening when the latter is too expensive (e.g., for a very large cohort).

Based on featured DMCs, we developed a panel of 12 methylation CpG sites that can well distinguish ESCC tumor from normal tissues and a linear model of 4 CpG sites that can classify patients into different risk groups in terms of overall survival time. Both models were validated in public ESCC datasets. Recently a 7-CpG diagnostic panel for ESCC has been developed using a large non-TCGA discovery cohort.[13] When applied to TCGA ESCC data, the model showed an AUC of 89%, lower than ours 96.6%. Without looking into the not-yet-released data of that study, we can only speculate on the reasons: (a) we processed all samples into freshly frozen tissues while they used formalin-fixed paraffin-embedded (FFPE) tissues; (b) we only included tumor samples with >75% neoplastic cells, while their filtering threshold was 50%; (c) we combined random forest and the least absolute shrinkage and selection operator (LASSO)-penalized logistic regression for marker screening, while they performed partial least square-discriminant analysis (PLS-DA); (d) we integrated gene expression data of the same patients while they used those of TCGA ESCC patients. A 9-CpG panel has been previously developed for ESCC prognosis,[38] but was based on Illumina's GoldenGate methylation array with only 1505 CpG sites. Only one marker in that panel showed correlation with its host gene.

Following the cis-effect hypothesis of DMCs, we investigated the host genes of the DMC markers in our models. All 12 diagnostic-

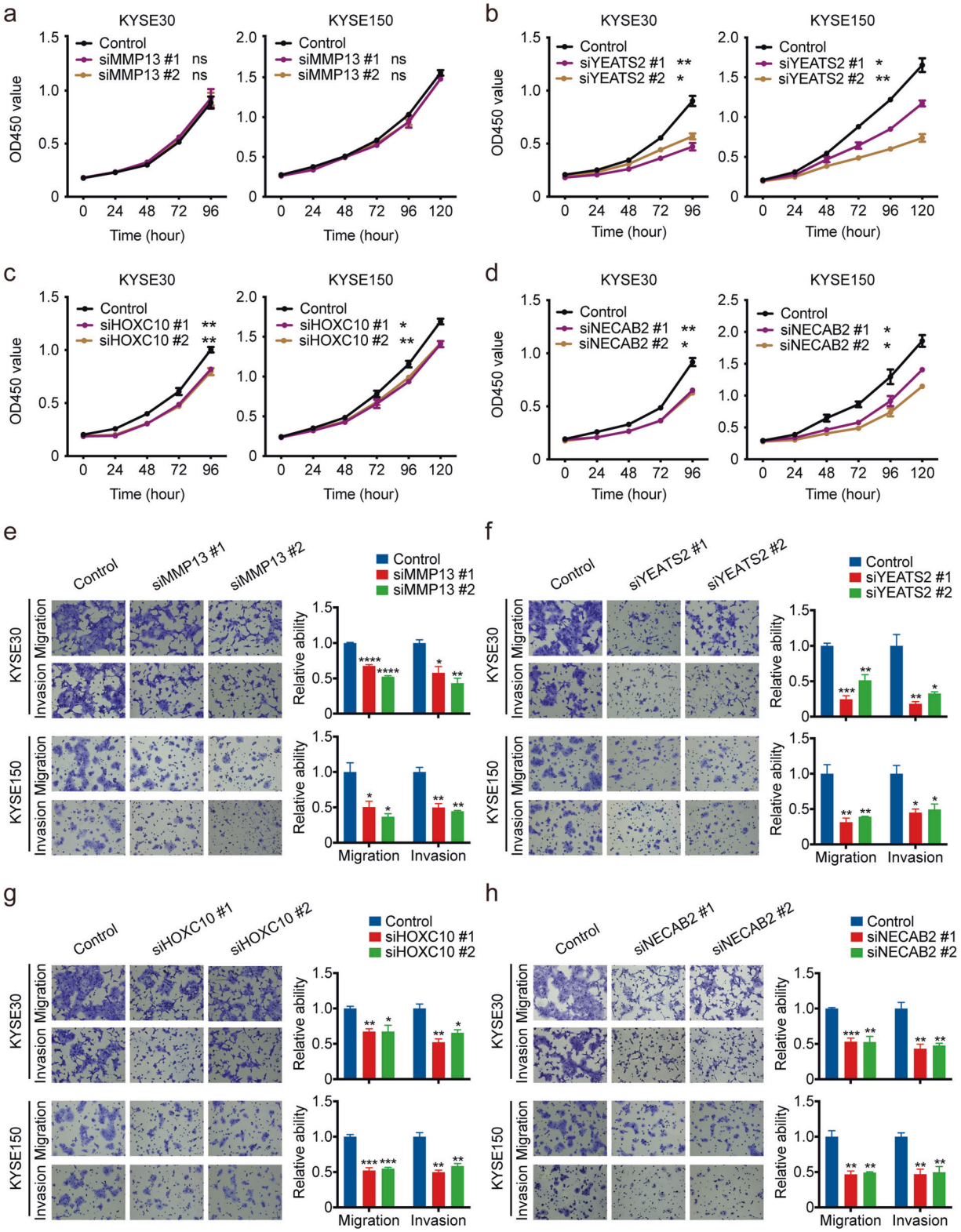


Fig. 6 Effects of silencing some genes in diagnostic and prognostic panels on ESCC cell phenotypes. a–d Silencing the expression of *MMP13* (**a**), *YEATS2* (**b**), *HOXC10* (**c**) and *NECAB2* (**d**) significantly suppressed KYSE30 and KYSE150 cell proliferation. **e–h** Silencing the expression of *MMP13* (**e**), *YEATS2* (**f**), *HOXC10* (**g**) and *NECAB2* (**h**) significantly suppressed KYSE30 and KYSE150 cell migration and invasion. Left panel shows representative cell migration and invasion images and right panel shows quantification statistics. Data represent mean ± SEM from 3 independent experiments. *$P < 0.05$; **$P < 0.01$; ***$P <$ 1.00e-3; ****$P <$ 1.00e-4 and ns not significant of Student's *t* test compared with corresponding control

marker host genes were dysregulated in ESCC samples compared with matched adjacent normal samples. Among them, *MMP13* encodes a member of the matrix metalloproteinase family that can be dysregulated in esophageal cancer.[48,49] *YEATS2* encodes a scaffolding subunit of the Ada-two-A-containing (ATAC) complex implicated in lung tumorigenesis.[50] *HDAC11* encodes a class IV histone deacetylase involved in multiple types of cancer.[51–53] *AFF3* is a potential tumor suppressor in lung cancer.[54] *PDE4D* and *CPS1* were dysregulated in a variety of cancer types.[55–60] *HOXC10* has been associated with many cancer types including ESCC.[34,61–64] The roles of *SYNE3, SLC8A3*, *LDB2* and *PACRG* in ESCC development remain unknown and thus warrant further investigations. Three out of four prognostic-marker host genes were dysregulated in our ESCC samples. *AREG* encodes the epidermal growth factor receptor (EGFR) ligand amphiregulin, an often-upregulated prognostic marker in several types of human cancer.[65–69] The methylation of *AREG* has been associated with survival in patients with astrocytoma.[69] The protein product of *CYFIP2* is involved in p53-dependent apoptosis induction.[70] Decreased expression of *CYFIP2* can promote cancer cell growth in vitro[71] and has been reported in gastric cancer.[72] The roles of *UBXN10* and *NECAB2* in ESCC progression remain to be elucidated. We demonstrated in vitro that *MMP13*, *YEATS2*, *HOXC10* and previously unreported *NECAB2* could contribute to ESCC progression when upregulated.

We validated the 16 DMC markers in the TCGA 450 K microarray data. Although they all had similarly significant methylation changes, 6 sites, including one at the functionally validated *MMP13*, changed so little that they would not have been discovered if we had screened the TCGA data. We also examined previously reported DNA methylation markers and their host genes[13,21–23,38,39] in our samples. Of 30 differentially methylated CpG regions/sites that involve 25 genes, six regions (in *PAX9*, *THSD4*, *TWIST1*, *EPB41L3*, *GPX3* and *COL14A1*, respectively) and 4 CpG sites (cg20655070, cg27062795 in *ZNF542* and cg04550052, cg04698114 in *SALL1*) had similarly significant methylation changes. The methylation status of one region (in *CDH5*) and 8 CpG sites was undetermined in our samples, as the 450 K microarray we used does not include corresponding probes. No significant methylation difference was detected regarding three regions (in *SIM2*, *MLH1* and *CDX1*, respectively) and 7 CpG sites (cg15830431, cg19396867, cg26671652, cg20295442, cg20912169, cg22383888, cg12973591 in *STK3*, *ZNF418*, *ADHFE1*, *EOMES* and *TFPI2*, respectively). These discrepancies may reflect ethnic divergency.

Though equipped with multi-omic data, we decided not to identify markers from other omics layers and then integrate them into current DMC-only diagnostic/prognostic panels. On the one hand, it might help explain more individual heterogeneity but not necessarily lead to more discriminating power. For example, we and others have found genomic alterations previously considered tumor-specific (e.g., driver mutations and copy number variations) in normal aging esophagus,[73] so incorporating these genomic features may end up adding noise. On the other hand, multi-analyte tests, i.e., checking markers from different omic layers, are presumably more complicated and expensive. In a clinical setting, comprehensiveness is rarely the priority and often traded-off for cost efficiency; fewer markers are preferred if they can do the same job. Finally, as mentioned earlier, DMC markers have unique advantages over other omics markers.[74]

The current study has several limitations. First, our DNA methylation profiling is limited by the fixed design of microarrays. Sequencing-based profiling may provide more insights due to improved base-pair resolution and better genome coverage. Second, marker screening and functional validation were limited to DMCs primarily affecting their host protein-coding genes in a cis manner, while DMCs can exert influences at a distance (i.e., in a trans manner) and on non-coding elements as well. Moreover, these influences may not be strictly one-to-one but rather form an interconnected network. Since our panels perform relatively well, we speculate that they may capture some central relations within this "network," which requires further investigation. Third, 73.63% (67/91) of the ESCC samples we used to develop the diagnostic and the prognostic models were at an advanced stage (III or IV). Although both models were validated in the TCGA ESCC set, 67.39% (62/92) of which are at an early stage (I or II), their efficacy in patients with early-stage ESCC or precancerous lesions needs additional evaluation. Lastly, the results of this study only implicate a potential functional role of DNA methylation in ESCC development and progression, which warrants further mechanistic investigations.

In conclusion, our characterization of genome-wide DNA methylation anomalies using a multi-omic approach in 91 Chinese ESCC patients has supported that aberrant DNA methylation is an important part of ESCC development and progression. This study has also targeted a small number of potentially functional methylation CpG sites able to distinguish tumors from normal tissues or classify patients into high or low-risk groups. Using these CpG sites, we have constructed DNA-methylation panels for molecular diagnosis and prognosis of ESCC and validated them in multiple public datasets. The panels are potentially useful for clinical care of ESCC and it would be interesting to evaluate their utilities on non-invasively collected, small amount of tumor DNA, such as those obtained using Cytosponge[75] or liquid biopsy.[76]

MATERIALS AND METHODS

Study subjects and biospecimens

Individuals with ESCC ($n = 91$) were recruited from Chinese Academy of Medical Sciences Cancer Hospital (CAMSCH; Beijing, China) and Zhejiang Cancer Hospital (ZCH; Hangzhou, China) between 2010 and 2014. All subjects underwent esophagostomy and had not been treated with chemotherapy or radiotherapy prior to the surgery. ESCC tumor and adjacent normal tissue (≥5 cm from the tumor margin) were collected from each individual as described previously.[25] Histological evaluation was conducted by two pathologists to ensure that tumor specimens contained an average of >75% tumor cell nuclei with <20% necrosis, whereas normal specimens contained no tumor cells. The demographic characteristics and clinical data of the study subjects were obtained from medical records. Written informed consent was obtained from every subject and this study was approved by the Institutional Review Board of CAMSCH and ZCH.

Cell lines and cell culture

Human ESCC cell lines KYSE30 and KYSE150 were generous gifts from Dr Y. Shimada at the Kyoto University. These cell lines were maintained in RPMI 1640 medium supplemented with 10% fetal bovine serum (FBS). Cell lines used in this study were authenticated by short tandem repeat profiling and were free of mycoplasma infection.

RNA interference

Small interfering RNA (siRNA) oligos targeting *MMP13*, *YEATS2*, *HOXC10* or *NECAB2* were provided by JTSBIO (Supplementary Table S4). The transfection of each siRNA was performed with Lipofectamine 3000 (Invitrogen). The specific sequences for target genes are provided in the supplementary information.

Quantitative real-time PCR analysis

Total RNA was extracted with Trizol reagent (Invitrogen) and the reverse transcription was performed using PrimeScript™ RT reagent kit (Takara). Quantitative real-time PCR (qRT-PCR) was performed in triplicate using TB Green Premix Ex Taq (Takara). The primer sequences used for qRT-PCR of interest genes are shown in Supplementary Table S4.

Western blot analysis

Proteins were extracted using RIPA lysis buffer (Solarbio, R0020) containing PMSF (Solarbio, P0100), phosphatase inhibitor cocktail I and II (MCE, HY-K0021 and HY-K0022). In total, lysate containing 10–20 μg of protein was separated on SDS-PAGE and transferred to PVDF membranes (Millipore). Antibodies against MMP13 (ab51072), HOXC10 (ab153904) and GAPDH (ab181602) were from Abcam while antibodies against YEATS2 (24717-1-AP) and NECAB2 (12257-1-AP) were from Proteintech. The signal was captured with a SuperSignal™ West Pico/Femto Chemiluminescent Substrate kit (Thermo Fisher, 34580) analyzedthrough the Amersham Imager 600.

Cell viability and migration or invasion assays

Cell viability was measured after incubation with CCK-8 (Dojindo). Invasion assays were performed in 24-well chambers (Corning) coated with Matrigel (BD Biosciences). Cells (20×10^4) in serum-free medium were added to the coated chamber and incubated for 18 or 24 h before fixed with methanol and stained with 0.5% crystal violet. Migration assays were performed in a similar fashion but without coating the filters with Matrigel.

DNA extraction and methylation data processing

Genomic DNA was isolated from tissue samples with Allprep DNA/RNA Kit (Qiagen) and arrayed using Infinium HumanMethylation450 BeadChips (450 K array, Illumina) to detect genome-wide methylation. We then conducted data preprocessing, normalization and calculation of β-value using the R package minfi[77] (version 1.26.2). We applied the following criteria for quality control: (i) probes with detection $P \geq 0.01$ in >5% of samples were removed from all samples; (ii) probes on the X or Y chromosome were removed; (iii) probes overlapping with single nucleotide polymorphisms (SNPs) were removed; (iv) probes mapped to multiple sites in human genome were removed. Finally, 429,717 probes were kept for further analysis.

The CpG probe annotation file was downloaded from the ENCODE Project database (http://genome.ucsc.edu/ENCODE/downloads.html). Each CpG probe is annotated with the corresponding gene, genomic region (TSS1500, 200–1500 bases upstream of the transcriptional start site [TSS]; TSS200, 0–200 bases upstream of the TSS; 5′UTR, within the 5′ untranslated region, between the TSS and the ATG start site; body, between the ATG and the stop codon; irrespective of the presence of introns, exons, TSS, or promoters; 3′UTR, between the stop codon and the poly A signal), the CpG island-associated regions (shore, 0–2 kb from island; shelf, 2–4 kb from island; N, upstream 5′ of CpG island; S, downstream 3′ of CpG island) and functional regions (enhancer, predicted enhancer elements as annotated in the original 450 K design are marked "true"; DHS, DNase I hypersensitivity site).[78]

Identification of differentially methylated CpG sites

We applied a two-sided Wilcoxon signed-rank test to identify CpG sites (DMCs) differentially methylated between paired tumor and normal samples. *P* values were adjusted for multiple testing using the Benjamini–Hochberg method to control the false discovery rate (FDR). We required that significant DMCs have FDR $q < 0.05$ and the absolute median methylation difference ($|MMD|$) > 0.20. We compared the of each DMC in ESCC and matched adjacent normal samples to determine its methylation status. A DMC in the ESCC genome of a specific patient is considered hyper-methylated or hypo-methylated if the β-value of this CpG site minus the β-value of the same site in that patient's matched adjacent normal tissue sample is greater or less than 0.20, respectively. We counted the fraction of hyper-methylated or hypo-methylated CpG sites as the frequencies of hyper- or hypo-methylation events for each patient.

Methylation quantitative trait loci (meQTL) analysis

The single nucleotide polymorphism (SNP) data of the 91 study subjects were obtained from our previous DNA sequencing study.[25] From a total of 6,092,313 SNPs that have minor allele frequency ≥5% and no deviation from the Hardy-Weinberg equilibrium ($P < 1.00e\text{-}6$), we only selected SNPs within a 100-kb window centering each DMC on the same chromosome for meQTL mapping. An additive linear regression model implemented in the R package MatrixEQTL (v.2.3) was used and only SNPs with FDR $q < 0.05$ were deemed significant. A hypergeometric test was then used to assess the statistical significance of the overlap between identified meQTLs and potential ESCC risk SNPs obtained from the CCGD-ESCC database.[37]

Identification of differentially expressed genes and gene set enrichment analysis

We identified genes differentially expressed between paired tumor and normal tissue samples using Student's *t* test on log2-transformed gene expression levels (quantified by Transcript per Million, TPM). Only genes had FDR $q < 0.05$ and the relative fold change of mean expression levels > 2 or < 0.50 (tumor versus normal) were deemed significant. We replaced Student's *t* test with Wilcoxon signed-rank test and 99.75% of the genes were still differentially expressed (FDR $q < 0.05$, relative fold change > 2 or < 0.50), including all the genes we used for downstream analyses. Gene ontology (GO) analysis was conducted using the enrichGO function implemented in the R package clusterProfiler (v. 3.8.1)[79] and only the top 10 enriched GO terms were plotted.

Identification of correlations between DNA methylation and gene expression

We examined the correlations between the methylation levels of DMCs and the expression levels of their corresponding genes using Spearman's rank correlation and considered a correlation statistically significant if FDR $q < 0.05$ and the absolute Spearman rank correlation coefficient $|r| > 0.30$. We further considered the consistency of direction between the methylation level of DMC and the corresponding gene expression level for a more rigorous screening. For DMCs located in promoter, we applied the following criteria: (a) negative correlations (Spearman $r < -0.30$, $P < 0.05$) between DMCs and their corresponding genes; (b) hyper-methylation of DMC corresponding silencing of gene expression or hypo-methylation of DMC corresponding upregulation of gene expression. For DMCs located in gene-body, we applied the inverse criteria: (a) positive correlations (Spearman $r > 0.30$, $P < 0.05$) between DMCs and their corresponding genes; (b) hyper-methylation of DMC corresponding upregulation of gene expression or hypo-methylation of DMC corresponding silencing of gene expression.

Development of a panel of DMCs for ESCC diagnosis

The panel was developed in 4 steps: (a) randomly divide 91 patients into the training ($n = 60$) and the validation ($n = 31$) sets with a 2:1 ratio; (b) From all the 1034 DMCs identified from 91 patients (method described above), select important variables for the training set using random forest analysis, with the feature dropping fraction of each iteration set at 1/3 according to the importance score; (c) use the least absolute shrinkage and selection operator (LASSO)-penalized logistic regression (a binomial model, 10-fold cross-validation) to further select the variables obtained in the previous step; (d) carry out the diagnostic model in the validation dataset and TCGA ESCC methylation dataset.

Development of a pane of DMCs for prognostic risk prediction

For each DMC identified in our patient set, we fitted a univariate Cox proportional hazard model with that DMC as the covariate and only retained DMCs with nominal $P < 0.05$. Then, for each

retained DMC, we fit a multivariate Cox proportional hazard model with that DMC as the predictor variable and age, sex, smoking status, drinking status and tumor TNM stage as covariates, and again, only retain DMCs with nominal $P < 0.05$. A sum of the methylation level of each remaining DMC multiplied by its respective natural logarithm of hazard ratio (HR) in our patient sample is the prognostic prediction model. And we applied the prognostic model in both our patients and TCGA ESCC patients.

Other analyses

Unsupervised hierarchical clustering based on the methylation difference between ESCC and adjacent normal tissue samples was conducted using the pheatmap function implemented in the R package pheatmap (v. 1.0.12). The R package survival (v. 3.2-7) and survminer (v. 0.4.8) were used for survival analysis. Overall survival time was estimated by the Kaplan–Meier method and the differences were examined by the log-rank test. Hazard ratios (HRs) and their 95% confidence intervals (CIs) were calculated with the Cox proportional hazards model. All statistical tests were two-sided tests and $P < 0.05$ was considered significant unless indicated. We used R 3.6.1 (https://www.r-project.org/).

DATA AVAILABILITY

The methylation data generated in this study are deposited in the OMIX, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (http://bigd.big.ac.cn/omix, accession number OMIX267). The other genetic and transcriptomic data of the same individuals are available through our earlier publications.[25] We obtained DNA methylation data (level 3) of 95 ESCCs and 14 normal esophageal tissue samples from The Cancer Genome Atlas (TCGA) (http://gdac.broadinstitute.org/) for comparative analysis. We also obtained methylation data from the Gene Expression Omnibus (GEO) database (GSE52826 and GSE77991), which include 4 ESCCs, 14 normal tissues adjacent to tumors from patients, and 21 esophageal mucosal tissues from healthy individuals. Methylation data of 16 ESCC markers in TCGA 22 cancer types were downloaded from SMART (Shiny methylation analysis resource tool) app.[80]

CODE AVAILABILITY

We used published software for all our analyses as indicated. Other accompanying code is available from the authors upon request.

ACKNOWLEDGEMENTS

Supported by National Natural Science Foundation of China (81988101 to D.L. and C.W.), National Science Fund for Distinguished Young Scholars (81725015 to C.W.), Chinese Academy Medical Sciences Innovation Fund for Medical Sciences (2021-I2M-1-013 to D.L., C.W. and W.T.; 2019-I2M-2-001 to D.L. and C.W.), Beijing Outstanding Young Scientist Program (BJJWZYJH01201910023027 to C.W.) and the National Key R&D Program of China (2021YFC2502000 to Y.L.).

AUTHOR CONTRIBUTIONS

D.L., C.W. and J.S. conceptualized and supervised this study. Y.X. and W.G. contributed to the study design and performed most experiments. C.M., W.L., Y.L., W.F., A.L. and Y.C. responded to clinical data, sample collection and preparation. Y.X., Y.L., X.W., H.Z. and Y.L. performed bioinformatics and statistical analysis. Y.X. and Y.L. prepared the manuscript. D.L., C.W. and J.S. reviewed and prepared the final manuscript. All authors read and approved the final manuscript.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41392-022-00873-8.

Competing interests: The authors declare no competing interests.

REFERENCES

1. Kamangar, F., Dores, G. M. & Anderson, W. F. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *J. Clin. Oncol.* **24**, 2137–2150 (2006).
2. Enzinger, P. C. & Mayer, R. J. Esophageal cancer. *N. Engl. J. Med.* **349**, 2241–2252 (2003).
3. Besharat, S. et al. Inoperable esophageal cancer and outcome of palliative care. *World J. Gastroenterol.* **14**, 3725–3728 (2008).
4. Wang, A. H. et al. Epidemiological studies of esophageal cancer in the era of genome-wide association studies. *World J. Gastrointest. Pathophysiol.* **5**, 335–343 (2014).
5. Schubeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).
6. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361**, 1336–1340 (2018).
7. Baylin, S. B. et al. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum. Mol. Genet.* **10**, 687–692 (2001).
8. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
9. Irizarry, R. A. et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
10. Walter, K. et al. Discovery and development of DNA methylation-based biomarkers for lung cancer. *Epigenomics* **6**, 59–72 (2014).
11. Okugawa, Y., Grady, W. M. & Goel, A. Epigenetic alterations in colorectal cancer: emerging biomarkers. *Gastroenterology* **149**, 1204–1225 e1212 (2015).
12. Tahara, T. & Arisawa, T. DNA methylation as a molecular biomarker in gastric cancer. *Epigenomics* **7**, 475–486 (2015).
13. Talukdar, F. R. et al. Genome-wide dna methylation profiling of esophageal squamous cell carcinoma from global high-incidence regions identifies crucial genes and potential cancer markers. *Cancer Res.* **81**, 2612–2624 (2021).
14. Laird, P. W. The power and the promise of DNA methylation markers. *Nat. Rev. Cancer* **3**, 253–266 (2003).
15. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome - biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).
16. Koch, A. et al. Analysis of DNA methylation in cancer: location revisited. *Nat. Rev. Clin. Oncol.* **15**, 459–466 (2018).
17. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *Lancet* **392**, 777–786 (2018).
18. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
19. Ma, K., Cao, B. & Guo, M. The detective, prognostic, and predictive value of DNA methylation in human esophageal squamous cell carcinoma. *Clin. Epigenetics* **8**, 43 (2016).
20. Lin, D. C., Wang, M. R. & Koeffler, H. P. Genomic and epigenomic aberrations in esophageal squamous cell carcinoma and implications for patients. *Gastroenterology* **154**, 374–389 (2018).
21. Pu, W. et al. Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC). *Clin. Epigenetics* **9**, 129 (2017).
22. Wang, C. et al. Identification of hyper-methylated tumor suppressor genes-based diagnostic panel for esophageal squamous cell carcinoma (ESCC) in a Chinese Han population. *Front. Genet.* **9**, 356 (2018).
23. Chen, C. et al. Genome-wide profiling of DNA methylation and gene expression in esophageal squamous cell carcinoma. *Oncotarget* **7**, 4507–4521 (2016).
24. Fraser, H. B., Lam, L. L., Neumann, S. M. & Kobor, M. S. Population-specificity of human DNA methylation. *Genome Biol.* **13**, R8 (2012).
25. Chang, J. et al. Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nat. Commun.* **8**, 15290 (2017).
26. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
27. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
28. Zhang, C. et al. The novel 19q13 KRAB zinc-finger tumour suppressor ZNF382 is frequently methylated in oesophageal squamous cell carcinoma and antagonises Wnt/beta-catenin signalling. *Cell Death Dis.* **9**, 573 (2018).
29. Su, J. et al. Homeobox oncogene activation by pan-cancer DNA hypermethylation. *Genome Biol.* **19**, 108 (2018).
30. Zhao, Y. et al. ZNF582 hypermethylation promotes metastasis of nasopharyngeal carcinoma by regulating the transcription of adhesion molecules Nectin-3 and NRXN3. *Cancer Commun. (Lond.)* **40**, 721–737 (2020).
31. Dong, Z. et al. Aberrant hypermethylation-mediated downregulation of antisense lncRNA ZNF667-AS1 and its sense gene ZNF667 correlate with progression and prognosis of esophageal squamous cell carcinoma. *Cell Death Dis.* **10**, 930 (2019).
32. Yan, C. et al. An esophageal adenocarcinoma susceptibility locus at 9q22 also confers risk to esophageal squamous cell carcinoma by regulating the function of BARX1. *Cancer Lett.* **421**, 103–111 (2018).

33. Lin, C. et al. Transcriptional and posttranscriptional regulation of HOXA13 by lncRNA HOTTIP facilitates tumorigenesis and metastasis in esophageal squamous carcinoma cells. *Oncogene* **36**, 5392–5406 (2017).
34. Suo, D. et al. HOXC10 upregulation confers resistance to chemoradiotherapy in ESCC tumor cells and predicts poor prognosis. *Oncogene* **39**, 5441–5454 (2020).
35. Cheng, Y. et al. KRAB zinc finger protein ZNF382 is a proapoptotic tumor suppressor that represses multiple oncogenes and is commonly silenced in multiple carcinomas. *Cancer Res.* **70**, 6516–6526 (2010).
36. Jiang, Y. Y. et al. TP63, SOX2, and KLF5 establish a core regulatory circuitry that controls epigenetic and transcription patterns in esophageal squamous cell carcinoma cell lines. *Gastroenterology* **159**, 1311–1327 e1319 (2020).
37. Peng, L. et al. CCGD-ESCC: a comprehensive database for genetic variants associated with esophageal squamous cell carcinoma in chinese population. *Genomics Proteom. Bioinforma.* **16**, 262–268 (2018).
38. Kuo, I. Y. et al. Prognostic CpG methylation biomarkers identified by methylation array in esophageal squamous cell carcinoma patients. *Int J. Med. Sci.* **11**, 779–787 (2014).
39. Li, X. et al. Identification of a DNA methylome profile of esophageal squamous cell carcinoma and potential plasma epigenetic biomarkers for early diagnosis. *PLoS ONE* **9**, e103162 (2014).
40. Adams, L. et al. Promoter methylation in cytology specimens as an early detection marker for esophageal squamous dysplasia and early esophageal squamous cell carcinoma. *Cancer Prev. Res. (Philos.)* **1**, 357–361 (2008).
41. Li, B. et al. Hypermethylation of multiple tumor-related genes associated with DNMT3b up-regulation served as a biomarker for early diagnosis of esophageal squamous cell carcinoma. *Epigenetics* **6**, 307–316 (2011).
42. Cheng, C. P. et al. Network-based analysis identifies epigenetic biomarkers of esophageal squamous cell carcinoma progression. *Bioinformatics* **30**, 3054–3061 (2014).
43. Cancer Genome Atlas Research N. et al. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).
44. Cao, W. et al. Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat. Commun.* **11**, 3675 (2020).
45. Teschendorff, A. E. et al. The multi-omic landscape of transcription factor inactivation in cancer. *Genome Med.* **8**, 89 (2016).
46. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
47. Singh, H., Khan, A. A. & Dinner, A. R. Gene regulatory networks in the immune system. *Trends Immunol.* **35**, 211–218 (2014).
48. Etoh, T. et al. Increased expression of collagenase-3 (MMP-13) and MT1-MMP in oesophageal cancer is related to cancer aggressiveness. *Gut* **47**, 50–56 (2000).
49. Osako, Y. et al. Regulation of MMP13 by antitumor microRNA-375 markedly inhibits cancer cell migration and invasion in esophageal squamous cell carcinoma. *Int J. Oncol.* **49**, 2255–2264 (2016).
50. Mi, W. et al. YEATS2 links histone acetylation to tumorigenesis of non-small cell lung cancer. *Nat. Commun.* **8**, 1088 (2017).
51. Feng, W. et al. Multiple histone deacetylases repress tumor suppressor gene ARHI in breast cancer. *Int J. Cancer* **120**, 1664–1668 (2007).
52. Thole, T. M. et al. Neuroblastoma cells depend on HDAC11 for mitotic cell cycle progression and survival. *Cell Death Dis.* **8**, e2635 (2017).
53. Gong, D., Zeng, Z., Yi, F. & Wu, J. Inhibition of histone deacetylase 11 promotes human liver cancer cell apoptosis. *Am. J. Transl. Res.* **11**, 983–990 (2019).
54. Zhang, D. L. et al. Genome-wide identification of transcription factors that are critical to non-small cell lung cancer. *Cancer Lett.* **434**, 132–143 (2018).
55. Chen, L. et al. miR-203a-3p promotes colorectal cancer proliferation and migration by targeting PDE4D. *Am. J. Cancer Res.* **8**, 2387–2401 (2018).
56. Qiang, Z. et al. Inhibition of TPL2 by interferon-alpha suppresses bladder cancer through activation of PDE4D. *J. Exp. Clin. Cancer Res.* **37**, 288 (2018).
57. Cardona, D. M., Zhang, X. & Liu, C. Loss of carbamoyl phosphate synthetase I in small-intestinal adenocarcinoma. *Am. J. Clin. Pathol.* **132**, 877–882 (2009).
58. Liu, H., Dong, H., Robertson, K. & Liu, C. DNA methylation suppresses expression of the urea cycle enzyme carbamoyl phosphate synthetase 1 (CPS1) in human hepatocellular carcinoma. *Am. J. Pathol.* **178**, 652–661 (2011).
59. Hsien Lai, S. et al. PDE4 subtypes in cancer. *Oncogene* **39**, 3791–3802 (2020).
60. Kim, J. et al. CPS1 maintains pyrimidine pools and DNA synthesis in KRAS/LKB1-mutant lung cancer cells. *Nature* **546**, 168–172 (2017).
61. Zhai, Y. et al. Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion. *Cancer Res.* **67**, 10163–10172 (2007).
62. Tan, Z. et al. Overexpression of HOXC10 promotes angiogenesis in human glioma via interaction with PRMT5 and upregulation of VEGFA expression. *Theranostics* **8**, 5143–5158 (2018).
63. Tang, X. L. et al. HOXC10 promotes the metastasis of human lung adenocarcinoma and indicates poor survival outcome. *Front Physiol.* **8**, 557 (2017).
64. Miwa, T. et al. Homeobox C10 influences on the malignant phenotype of gastric cancer cell lines and its elevated expression positively correlates with recurrence and poor survival. *Ann. Surg. Oncol.* **26**, 1535–1543 (2019).
65. Busser, B. et al. The multiple roles of amphiregulin in human cancer. *Biochim. Biophys. Acta* **1816**, 119–131 (2011).
66. Addison, C. L. et al. Plasma transforming growth factor alpha and amphiregulin protein levels in NCIC Clinical Trials Group BR.21. *J. Clin. Oncol.* **28**, 5247–5256 (2010).
67. Li, X. D. et al. Amphiregulin and epiregulin expression in colorectal carcinoma and the correlation with clinicopathological characteristics. *Onkologie* **33**, 353–358 (2010).
68. Yamada, M. et al. Amphiregulin is a promising prognostic marker for liver metastases of colorectal cancer. *Clin. Cancer Res.* **14**, 2351–2356 (2008).
69. Steponaitis, G. et al. Significance of amphiregulin (AREG) for the outcome of low and high grade astrocytoma patients. *J. Cancer* **10**, 1479–1488 (2019).
70. Jackson, R. S. 2nd, Cho, Y. J., Stein, S. & Liang, P. CYFIP2, a direct p53 target, is leptomycin-B sensitive. *Cell Cycle* **6**, 95–103 (2007).
71. Jiao, S. et al. Inhibition of CYFIP2 promotes gastric cancer cell proliferation and chemoresistance to 5-fluorouracil through activation of the Akt signaling pathway. *Oncol. Lett.* **13**, 2133–2140 (2017).
72. Cheng, A. S. et al. Helicobacter pylori causes epigenetic dysregulation of FOXD3 to promote gastric carcinogenesis. *Gastroenterology* **144**, 122–133 e129 (2013).
73. Li, R. et al. A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* **597**, 398–403 (2021).
74. Deger, T. et al. High-throughput and affordable genome-wide methylation profiling of circulating cell-free DNA by methylated DNA sequencing (MeD-seq) of LpnPI digested fragments. *Clin. Epigenetics* **13**, 196 (2021).
75. Januszewicz, W. et al. Safety and acceptability of esophageal cytosponge cell collection device in a pooled analysis of data from individual patients. *Clin. Gastroenterol. Hepatol.* **17**, 647–656 e641 (2019).
76. Li, W. & Zhou, X. J. Methylation extends the reach of liquid biopsy in cancer detection. *Nat. Rev. Clin. Oncol.* **17**, 655–656 (2020).
77. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
78. Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
79. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
80. Li, Y., Ge, D. & Lu, C. The SMART App: an interactive web application for comprehensive DNA methylation analysis and visualization. *Epigenetics Chromatin* **12**, 71 (2019).