Molecular Oncology

FEBS PRESS science publishing by scientists

HPV-related methylation-based reclassification and risk stratification of cervical cancer

Si Yang[1,2], Ying Wu[1,2], Shuqian Wang[2], Peng Xu[1], Yujiao Deng[1,2], Meng Wang[1], Kang Liu[3], Tian Tian[1], Yuyao Zhu[1,2], Na Li[1,2], Linghui Zhou[1,2], Zhijun Dai[2] and Huafeng Kang[1]

1 Department of Oncology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
2 Department of Breast Surgery, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China
3 Department of Hepatobiliary Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Keywords
cervical cancer; DNA methylation; human papillomavirus; prognosis; risk stratification; subtype

Correspondence
H. Kang, Department of Oncology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710004, China
Tel: +86 13759887377
E-mail: kanghuafeng1973@126.com
and
Z. Dai, Department of Breast Surgery, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou 310003, China
Tel: +86 13991393919
E-mail: dzj0911@126.com

Si Yang and Ying Wu contributed equally to this work

(Received 23 December 2019, revised 1 April 2020, accepted 9 May 2020, available online 2 June 2020)

doi:10.1002/1878-0261.12709

Human papillomavirus (HPV) is a clear etiology of cervical cancer (CC). However, the associations between HPV infection and DNA methylation have not been thoroughly investigated. Additionally, it remains unknown whether HPV-related methylation signatures can identify subtypes of CC and stratify the prognosis of CC patients. DNA methylation profiles were obtained from The Cancer Genome Atlas to identify HPV-related methylation sites. Unsupervised clustering analysis of HPV-related methylation sites was performed to determine the different CC subtypes. CC patients were categorized into cluster 1 (Methylation-H), cluster 2 (Methylation-M), and cluster 3 (Methylation-L). Compared to Methylation-M and Methylation-L, Methylation-H exhibited a significantly improved overall survival (OS). Gene set enrichment analysis (GSEA) was conducted to investigate the functions that correlated with different CC subtypes. GSEA indicated that the hallmarks of tumors, including KRAS signaling, TNFα signaling via NF-κB, inflammatory response, epithelial–mesenchymal transition, and interferon-gamma response, were enriched in Methylation-M and Methylation-L. Based on mutation and copy number variation analyses, we found that aberrant mutations, amplifications, and deletions among the MYC, Notch, PI3K-AKT, and RTK-RAS pathways were most frequently detected in Methylation-H. Additionally, mutations, amplifications, and deletions within the Hippo, PI3K-AKT, and TGF-β pathways were presented in Methylation-M. Genes within the cell cycle, Notch, and Hippo pathways possessed aberrant mutations, amplifications, and deletions in Methylation-L. Moreover, the analysis of tumor microenvironments revealed that Methylation-H was characterized by a relatively low degree of immune cell infiltration. Finally, a prognostic signature based on six HPV-related methylation sites was developed and validated. Our study revealed that CC patients could be classified into three heterogeneous clusters based on HPV-related methylation signatures.

Abbreviations
aDCs, activated dendritic cells; AUC, the area under the curve; CC, cervical cancer; CI, confidence interval; CNV, copy number variation; DCs, dendritic cells; DMP, differentially methylated probe; ESTIMATE, Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data; GSEA, gene set enrichment analysis; GX, unknown histological grade; HPV, human papillomavirus; HR, hazard ratio; iDCs, immature dendritic cells; LASSO, least absolute shrinkage and selection operator; NES, normalized enrichment score; NK, natural killer; NX, unknown N stage; OS, overall survival; PCA, principal component analysis; pDCs, plasmacytoid dendritic cells; ROC, receiver operating characteristic; SsGSEA, single sample gene set enrichment analysis; TCGA, The Cancer Genome Atlas; Tcm cells, central memory T cells; Tem cells, effector memory T cells; Tfh cells, follicular helper T cells; Tgd cells, gamma delta T cells; Th1 cells, type 1T helper cells; Treg cells, regulatory T cells; TSG, tumor suppressor gene; TX, unknown T stage.

Additionally, we derived a prognostic signature using six HPV-related methylation sites that stratified the OS of patients with CC into high- and low-risk groups.

1. Introduction

Cervical cancer (CC) is a major public health concern and is the fourth most frequently diagnosed cancer and the fourth leading cause of cancer-related death in women worldwide (Bray *et al.*, 2018). A persistent infection with oncogenic human papillomavirus (HPV) can lead to cervical precancerous lesions that may ultimately develop into cancer (Crosbie *et al.*, 2013). There are more than 100 identified HPV genotypes, and types 16 and 18 are the most common in CC (Munoz *et al.*, 2006). HPV plays a significant role in the pathogenesis of CC; it affects apoptosis, cell cycle, cell adhesion, and DNA repair mechanisms within the host cell and can also activate immune responses (Coussens and Werb, 2002; Whiteside *et al.*, 2008). The integration of HPV virus into the host genome often occurs within the transcribed genomic region, and this mechanism is utilized by the virus to increase the expression of certain viral products, including the E6 and E7 viral oncogenes (Schmitz *et al.*, 2012; Ziegert *et al.*, 2003). Additionally, the integration of HPV virus is strongly associated with the development of CC (Li *et al.*, 2008).

Aberrant DNA methylation is an epigenetic hallmark of tumors and leads to tumor development and progression by silencing tumor suppressor genes (TSGs) and activating oncogenes (Das and Singal, 2004; Egger *et al.*, 2004). The characteristics of DNA methylation make epigenetic changes ideal and clinically applicable biomarkers for diagnosis or use as prognostic indicators in cancer (Keeley *et al.*, 2013). A growing number of studies have shown that aberrant DNA methylation plays a significant role in tumor progression, and DNA methylation can serve as a biomarker for predicting the prognosis of patients with a variety of tumors (Guo *et al.*, 2004; Roh *et al.*, 2016; Zhou *et al.*, 2014). In CC, aberrant DNA methylation can occur on the integrated viral DNA, but it can also occur within the host cell genome (Yanatatsaneejit *et al.*, 2011). HPV infection has been observed to be correlated with the regulation of DNA methylation in CC. Both E6 and E7 oncoproteins encoded by HPV type 16 affect the DNA methyltransferase DNMT1 (Au Yeung *et al.*, 2010; Burgers *et al.*, 2007). The E7 protein directly combines with DNMT1 and stimulates its DNA methyltransferase activity (Burgers *et al.*, 2007). On the other hand, E6 protein has been reported to upregulate the DNMT1 through suppression of p53 (Au Yeung *et al.*, 2010). The upregulation of DNA methyltransferases by HPV oncoproteins can increase methylation of the host cell genome and repress transcription of TSGs. HPV can drive progression of CC through the aberrant DNA methylation in plenty of TSGs, such as E-cadherin (Laurson *et al.*, 2010), p53 (Moody and Laimins, 2010), and RB1 (Yim and Park, 2005).

Many recent studies examining CC investigated only aberrant DNA methylation of one or a few genes using relatively small sample cohorts. These studies also ignored the association of HPV infection with DNA methylation. There are variations in the DNA methylation profiles, suggesting that novel methylation signatures are required for diagnosis and predicting outcomes of CC. Therefore, this study was conducted to determine the epigenetic alterations involved in CC progression and HPV infection. The DNA methylation profiles of CC samples from The Cancer Genome Atlas (TCGA) were analyzed. HPV-related methylation sites were obtained in the present study. We then performed unsupervised hierarchical clustering of HPV-related methylation sites to determine the subgroups of CC patients. The gene expression RNA-sequencing, mutation, and copy number variation (CNV) profiles in CC patients within the different methylation subgroups were investigated. Finally, a new signature possessing predictive power based on HPV-related methylation sites was developed and validated to stratify the prognosis of CC.

2. Materials and Methods

2.1. Data collection and processing

The Cancer Genome Atlas DNA methylation data from 309 CC samples and three adjacent samples based on the Illumina HumanMethylation 450 (450K) platform (Illumina Inc., San Diego, CA, USA) were downloaded from UCSC xena (https://xena.ucsc.edu/). The genomic annotation of the CpG sites was based on GRCh38. The methylation levels of the CpG sites

were estimated as beta-values and calculated as M/(M + U), where M is the signal from methylated beads, and U is the signal from un-methylated beads at the targeted CpG site (Bibikova *et al.*, 2011). For each CpG site, their beta-values ranged from 0 (no DNA methylation) to 1 (100% DNA methylation) (Bibikova *et al.*, 2011).

Clinical information for 307 CC patients was obtained from UCSC xena. A total of 13 CC patients were then excluded because their survival time was zero. We extracted the information regarding HPV infection status from Table S2 of the study published by TCGA (Cancer Genome Atlas Research Network, 2017).

The RNA-sequencing gene expression, somatic mutation, and CNV profiles for 306, 289, and 297 patients with CC, respectively, were obtained from TCGA data portal (https://portal.gdc.cancer.gov/). Somatic mutation data, which stored in the form of Mutation Annotation Format, were analyzed and summarized using maftools (Mayakonda *et al.*, 2018). Significant amplification or deletion alterations were determined using GISTIC 2.0 based on a robust algorithm to detect recurrent somatic CNVs by evaluating the frequency and amplitude of corresponding events (Mermel *et al.*, 2011).

2.2. Identification and screening of HPV-related methylation sites

ChAMP was used to perform quality control, standardization, and calculation of methylation sites and regions (Tian *et al.*, 2017). By using the ChAMP package (parameters: adjusted P-value < 0.05, |deltabeta| > 0.2), differentially methylated probes (DMPs) between CC and adjacent tissue were identified. DMPs between HPV-positive and HPV-negative CC tissue were also identified based on the same criterion. The intersection between these two groups of DMPs was identified as HPV-related methylation sites in the present study. Probes exhibiting deltabeta > 0.2 and adjusted P-value < 0.05 were characterized as hypermethylated, and those exhibiting deltabeta < −0.2 and adjusted P-value < 0.05 were characterized as hypomethylated.

2.3. Unsupervised hierarchical cluster analysis

To identify subtypes of CC patients, we performed unsupervised hierarchical clustering based on DNA methylation data. Clustering was performed using the beta-values of the HPV-related methylation sites with prognostic value.

2.4. Development and validation of the HPV-related methylation signature

For further analyses, 294 patients possessing survival data were screened to investigate the relationship between DNA methylation levels and OS in CC. These 294 patients with CC were divided randomly and equally into two datasets (a training dataset and a testing dataset). The training dataset was used for identifying and establishing a prognostic signature, and the testing dataset was used for validating its predictive effectiveness (Yang *et al.*, 2019). First, a univariate Cox regression analysis was performed to determine HPV-related methylation sites with prognostic value in the training dataset. If the P-value was < 0.01, the corresponding methylated sites were considered as candidate methylated sites. Through preliminary screening, there may be excess candidate methylated sites. Therefore, least absolute shrinkage and selection operator (LASSO)-penalized Cox proportional hazards regression analysis was conducted to further reduce candidate methylated sites by using the R package 'glmnet' (Friedman *et al.*, 2010; Wang *et al.*, 2019b); LASSO is a popular algorithm that adopts explicable prediction rules and can solve the collinearity problem by dimension reduction (Gui and Li, 2005). Third, a stepwise multivariate Cox regression analysis was performed to further screen the methylated sites. An optimal predictive model was selected, similar to that used for the lowest Akaike information criterions value. The HPV-related methylation sites in the model were utilized to establish the risk signature. The risk score was calculated using the following formula: Risk score = beta-value of methylated site 1 × coefficient + beta-value of methylated site 2 × coefficient + … beta-value of methylated site n × coefficient. The risk score for each patient in the training dataset, testing dataset, and entire dataset was calculated based on this formula. Based on the median cutoff of the risk score, patients with CC were grouped into high- and low-risk groups. Survival analysis was performed to evaluate the OS difference between high- and low-risk groups that were stratified according to the signature using the 'survival' R package (Holleczek and Brenner, 2013). To validate the prognostic capability of this signature, we calculated area under the curve (AUC) using the 'timeROC' R package (Lorent *et al.*, 2014). The high AUC suggested accurate predictive capability of the signature (Lorent *et al.*, 2014).

2.5. Gene set enrichment analysis (GSEA)

To explore differences in potential biological processes in CC patients from different clusters, GSEA was

performed using the hallmark gene sets (h.all.v7.0.symbols), which were obtained from the Molecular Signatures Database (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). The hallmark gene sets display coordinate expression and represent well-defined biological processes, providing a clearer biological space for GSEA (Halvorsen *et al.*, 2019; Liberzon *et al.*, 2015). The 'fgsea' R package was used, and 10 000 permutations were performed for each parameter analyzed to calculate the enrichment scores based on the threshold of adjusted P-value < 0.05 (Sergushichev, 2016).

2.6. Single sample gene set enrichment analysis (ssGSEA)

The immune infiltration levels of 24 different immune cell types were estimated by performing ssGSEA in the R package 'gsva' (http://www.bioconductor.org/packages/release/bioc/html/GSVA.html). The marker gene set for 24 types of immune cells was obtained from a previous study (Bindea *et al.*, 2013). The ssGSEA algorithm transforms marker gene expression patterns into quantities of immune cell populations in individual tumor samples (Rooney *et al.*, 2015). This algorithm could identify 24 types of immune cells, including innate immune cells [natural killer (NK) cells, NK CD56dim cells, NK CD56bright cells, dendritic cells (DCs), activated DCs (aDCs), plasmacytoid DCs (pDCs), immature DCs (iDCs), neutrophils, mast cells, eosinophils, and macrophages] and adaptive immune cells [B cells, CD8+ T cells, cytotoxic cells, T cells, T helper cells, central memory T cells (Tcm cells), effector memory T cells (Tem cells), follicular helper T cells (Tfh cells), gamma delta T cells (Tgd cells), type 1T helper cells (Th1 cells), type 2T helper cells, type 17T helper cells, and regulatory T cells (Treg cells)] (Zhang *et al.*, 2018).

2.7. Tumor microenvironments analysis

Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data (ESTIMATE) algorithm was used to calculate immune and stromal scores to predict the infiltration of tumor microenvironment cells, by analyzing specific gene expression signature of immune and stromal cells (Yoshihara *et al.*, 2013).

2.8. Statistical analyses

All statistical analyses were conducted by using GRAPHPAD PRISM 7(GraphPad Software Inc., San Diego, CA, USA) and R software (version 3.5.2, R Foundation for Statistical Computing, Vienna, Austria) unless otherwise stated. Univariate and multivariate Cox regression analyses were conducted to investigate the prognostic value of HPV-related methylation signature and some clinicopathological variables. Volcano plots were created using R package 'ggplot2'. Heatmaps were created using R package 'pheatmap'. Violin plots were created using R package 'vioplot'. Forest plots were created using R package 'forestplot'. All statistical results with a P-value < 0.05 were considered significant.

3. RESULTS

3.1. Identification and screening of HPV-related methylation sites

After removing a number of undetected methylated probes, a total of 312 samples (309 CC and three adjacent normal samples) and 372 137 DNA methylated sites were obtained from TCGA. Additionally, a total of 178 samples (169 HPV-positive and nine HPV-negative CC samples) and 378 494 DNA methylated sites were obtained. By performing ChAMP with the adjusted P-value < 0.05 and |deltabeta| > 0.2, we identified 35 678 DMPs between tumor and normal tissue (Fig. 1A). A total of 48 190 DMPs were screened between HPV-positive and HPV-negative CC tissues (Fig. 1B). By intersecting these two groups of DMPs, we acquired 9249 HPV-related methylation sites (Fig. S1). Then, 294 CC patients with survival data containing information on 9249 HPV-related methylation sites were included in further analysis. These 294 patients were divided randomly and equally into a training dataset and a testing dataset. The clinicopathological characteristics of patients are summarized in Table S1. There were no statistically significant differences between these two groups.

3.2. Identification of three methylation clusters of CC exhibiting distinct survival outcomes

Univariate Cox regression analysis was conducted to screen DNA methylation sites that related to overall survival (OS) by using the HPV-related methylation sites as variables in the training dataset. A total of 191 HPV-related methylation sites that were related to OS in the training dataset (P < 0.05) were picked out for unsupervised clustering analysis. Consequently, 294 patients with CC were categorized into three clusters (Fig. 2A). The patients in cluster 1 (Methylation-H) exhibited frequent hypermethylation among the 191

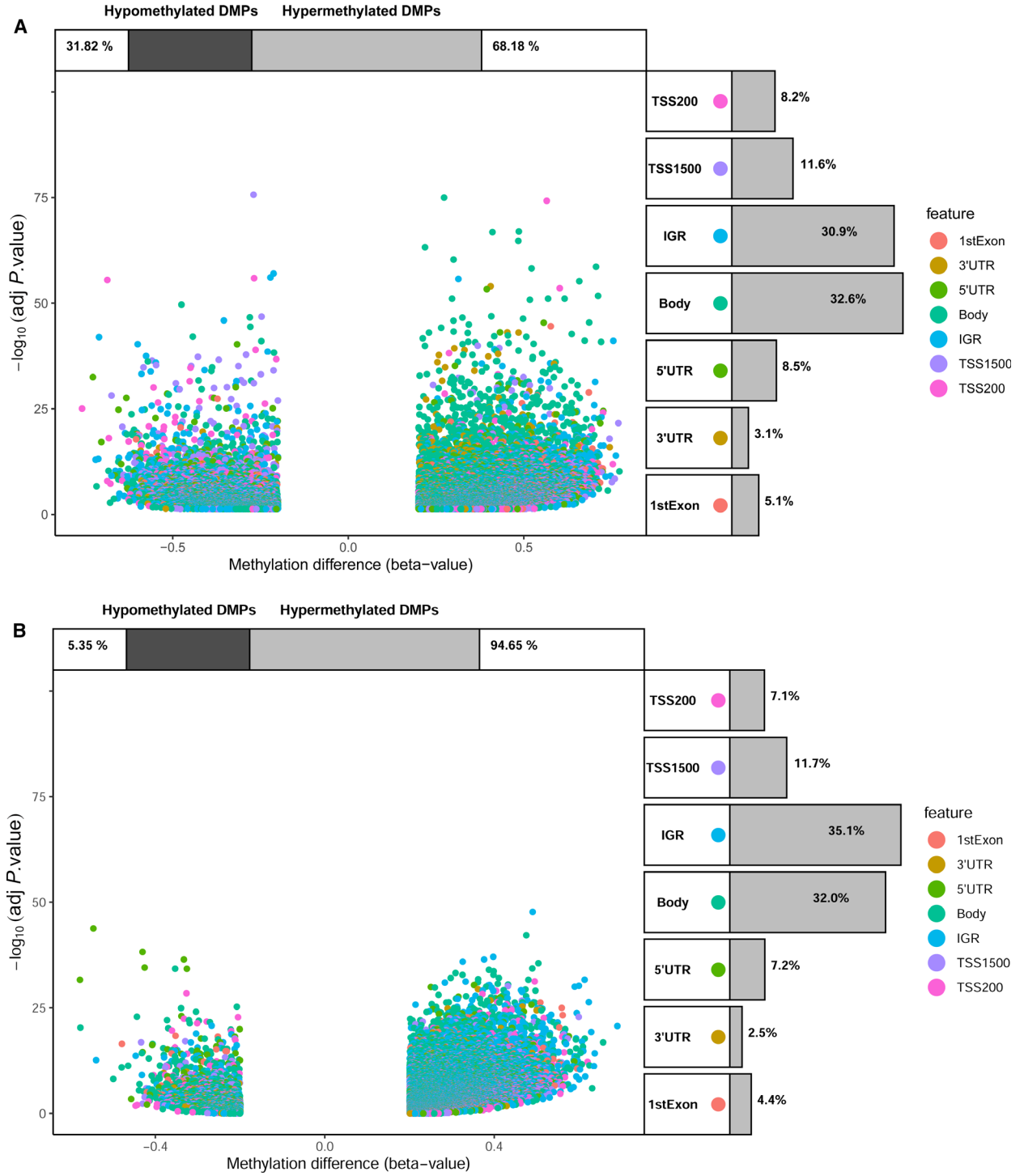


Fig. 1. DMP analyses in cases and controls. (A) DMP analysis between CC and normal tissue. (B) DMP analysis between HPV-positive and HPV-negative CC tissues. Volcano plots of DMPs and position of methylation probes in relation to the gene (IGR, intergenic region; TSS, transcription start site; UTR) are displayed. The percentages of hypomethylated and hypermethylated DMPs are displayed on top. The proportions of different genomic features are shown on the right.

methylation sites. The methylation level of cluster 3 (Methylation-L) was the lowest, and the methylation level of cluster 2 (Methylation-M) was intermediate. Compared to Methylation-M or Methylation-L, Methylation-H exhibited a significantly higher OS ($P = 0.009$, Fig. 2B). To check for cluster stability, clustering was compared for 9249 HPV-related methylation sites (Fig. S2). The formation of clusters was robust for the 9249 HPV-related methylation sites selected for calculation. The results demonstrated the presence of three different methylation clusters. Principal component analysis (PCA), which was further employed to compare the transcriptional profiles among the three clusters, displayed a clear distinction among these clusters. In detail, PCA showed that the samples from the three clusters were well separated from each other (Fig. 2C).

3.3. Biological processes and mechanisms related to different clusters of CC

Gene set enrichment analysis was performed to investigate the underlying biological processes and mechanisms related to different clusters of CC. The results revealed that the hallmarks of tumors, including KRAS signaling [normalized enrichment score (NES) = 1.44, adjusted P-value < 0.05], coagulation (NES = 1.57, adjusted P-value < 0.05), TNFα signaling via NF-κB (NES = 1.52, adjusted P-value < 0.05), inflammatory response (NES = 1.60, adjusted P-value < 0.01), and epithelial–mesenchymal transition (NES = 1.84, adjusted P-value < 0.01), were significantly and positively associated with Methylation-L (Fig. 3A). E2F targets (NES = −1.65, adjusted P-value < 0.01) were significantly and negatively associated with Methylation-L compared to Methylation-H (Fig. 3A). Additionally, interferon-gamma response (NES = 1.75, adjusted P-value < 0.05) and TNFα signaling via NF-κB (NES = 1.66, adjusted P-value < 0.05) were significantly and positively associated with Methylation-M compared to Methylation-H (Fig. 3B).

3.4. Analysis of mutations and CNVs in CC patients within the three clusters

We further investigated the genomic alterations that could be correlated with different survival outcomes in the three clusters, with an aim to identify potential drug targets to reverse the poor prognosis of Methylation-M and Methylation-L.

The mutation profiles in CC patients within the three methylation clusters were investigated. Among the 294 CC patients, 276 possessed available somatic mutation data. The top 30 most frequently mutated genes in the three clusters are presented in Fig. 4A–C. The mutation frequencies of 10 common oncogenic pathways among the three clusters were calculated (Figs 4D and S3–S8). Mutations among the MYC, Notch, PI3K-AKT, and RTK-RAS signaling pathways were most frequently detected in Methylation-H (Figs 4D and S3–S8). The cell cycle, Hippo, P53, and Wnt signaling pathways exhibited higher mutation frequencies in Methylation-L (Figs 4D and S3–S8). The mutation frequencies of the Hippo and TGF-β pathways were high in Methylation-M (Figs 4D and S3–S8). Additionally, differences in somatic CNV among the CC patients in the three clusters were evaluated using GISTIC 2.0. Among the 294 CC patients, 282 patients possessed CNV data. CNV analysis demonstrated that amplifications of 8q24.21 [*MYC* (oncogenic gene in MYC pathway)], 9p24.1 [*JAK2* (oncogenic gene in RTK-RAS pathway)], 17q12 [*ERBB2* (oncogenic gene in RTK-RAS pathway)], and 17q25.1 [*RPS6KB1* (oncogenic gene in PI3K-AKT pathway)] as well as deletions of 11q25 [*CBL* (TSG in RTK-RAS pathway)], 10q23.31 [*PTEN* (TSG in PI3K-AKT pathway)], 4q34.1 [*FBXW7* (TSG in Notch pathway)], and 15q15.1 [*MGA* (TSG in MYC pathway)] were identified in Methylation-H (Figs 5A,B and S9). Moreover, amplifications of 11q22.1 [*YAP1* (oncogenic gene in Hippo pathway)], 7p11.2 [*EGFR* (oncogenic gene in PI3K-AKT pathway)], as well as deletions of 4q35.2 [*FAT1* (TSG in Hippo pathway)], 10q23.31 [*PTEN* (TSG in PI3K-AKT pathway)], 17q25.3 [*CSNK1D* (TSG in Hippo pathway)], and 3p24.1 [*TGFBR2* (TSG in TGF-β pathway)] were identified in Methylation-M (Figs 5C,D and S9). Finally, amplifications of 11q22.1 [*YAP1* (oncogenic gene in Hippo pathway)] and 19q12 [*CCNE1* (oncogenic gene in cell cycle pathway)] and deletions of 13q14.2 [*RB1* (TSG in cell cycle pathway)] and 1p36.11 [*HES2/3/4/5* (TSGs in Notch pathway)] were identified in Methylation-L (Figs 5E,F and S9).

3.5. Construction of the HPV-related methylation signature

Univariate Cox regression analysis identified 20 HPV-related methylation sites with prognostic significance in the training dataset ($P < 0.01$, Fig. S10a). After LASSO regression analysis, 11 of the 20 HPV-related methylation sites were selected (Fig. S10b). Subsequently, a stepwise multivariate Cox regression analysis was performed for

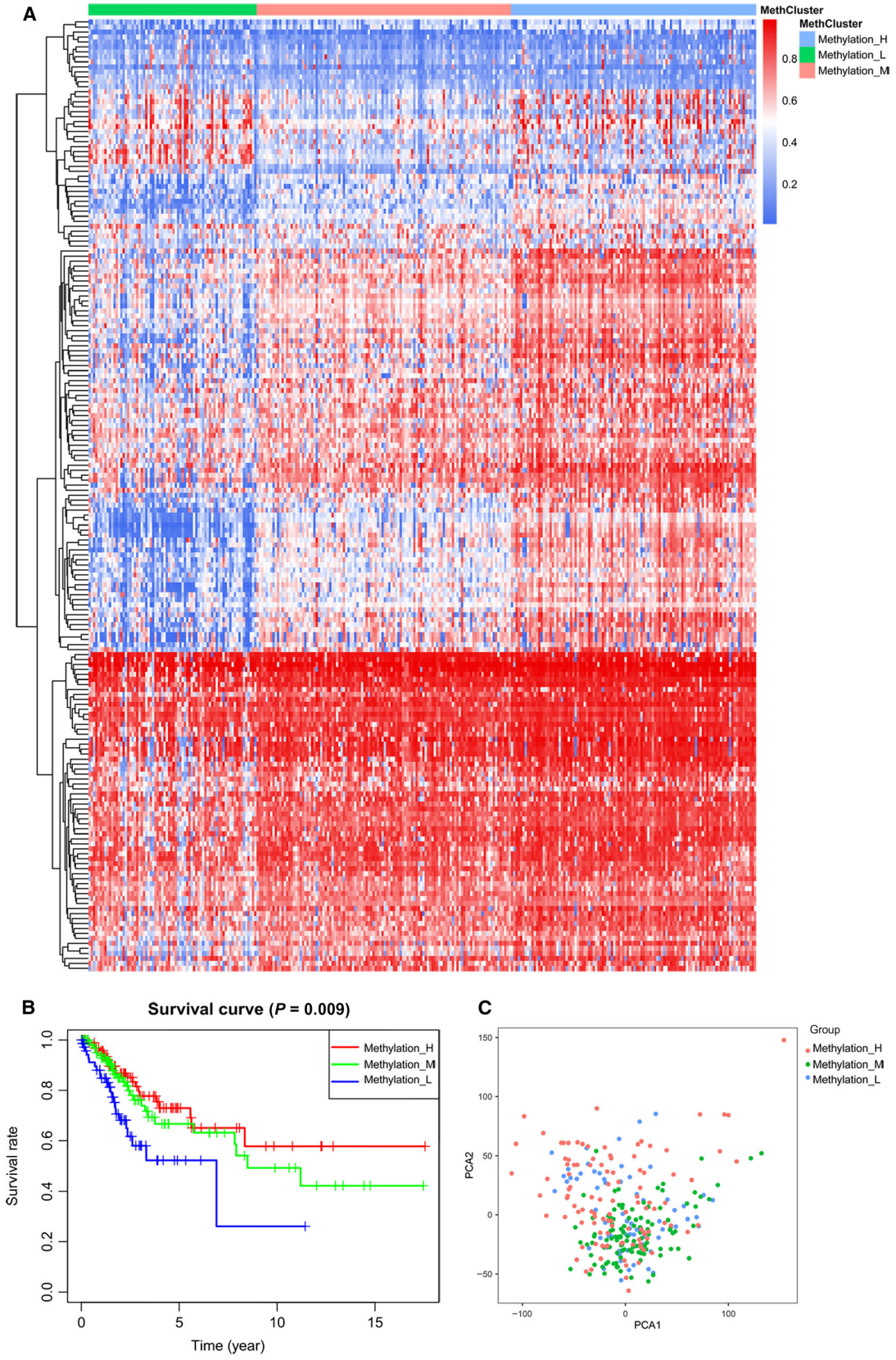

A
MethCluster
MethCluster
Methylation_H
Methylation_L
Methylation_M
0.8
0.6
0.4
0.2
B
Survival curve (P = 0.009)
Methylation_H
Methylation_M
Methylation_L
Survival rate
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
Time (year)
C
Group
Methylation_H
Methylation_M
Methylation_L
PCA2
150
100
50
0
−50
−100
0
100
PCA1

Fig. 2. Identification of three methylation clusters of CC with distinct survival outcomes. (A) Heatmap of three methylation clusters generated by performing unsupervised hierarchical clustering. (B) Kaplan–Meier curves of OS of three methylation clusters. (C) PCA of the total RNA expression profiles in the TGGA CC dataset.

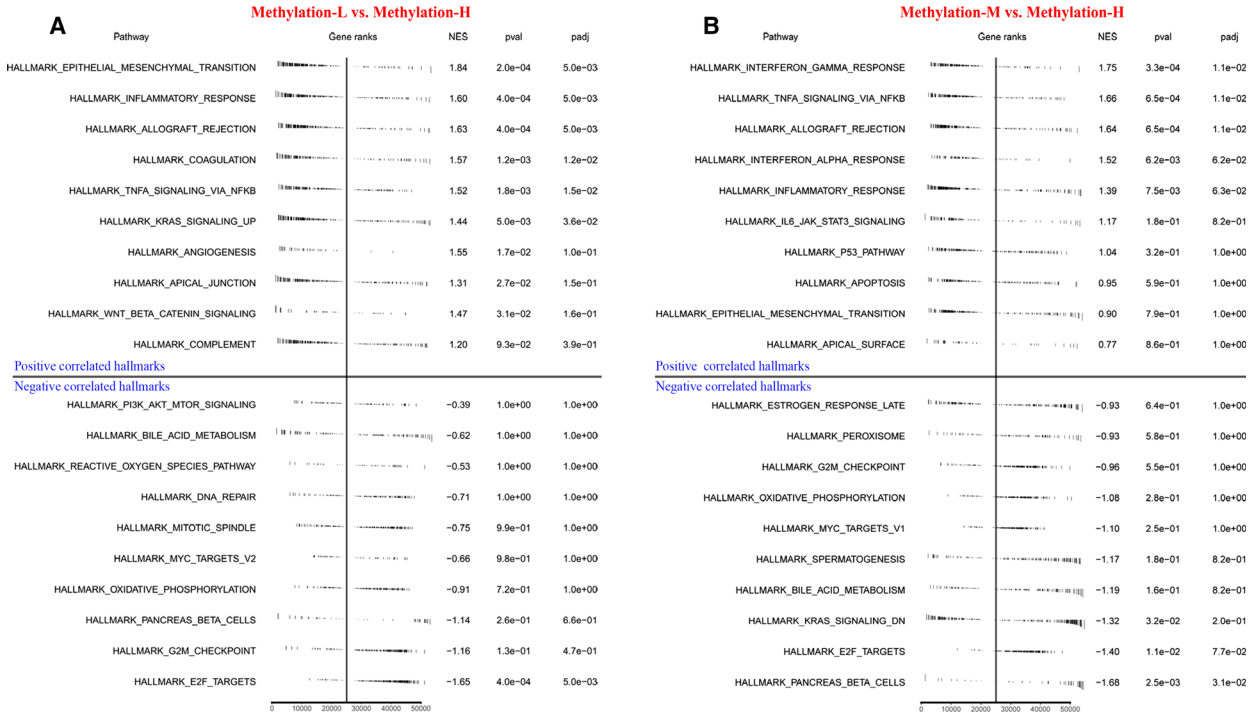


Fig. 3. Summary of GSEA results with hallmark gene sets in Methylation-L (A) or Methylation-M (B), compared to Methylation-H. Six positively enriched hallmarks and one negatively enriched hallmark with adjusted *P*-value < 0.05 were identified in Methylation-L. Three positively enriched hallmarks and one negatively enriched hallmark with adjusted *P*-value < 0.05 were identified in Methylation-M.

these 11 methylation sites. Finally, six methylation sites (cg23170347, cg16376000, cg13759702, cg01727408, cg05008070, and cg07227049) were identified to construct the optimal prognostic model (Fig. S10c). The risk score formula based on the DNA methylation levels and coefficients of these six HPV-related methylation sites was calculated as follows: Risk score = 1.9939 × beta-value of cg13759702 − 1.6941 × beta-value of cg23170347 − 1.5290 × beta-value of cg16376000 − 3.9910 × beta-value of cg01727408 − 2.4146 × beta-value of cg05008070 − 4.8805 × beta-value of cg07227049. The DNA methylation level of cg13759702 was correlated with high risk, while the DNA methylation levels of cg23170347, cg16376000, cg01727408, cg05008070, and cg07227049 were correlated with low risk. The genes corresponding to five methylation sites (cg05008070, cg07227049, cg13759702, cg16376000, and cg23170347) were Disheveled Binding Antagonist of Beta Catenin 1 (*DACT1*), VRK Serine/Threonine Kinase 2 (*VRK2*), Melanotransferrin (*MELTF*), Fibroblast Growth Factor 12 (*FGF12*), and Prickle Planar Cell Polarity Protein 2 (*PRICKLE2*); all of these genes were protein coding genes. The list of the six HPV-related methylation sites, their chromosomal locations, *P*-values, and the coefficients obtained in multivariate Cox regression analysis are provided in Table 1. Pearson's correlation test was conducted to measure the correlations between the expression of the above five protein coding genes and the methylation levels of the corresponding methylation sites (Fig. S11). Significant and negative correlations were observed between the expression of the five genes and the methylation level of corresponding methylation sites ($P < 0.001$, $R < -0.10$, Fig. S11).

3.6. Evaluating the predictive capability of the HPV-related methylation signature

Kaplan–Meier survival analyses were conducted in the training and testing datasets to evaluate the predictive capability of our HPV-related methylation signature.

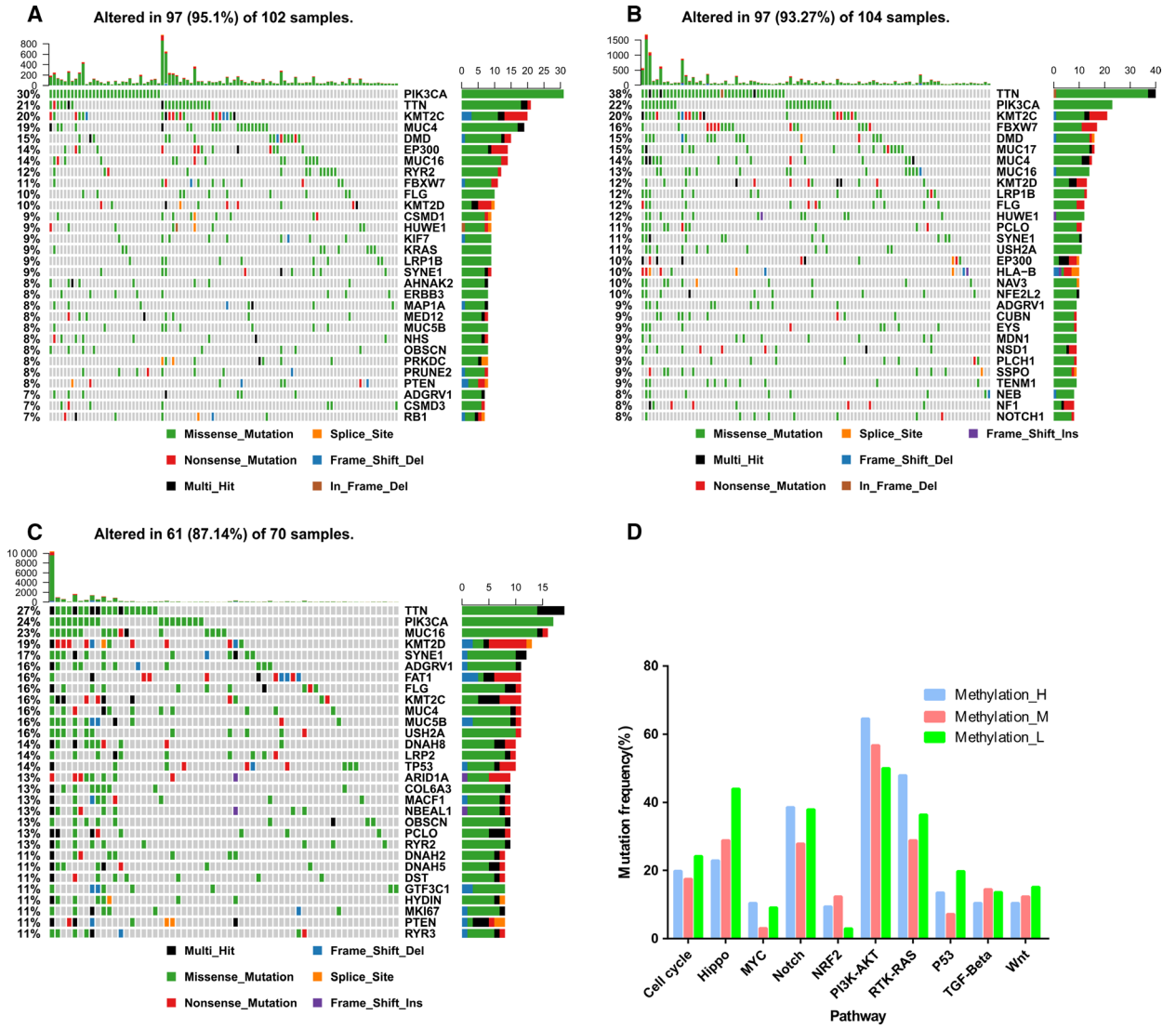


Fig. 4. Comparison of mutations among the three methylation clusters of CC. The top 30 most frequently mutated genes in the CC patients of Methylation-H (A), Methylation-M (B), and Methylation-L cluster (C). (D) The mutation frequencies of ten common oncogenic pathways among three clusters.

In the training dataset, CC patients were stratified into either high-risk ($n = 74$) or low-risk ($n = 74$) group. These two groups possessed distinct survival outcomes ($P < 0.0001$, Fig. 6A). Low-risk patients exhibited better OS compared to high-risk patients. The 5-year AUC for the six-DNA methylation signature was 0.899; the 3-year AUC was 0.888 (Fig. 6B). A similar result was observed in the testing dataset and the entire dataset. The risk score was calculated for patients in the testing dataset and the entire dataset, and then, each patient was marked as high risk or low risk, as previously described (Yang *et al.*, 2019). There were 73 high-risk patients and 73 low-risk patients within the testing dataset. The survival for low-risk patients was improved than that for the high-risk patients (Fig. 6C, $P < 0.0001$). The AUC at 5 years was 0.74, and the 3-year AUC was 0.728 in the testing dataset (Fig. 6D). There were 147 high-risk patients and 147 low-risk patients in the entire dataset. The low-risk patients exhibited longer median survival than the high-risk patients (2.50 vs. 1.48 years, $P < 0.0001$, Fig. 6E). In the entire dataset, the 5-year AUC was 0.813, and the 3-year AUC was 0.807 (Fig. 6F). The distribution of risk score, survival status, and heatmap of the six methylation sites for patients with CC in the training, testing, and entire datasets are displayed in

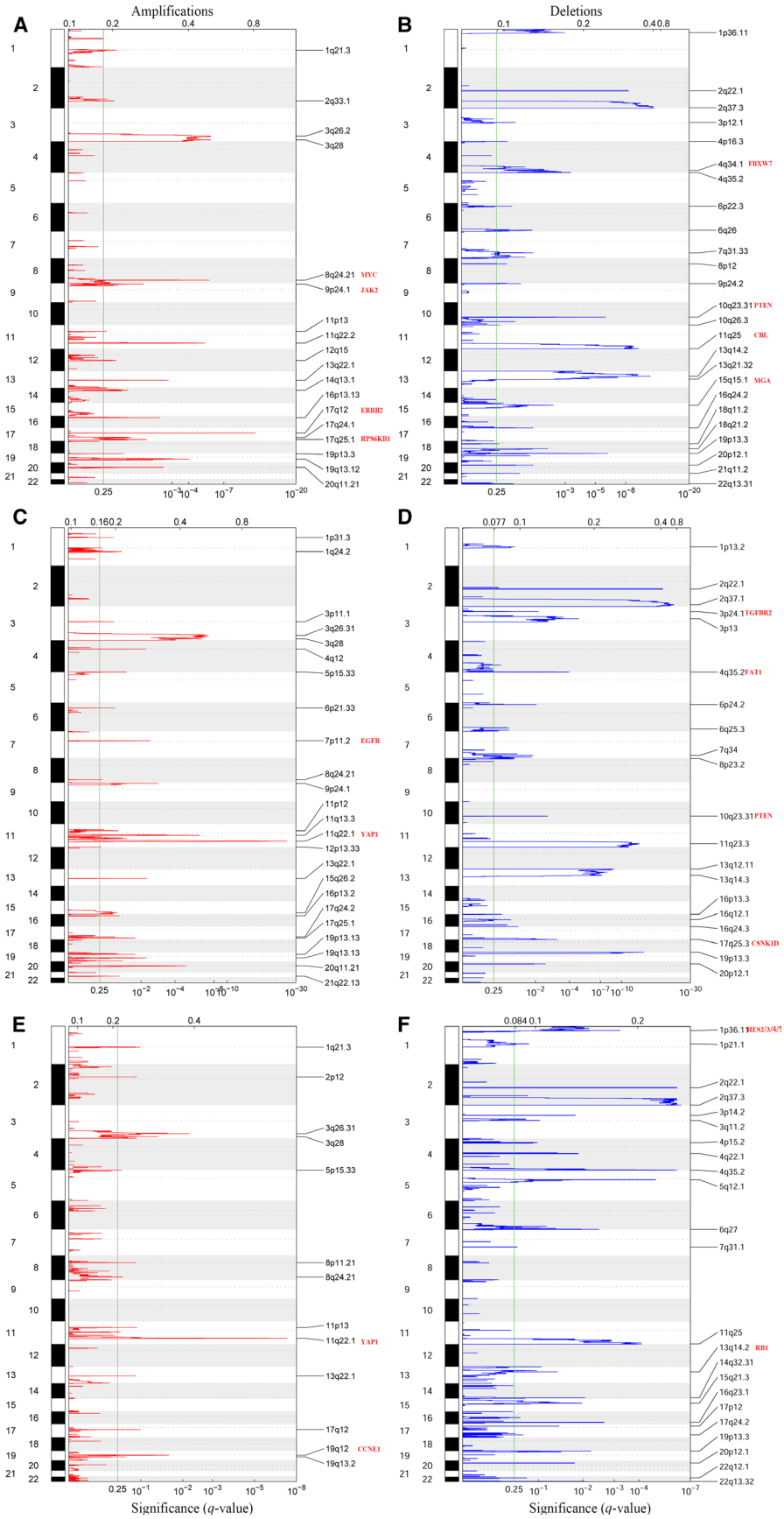


Fig. 5. GISTIC 2.0 amplifications and deletions in Methylation-H (A, B), Methylation-M (C, D), and Methylation-L (E, F) cluster. Chromosomal locations of peaks of significantly recurring focal amplifications (red) and deletions (blue) are displayed. The *q*-values, representing the statistical significance, are displayed along the bottom. Regions with *q*-values < 0.25 (green lines) were considered significantly altered. The locations of the peak regions of maximal copy number change and the known cancer-related genes within those peaks are indicated to the right of each panel.

Table 1. Six HPV-related methylation sites in the signature. NA, not available.

Probe ID	Chromosomal location	Gene symbol	Gene type	CGI coordinate	Feature type	Coefficient[a]	*P* value[a]
cg01727408	chr16: 85575465–85575466	NA	NA	chr16:85586333-85586656	NA	−3.9910	0.0007
cg05008070	chr14: 58639944–58639945	*DACT1*	Protein coding	chr14:58637581-58638859	S_Shore	−2.4146	0.0599
cg07227049	chr2: 58107873–58107874	*VRK2*	Protein coding	chr2:58046507-58047287	NA	−4.8805	< 0.0001
cg13759702	chr3: 197001599–197001600	*MELTF*	Protein coding	chr3:197002087-197004007	N_Shore	1.9939	0.0349
cg16376000	chr3: 192409541–192409542	*FGF12*	Protein coding	chr3:192408032-192410205	Island	−1.5290	0.0241
cg23170347	chr3: 64268119–64268120	*PRICKLE2*	Protein coding	chr3:64267857-64268143	Island	−1.6941	0.0958

[a]In multivariate Cox regression analysis.

Figs S12-S14. Furthermore, the DNA methylation levels of the six methylation sites in the high- and low-risk patients in the entire dataset were measured. We found that high-risk patients possessed significantly higher methylation levels for cg13759702 and significantly lower methylation levels for the other four methylation sites, with the exception of cg01727408, in the entire dataset (Fig. S15, $P < 0.001$).

3.7. Independence of the HPV-related methylation signature in the OS prediction from clinicopathological factors

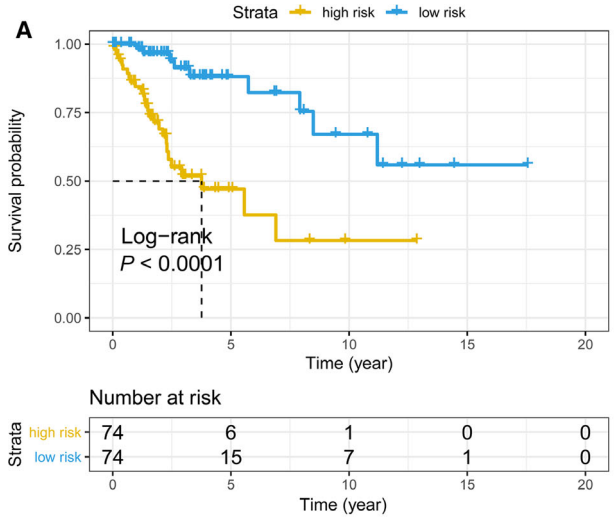
Univariate (Fig. S16a) and multivariate Cox regression analyses (Fig. S16b) were carried out to explore if the HPV-related methylation signature was an independent predictive indicator for the OS of CC patients. The results were adjusted for certain clinicopathological variables including age, grade, pathologic stage, clinical stage, and tumor status. The sample size was small after we excluded samples with unknown M stage ($n = 169$, >50%) and unknown HPV status ($n = 118$, 40.14%); therefore, M stage and HPV status were not included in the univariate and multivariable models. As a result, our signature could serve as an independent prognostic indicator within the entire dataset in the multivariate analysis (HR (95% CI) = 1.096(1.037–1.159), $P = 0.001$, Fig. S16b). To assess the independence of this HPV-related methylation signature, CC patients were reclassified according to different clinicopathological characteristics (Table 2). The results revealed that the signature was independent of age, clinical stage, histologic grade, T stage, lymph node metastasis, and tumor status; the signature was effective to stratify the prognosis of patients with CC.

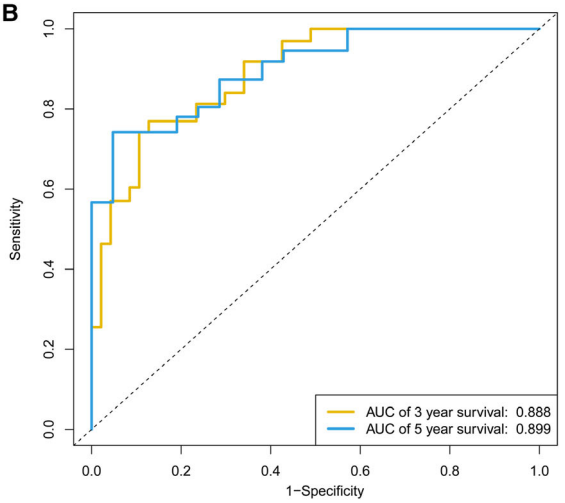
3.8. Immune landscape of cervical cancer patients within different subgroups

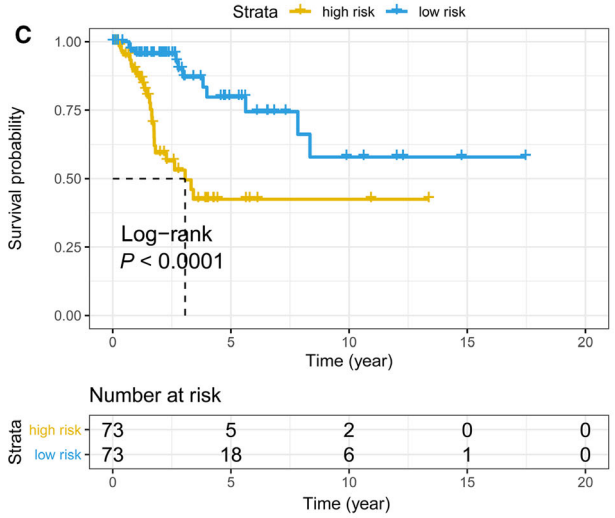
To compare the differences in the proportions of 24 immune cells between CC low- and high-risk patients and to explore the heterogeneity of immune infiltration in CC of the three methylation clusters, ssGSEA was conducted to estimate the relative proportion of the 24 immune cells in individual CC patient. The heatmap revealed the tumor-infiltrating immune-cell landscape of 294 CC patients from TCGA (Fig. 7).

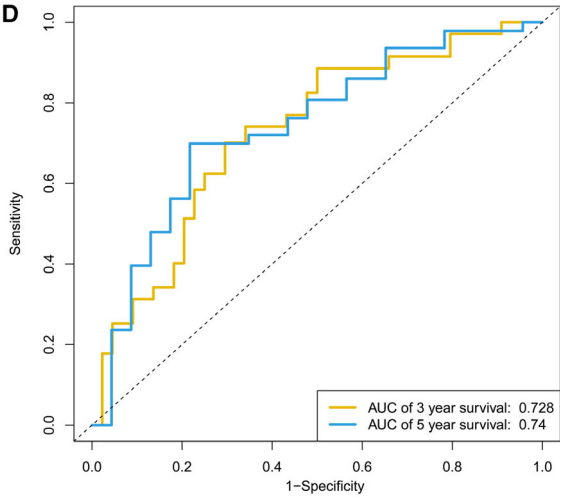
Among the 13 adaptive immune cell types, the low-risk group possessed significantly high proportions of B cells, CD8 T cells, cytotoxic cells, T cells, T helper cells, Tcm cells, Tfh cells, Th1 cells, and Treg cells ($P < 0.05$, Fig. S17a) compared to those of the high-risk group. Among the 11 innate immune cell types, the low-risk group possessed significantly higher proportions of DCs, aDCs, pDCs, iDCs, and neutrophils compared to those of the high-risk group ($P < 0.05$, Fig. S17b). Additionally, Methylation-H was characterized by relatively low infiltration of adaptive immune cells and innate immune cells, including CD8 T cells, cytotoxic cells, T cells, Tem cells, Tfh cells,

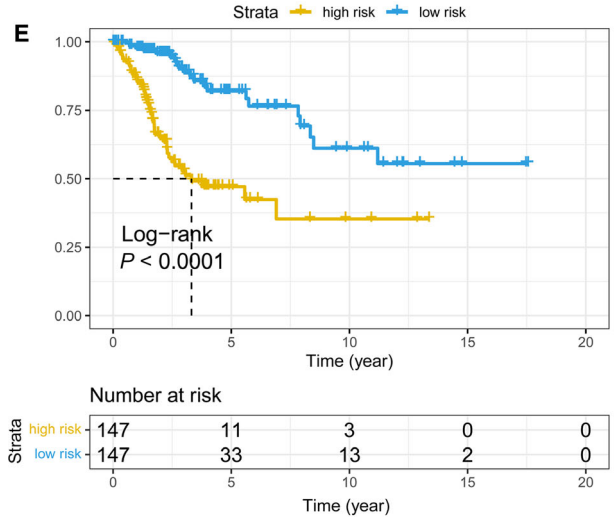
Fig. 6. The prognostic role of the HPV-related methylation signature in the training, testing, and entire datasets. Kaplan–Meier OS curves for patients assigned to high- and low-risk groups based on the risk score in the training (A), testing (C), and entire datasets (E) are shown. Time-dependent ROC curves in the training (B), testing (D), and entire datasets (F) are displayed.


A
Strata high risk low risk
Survival probability
1.00
0.75
0.50
0.25
0.00
Log−rank
P < 0.0001
0 5 10 15 20
Time (year)
Number at risk
Strata
high risk 74 6 1 0 0
low risk 74 15 7 1 0
0 5 10 15 20
Time (year)


B
Sensitivity
0.0 0.2 0.4 0.6 0.8 1.0
1−Specificity
AUC of 3 year survival: 0.888
AUC of 5 year survival: 0.899


C
Strata high risk low risk
Survival probability
1.00
0.75
0.50
0.25
0.00
Log−rank
P < 0.0001
0 5 10 15 20
Time (year)
Number at risk
Strata
high risk 73 5 2 0 0
low risk 73 18 6 1 0
0 5 10 15 20
Time (year)


D
Sensitivity
0.0 0.2 0.4 0.6 0.8 1.0
1−Specificity
AUC of 3 year survival: 0.728
AUC of 5 year survival: 0.74


E
Strata high risk low risk
1.00
0.75
0.50
0.25
0.00
Log−rank
P < 0.0001
0 5 10 15 20
Time (year)
Number at risk
Strata
high risk 147 11 3 0 0
low risk 147 33 13 2 0
0 5 10 15 20
Time (year)

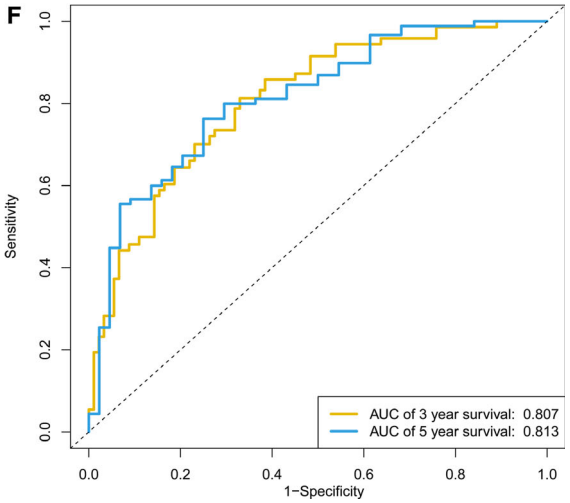

F
Sensitivity
0.0 0.2 0.4 0.6 0.8 1.0
1−Specificity
AUC of 3 year survival: 0.807
AUC of 5 year survival: 0.813

Table 2. Results of Kaplan–Meier and ROC analyses based on different subgroups. ROC, receiver operating characteristic.

Variables	Group	Sample size	Kaplan–Meier, *P* value	3-year AUC	5-year AUC
Age(years)	≤ 50	182	< 0.0001	0.815	0.778
	> 50	112	< 0.0001	0.806	0.859
Clinical stage	I/II	225	< 0.0001	0.790	0.814
	III/IV	63	0.0007	0.851	0.831
Histologic grade	G1/2	148	< 0.0001	0.785	0.823
	G3/4	119	0.0023	0.838	0.824
T stage	T1	169	0.0012	0.800	0.760
	T2/3/4	104	< 0.0001	0.845	0.936
Lymph node metastasis	No	164	0.0002	0.814	0.773
	Yes	64	0.0180	0.795	0.789
Tumor status	Tumor free	199	0.0340	0.838	0.789
	With tumor	80	0.0004	0.776	0.903

Tgd cells, Th1 cells, Treg cells, DCs, aDCs, iDCs, neutrophils, and macrophages ($P < 0.05$, Fig. S18). Consistent with this, the analysis of tumor microenvironments demonstrated that Methylation-H exhibited a lower ESTIMATE score, immune score, stromal score, and a higher tumor purity compared to the other two clusters ($P < 0.05$, Fig. S19). These results indicated that compared to the other two clusters, Methylation-H possessed a different immune phenotype that was featured by less immune infiltration and lower immune activation.

4. Discussion

Accurate subtype identification, prognostic stratification, and characterization of the underlying mechanism are crucial for our understanding of cancers and for the guidance of treatment management and personalized therapies. DNA methylation profiling is a recent method that has been used to improve tumor classification, and this technique has already led to re-definitions and sub-classifications of various tumors such as glioblastoma (Noushmehr *et al.*, 2010), head and neck squamous carcinoma (Brennan *et al.*, 2017), adrenocortical carcinoma (Barreau *et al.*, 2013), and hepatocellular carcinoma (Li *et al.*, 2019). Furthermore, several prognostic signatures based on DNA methylation sites have been reported to stratify the prognosis of various cancers such as cutaneous melanoma (Guo *et al.*, 2019), ovarian cancer (Guo *et al.*, 2018), lung adenocarcinoma (Wang *et al.*, 2019a), and breast cancer (Tao *et al.*, 2019). These studies support that DNA methylation sites can serve as promising biomarkers for subtype identification and prognostic stratification in cancer. HPV infection is a key oncogenic driver in CC. HPV infection causes epigenetic reprogramming of the host cell during malignant transformation, subsequently resulting in distinct HPV-related epigenetic phenotypes. Therefore, this study was intended to determine the epigenetic alterations involved in CC progression and HPV infection. Additionally, HPV-related DNA methylation signatures were explored to identify the different CC subtypes in patients and to stratify the prognosis of CC.

4.1. Subtype identification

Based on the HPV-related methylation sites, CC patients were classified into three heterogeneous clusters. Compared to Methylation-M and Methylation-L, Methylation-H exhibited a significantly improved OS. The hallmarks of tumors, including KRAS signaling, TNFα signaling via NF-κB, inflammatory response, epithelial–mesenchymal transition, and interferon-gamma response, were enriched in Methylation-M and Methylation-L. A great deal of research work has suggested that these biological processes or pathways play a significant role in tumorigenesis and the progression of CC (Kang *et al.*, 2007; Kloth *et al.*, 2005; Lages *et al.*, 2011; Lee *et al.*, 2008). We reasoned that the HPV-related methylation signature might take an important part in CC via the above biological processes or pathways. Based on mutation and CNV analyses, we found that mutations among the MYC, Notch, PI3K-AKT, and RTK-RAS pathways were most frequently detected in Methylation-H. Concurrently, the amplifications of oncogenes, such as *JAK2* and *ERBB2* in the RTK-RAS pathway and *RPS6KB1* in the PI3K-AKT pathway, and the deletions of TSGs, such as *CBL* in the RTK-RAS pathway and *PTEN* in the PI3K-AKT pathway, were identified in Methylation-H. Therefore, we speculated that hyper-activated RTK-RAS or PI3K-AKT pathways in tumor may take an important part in Methylation-H. The mutation frequencies of the Hippo and TGF-β pathways were high in Methylation-M. The amplifications of oncogenic genes, such as *YAP1* within the Hippo pathway and *EGFR* within the PI3K-AKT pathway, and the deletions of TSGs, such as *FAT1* and *CSNK1D* within the Hippo pathway, *TGFBR2* within the TGF-β pathway, and *PTEN* within the PI3K-AKT pathway, were identified in Methylation-M. Moreover, the cell cycle and Hippo signaling pathways exhibited higher mutation frequencies in Methylation-L. The amplifications of oncogenic genes, such as *YAP1* within the Hippo

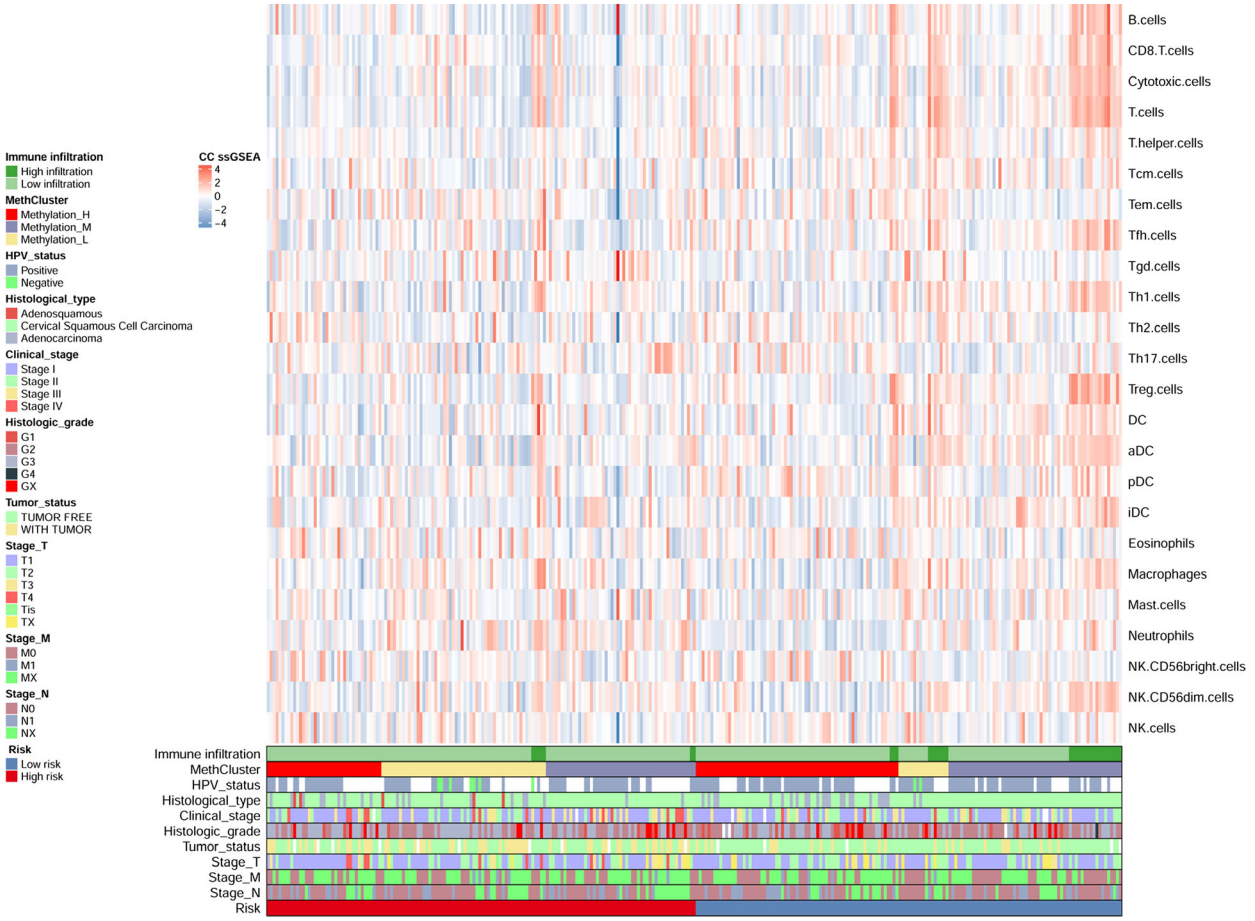


Fig. 7. Immune landscape of CC patients within different subgroups. The heatmap showed single sample GSEA scores from 24 immune cell types of 294 patients from TCGA. HPV status, clinical stage, tumor status, histologic grade, T/N/M stage, histological type, methylation cluster, and risk were annotated in the lower panel. Hierarchical clustering was performed with Euclidean distance and Ward linkage. Two distinct immune infiltration clusters, here termed high infiltration and low infiltration, were defined.

pathway and *CCNE1* within the cell cycle pathway, and the deletions of TSGs, such as *RB1* within the cell cycle pathway, were identified in Methylation-L. Therefore, we speculated that the Hippo, TGF-β, and cell cycle pathways might be responsible for the poor outcome observed in Methylation-M and Methylation-L. Overall, our analyses revealed that certain biological processes, pathways, and genomic alterations may result in a worse OS in Methylation-L and Methylation-M. Future studies are required to elucidate the role of these biological processes, pathways, and genomic alterations in HPV-related epigenetic phenotypes that specifically drive cancer development.

4.2. Risk stratification

We constructed and verified a prognostic risk signature using six HPV-related methylation sites (cg01727408, cg05008070, cg07227049, cg13759702, cg16376000, and cg23170347) that stratified CC patients into high- and low-risk groups. The genes corresponding to the five methylation sites (cg05008070, cg07227049, cg13759702, cg16376000, and cg23170347) were *DACT1*, *VRK2*, *MELTF*, *FGF12*, and *PRICKLE2*, which were all protein coding genes. *DACT1*, *VRK2*, *MELTF*, and *FGF12* have been implicated in cancers other than CC (Dmitriev *et al.*, 2015; Fernandez *et al.*, 2010; Guo *et al.*, 2017), and *PRICKLE2* has been reported to correlate with CC (Senchenko *et al.*, 2013). Future studies will be required to elucidate the functional impact of aberrant methylation of these five genes in CC development. The corresponding encoded proteins affected by aberrant methylation may also represent promising drug targets for cancer therapy.

Our HPV-related methylation signature could still act as an independent prognostic predictor, after

adjusting for certain clinicopathological variables. Subgroup analyses further highlighted that the signature possessed strong and independent predictive power when CC patients were regrouped according to different clinicopathological characteristics. Additionally, this signature possessed higher predictive performance for patients in advanced T and clinical stages (3-year AUCs were 0.845 and 0.851, respectively). Based on this, combining this signature with other clinical factors could serve as a promising tool for the prognosis of CC patients.

Finally, GSEA further revealed the connection between the signature and immune systems. Therefore, ssGSEA was conducted to estimate the relative proportion of the 24 immune cells in individual patient with CC. We aimed to compare the differences in the proportions of 24 immune cells between low-risk and high-risk patients with CC and to explore the heterogeneity of immune infiltration in CC within the three methylation clusters. Consequently, the low-risk group possessed a significantly higher proportion of immune cells that were involved in adaptive and innate immune responses compared to that of the high-risk group. Contrary to expectations, Methylation-H was characterized by relatively low infiltration of adaptive immune cells and innate immune cells.

4.3. Strength and limitations

To our knowledge, this is the first study to explore HPV-related DNA methylation signatures to identify the different subtypes of CC and to stratify the prognosis of CC. The molecular differences between the identified subtypes may allow these subtypes to be targeted separately under specific therapeutic approaches. In terms of clinical utility, a novel risk signature based on six HPV-related methylation sites was identified and verified. This signature can be tested as a prognostic tool to determine patients at high risk with the potential for multimodal therapy.

This study has a few limitations. Firstly, the sample size containing information with HPV status and HPV subtypes was relatively small, and we were unable to explore the association between the HPV-related methylation signature and HPV subtypes. Secondly, an ideal prognostic signature is one that can also efficiently risk-stratify in other independent datasets, and we could not yet locate another dataset to further validate the performance of our six-DNA methylation signature. Lastly, the TCGA dataset enrolled for analysis was primarily collected from patients with CC in Western countries and lacked data from Asian countries.

5. Conclusions

In conclusion, our study revealed that CC patients could be classified into three heterogeneous clusters based on the HPV-related methylation sites. Specific biological processes, pathways, and genomic alterations could be correlated with the different outcomes in the three clusters. Additionally, we derived a prognostic risk signature using six HPV-related methylation sites that stratified the patients with CC into high- and low-risk groups. This study provides new insight into epigenetic biomarkers that could help to improve subtype identification, risk stratification, and treatment management.

Acknowledgements

We thank the other members of our research team for their assistance.

Conflict of interest

The authors declare no conflict of interest.

Data accessibility

TCGA DNA methylation data from 309 CC samples and 3 adjacent normal samples based on the Illumina HumanMethylation 450 (450K) platform (Illumina Inc.) were downloaded from UCSC xena (https://xena.ucsc.edu/). Clinical information for 307 CC patients was obtained from UCSC xena. The RNA-sequencing gene expression profiles for 306 CC patients, somatic mutation profiles for 289 CC patients, and CNV profiles for 297 CC patients were downloaded from TCGA data portal (https://portal.gdc.cancer.gov/). The hallmark gene sets (h.all.v7.0.symbols) were obtained from the Molecular Signatures Database (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp).

Author contributions

SY and YW collected and analyzed the data, and wrote the manuscript. SQW, PX, and YJD analyzed the data and reviewed the manuscript. MW, KL, and TT participated in analyzing the data. YYZ, NL, and LHZ participated in preparation of the figures and tables. ZJD and HFK designed the study and revised the manuscript. All the authors read and approved the final manuscript.

Ethics approval

Not applicable.

Consent for publication

Not applicable.

References

Au Yeung CL, Tsang WP, Tsang TY, Co NN, Yau PL and Kwok TT (2010) HPV-16 E6 upregulation of DNMT1 through repression of tumor suppressor p53. *Oncol Rep* **24**, 1599–1604.

Barreau O, Assie G, Wilmot-Roussel H, Ragazzon B, Baudry C, Perlemoine K, Rene-Corail F, Bertagna X, Dousset B, Hamzaoui N *et al.* (2013) Identification of a CpG island methylator phenotype in adrenocortical carcinomas. *J Clin Endocrinol Metab* **98**, E174–E184.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295.

Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, Angell H, Fredriksen T, Lafontaine L, Berger A *et al.* (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394–424.

Brennan K, Koenig JL, Gentles AJ, Sunwoo JB and Gevaert O (2017) Identification of an atypical etiological head and neck squamous carcinoma subtype featuring the CpG island methylator phenotype. *EBioMedicine* **17**, 223–236.

Burgers WA, Blanchon L, Pradhan S, de Launoit Y, Kouzarides T and Fuks F (2007) Viral oncoproteins target the DNA methyltransferases. *Oncogene* **26**, 1650–1655.

Cancer Genome Atlas Research Network (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384.

Coussens LM and Werb Z (2002) Inflammation and cancer. *Nature* **420**, 860–867.

Crosbie EJ, Einstein MH, Franceschi S and Kitchener HC (2013) Human papillomavirus and cervical cancer. *Lancet* **382**, 889–899.

Das PM and Singal R (2004) DNA methylation and cancer. *J Clin Oncol* **22**, 4632–4642.

Dmitriev AA, Rosenberg EE, Krasnov GS, Gerashchenko GV, Gordiyuk VV, Pavlova TV, Kudryavtseva AV, Beniaminov AD, Belova AA, Bondarenko YN *et al.* (2015) Identification of novel epigenetic markers of prostate cancer by NotI-microarray analysis. *Dis Markers* **2015**, 1–13.

Egger G, Liang G, Aparicio A and Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429**, 457–463.

Fernandez IF, Blanco S, Lozano J and Lazo PA (2010) VRK2 inhibits mitogen-activated protein kinase signaling and inversely correlates with ErbB2 in human breast cancer. *Mol Cell Biol* **30**, 4687–4697.

Friedman J, Hastie T and Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**, 1–22.

Gui J and Li H (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001–3008.

Guo M, Akiyama Y, House MG, Hooker CM, Heath E, Gabrielson E, Yang SC, Han Y, Baylin SB, Herman JG *et al.* (2004) Hypermethylation of the GATA genes in lung cancer. *Clin Cancer Res* **10**, 7917–7924.

Guo W, Zhu L, Yu M, Zhu R, Chen Q and Wang Q (2018) A five-DNA methylation signature act as a novel prognostic biomarker in patients with ovarian serous cystadenocarcinoma. *Clin Epigenetics* **10**, 142.

Guo W, Zhu L, Zhu R, Chen Q, Wang Q and Chen JQ (2019) A four-DNA methylation biomarker is a superior predictor of survival of patients with cutaneous melanoma. *Elife* **8**, e44310.

Guo YL, Shan BE, Guo W, Dong ZM, Zhou Z, Shen SP, Guo X, Liang J and Kuang G (2017) Aberrant methylation of DACT1 and DACT2 are associated with tumor progression and poor prognosis in esophageal squamous cell carcinoma. *J Biomed Sci* **24**, 6.

Halvorsen AR, Ragle Aure M, Ojlert AK, Brustugun OT, Solberg S, Nebdal D and Helland A (2019) Identification of microRNAs involved in pathways which characterize the expression subtypes of NSCLC. *Mol Oncol* **13**, 2604–2615.

Holleczek B and Brenner H (2013) Model based period analysis of absolute and relative survival with R: data preparation, model fitting and derivation of survival estimates. *Comput Methods Programs Biomed* **110**, 192–202.

Kang S, Kim HS, Seo SS, Park SY, Sidransky D and Dong SM (2007) Inverse correlation between RASSF1A hypermethylation, KRAS and BRAF mutations in cervical adenocarcinoma. *Gynecol Oncol* **105**, 662–666.

Keeley B, Stark A, Pisanic TR 2nd, Kwak R, Zhang Y, Wrangle J, Baylin S, Herman J, Ahuja N, Brock MV *et al.* (2013) Extraction and processing of circulating DNA from large sample volumes using methylation on beads for the detection of rare epigenetic events. *Clin Chim Acta* **425**, 169–175.

Kloth JN, Fleuren GJ, Oosting J, de Menezes RX, Eilers PHC, Kenter GG and Gorter A (2005) Substantial

changes in gene expression of Wnt, MAPK and TNFalpha pathways induced by TGF-beta1 in cervical cancer cell lines. *Carcinogenesis* **26**, 1493–1502.

Lages EL, Belo AV, Andrade SP, Rocha MA, de Freitas GF, Lamaita RM, Traiman P and Silva-Filho AL (2011) Analysis of systemic inflammatory response in the carcinogenic process of uterine cervical neoplasia. *Biomed Pharmacother* **65**, 496–499.

Laurson J, Khan S, Chung R, Cross K and Raj K (2010) Epigenetic repression of E-cadherin by human papillomavirus 16 E7 protein. *Carcinogenesis* **31**, 918–926.

Lee MY, Chou CY, Tang MJ and Shen MR (2008) Epithelial-mesenchymal transition in cervical cancer: correlation with tumor progression, epidermal growth factor receptor overexpression, and snail up-regulation. *Clin Cancer Res* **14**, 4743–4750.

Li G, Xu W, Zhang L, Liu T, Jin G, Song J, Wu J, Wang Y, Chen W, Zhang C *et al.* (2019) Development and validation of a CIMP-associated prognostic model for hepatocellular carcinoma. *EBioMedicine* **47**, 128–141.

Li W, Wang W, Si M, Han L, Gao Q, Luo A, Li Y, Lu Y, Wang S and Ma D (2008) The physical state of HPV16 infection and its clinical significance in cancer precursor lesion and cervical carcinoma. *J Cancer Res Clin Oncol* **134**, 1355–1361.

Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP and Tamayo P (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425.

Lorent M, Giral M and Foucher Y (2014) Net time-dependent ROC curves: a solution for evaluating the accuracy of a marker to predict disease-related mortality. *Stat Med* **33**, 2379–2389.

Mayakonda A, Lin DC, Assenov Y, Plass C and Koeffler HP (2018) Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747–1756.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R and Getz G (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41.

Moody CA and Laimins LA (2010) Human papillomavirus oncoproteins: pathways to transformation. *Nat Rev Cancer* **10**, 550–560.

Munoz N, Castellsague X, de Gonzalez AB and Gissmann L (2006) Chapter 1: HPV in the etiology of human cancer. *Vaccine* **24**(Suppl 3), S3/1–S10.

Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP *et al.* (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522.

Roh MR, Gupta S, Park KH, Chung KY, Lauss M, Flaherty KT, Jonsson G, Rha SY and Tsao H (2016) Promoter methylation of PTEN is a significant prognostic factor in melanoma survival. *J Invest Dermatol* **136**, 1002–1011.

Rooney MS, Shukla SA, Wu CJ, Getz G and Hacohen N (2015) Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61.

Schmitz M, Driesch C, Jansen L, Runnebaum IB and Durst M (2012) Non-random integration of the HPV genome in cervical cancer. *PLoS One* **7**, e39632.

Senchenko VN, Kisseljova NP, Ivanova TA, Dmitriev AA, Krasnov GS, Kudryavtseva AV, Panasenko GV, Tsitrin EB, Lerman MI, Kisseljov FL *et al.* (2013) Novel tumor suppressor candidates on chromosome 3 revealed by NotI-microarrays in cervical cancer. *Epigenetics* **8**, 409–420.

Sergushichev AA (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv* 060012. "[PREPRINT]"

Tao C, Luo R, Song J, Zhang W and Ran L (2019) A seven-DNA methylation signature as a novel prognostic biomarker in breast cancer. *J Cell Biochem* **121**, 2385–2393.

Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A and Teschendorff AE (2017) ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33**, 3982–3984.

Wang Y, Deng H, Xin S, Zhang K, Shi R and Bao X (2019a) Prognostic and predictive value of three DNA methylation signatures in lung adenocarcinoma. *Front Genet* **10**, 349.

Wang H, Lengerich BJ, Aragam B and Xing EP (2019b) Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics* **35**, 1181–1187.

Whiteside MA, Siegel EM and Unger ER (2008) Human papillomavirus and molecular considerations for cancer risk. *Cancer* **113**, 2981–2994.

Yanatatsaneejit P, Mutirangura A and Kitkumthorn N (2011) Human papillomavirus's physical state and cyclin A1 promoter methylation in cervical cancer. *Int J Gynecol Cancer* **21**, 902–906.

Yang S, Wu Y, Deng Y, Zhou L, Yang P, Zheng Y, Zhang D, Zhai Z, Li N, Hao Q *et al.* (2019) Identification of a prognostic immune signature for cervical cancer to predict survival and response to immune checkpoint inhibitors. *Oncoimmunology* **8**, e1659094.

Yim EK and Park JS (2005) The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. *Cancer Res Treat* **37**, 319–324.

Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA *et al.* (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612.

Zhang L, Zhao Y, Dai Y, Cheng JN, Gong Z, Feng Y, Sun C, Jia Q and Zhu B (2018) Immune landscape of colorectal cancer tumor microenvironment from different primary tumor location. *Front Immunol* **9**, 1578.

Zhou F, Tao G, Chen X, Xie W, Liu M and Cao X (2014) Methylation of OPCML promoter in ovarian cancer tissues predicts poor patient survival. *Clin Chem Lab Med* **52**, 735–742.

Ziegert C, Wentzensen N, Vinokurova S, Kisseljov F, Einenkel J, Hoeckel M and von Knebel Doeberitz M (2003) A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. *Oncogene* **22**, 3977–3984.

Supporting information

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

Fig. S1. Venn diagram for the intersections between 48 190 DMPs (HPV-positive vs. HPV-negative) and 35 678 DMPs (tumor vs. normal). vs.: versus; DMPs: differentially methylated probes.

Fig. S2. Unsupervised clustering analysis of 9249 HPV-related methylation sites.

Fig. S3. Comparison of mutations among the three methylation clusters of cervical cancer. The mutation frequencies of the cell cycle (a), MYC (b), NRF2 (c), TGF-β (d), and P53 (e) signaling pathways among the three clusters are shown.

Fig. S4. The mutation frequencies of the Hippo signaling pathway among the three clusters are shown.

Fig. S5. The mutation frequencies of the Notch signaling pathway among the three clusters are shown.

Fig. S6. The mutation frequencies of the RTK-RAS signaling pathway among the three clusters are shown.

Fig. S7. The mutation frequencies of the PI3K-AKT signaling pathway among the three clusters are shown.

Fig. S8. The mutation frequencies of the Wnt signaling pathway among the three clusters are shown.

Fig. S9. Comparison of copy number variations among three methylation clusters of cervical cancer. (a) Copy number gistic score for Methylation-H, Methylation-M, and Methylation-L cluster. (b) Copy number frequency for Methylation-H, Methylation-M, and Methylation-L cluster. Copy number gistic score/-copy number frequency is indicated on the y-axis and chromosome on the x-axis. Individual chromosomes are separated by dotted lines with 'red' indicating copy number gain and 'blue' indicating copy number loss.

Fig. S10. The process of developing a prognostic signature containing six HPV-related methylation sites. The hazard ratios (HR), 95% confidence intervals (CI) calculated by univariate Cox regression (a), the results of LASSO regression (b), and the coefficients calculated by multivariate Cox regression analysis (c) are shown.

Fig. S11. Association between the expression of five genes and the methylation levels of the corresponding methylation sites. Level of gene expression is reported as $\log_2$-transformed FPKM, and the methylation levels of methylation sites were beta-values.

Fig. S12. The distribution of risk score, survival status, and the heatmap of six methylation sites for cervical cancer patients in the training dataset.

Fig. S13. The distribution of risk score, survival status, and the heatmap of six methylation sites for cervical cancer patients in the testing dataset.

Fig. S14. The distribution of risk score, survival status, and the heatmap of six methylation sites for cervical cancer patients in the entire dataset.

Fig. S15. Boxplots of beta-value in samples of patients in high- and low-risk groups in the entire dataset. Mann–Whitney U test was used to determine the differences between the two groups.

Fig. S16. Univariate and multivariate Cox regression analyses of the association between clinicopathological factors, risk score and overall survival of patients in the TCGA cervical cancer dataset.

Fig. S17. The relative abundance of the 24 immune cells types in cervical cancer high-risk and low-risk groups. A green violin represents the low-risk group. A red violin represents the high-risk group. The white points inside the violin represent median values. Wilcoxon test was implemented to evaluate the differences in infiltration levels of the 24 immune cell types between the two groups.

Fig. S18. The relative abundance of the 24 immune cells among the three clusters. Kruskal–Wallis test was used to determine the differences between the three clusters. ns no significance, $*P < 0.05$, $**P < 0.01$, and $*** P < 0.001$.

Fig. S19. The analysis of tumor microenvironments among the three clusters. (a) Distribution of ESTIMATE scores of three clusters. (b) Distribution of immune scores of three clusters. (c) Distribution of stromal scores of three clusters. (d) Distribution of tumor purity of three clusters. Kruskal–Wallis test was used to determine the differences between the three clusters. ns no significance, $**P < 0.01$, and $*** P < 0.001$.

Table S1. Clinical variables in the training and testing datasets.

Table S2. HPV infection status of 178 cervical cancer patients from TCGA.