An 8-DNA methylation signature predicts recurrence risk of cervical cancer

Jing-Hang Ma,^{a, b} Yu Huang,^a Lu-Yao Liu,^a and Zhen Feng^{a, *}

^a First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China.

^b Department of Gynecology, First Affiliated Hospital, Wenzhou Medical University, Wenzhou 325000, China.

* fengzhen@wmu.edu.cn

Abstract

Purpose: Cervical cancer is the leading cause of cancer-associated mortalities in female worldwide. DNA methylation have been demonstrated to have a regulatory role in the progression and could be novel biomarkers to predict the recurrence of the disease.

Methods: DNA methylation and RNA expression data of cervical cancer were downloaded from TCGA. DMGs and DEGs were screened and extracted by correlation analysis. The SVM-based recurrence prediction model was established using the selected DMGs. Cox regression analysis and ROC analysis were used as self-evaluations. The GEO database was applied for external validation. Functional enrichment was determined by GO and KEGG analysis.

Results: The eight-DNA methylation signature could identify patients with a high risk of recurrence (AUC=0.833). The SVM score was an independent risk factor for recurrence (HR = 0.418; 95% CI: 0.26-0.67, p-value <0.05). The independent GEO database further supported the

result.

Conclusion: The predictive function of the 8-DNA methylation signature for recurrence of cervical cancer was revealed in this study, which may help identify high-risk patients of recurrence and benefit the clinical treatment.

Introduction

Cervical cancer is the leading cause of cancer-related deaths in women, with more than 500,000 patients diagnosed and 260,000 deaths each year worldwide [1]. Surgery is the main treatment for early-stage cervical cancer patients. For those with a high risk of recurrence, radiotherapy alone or in combination with chemotherapy is also needed [2]. It has been reported that the recurrence rate of cervical cancer is 20%–30%, which increases with disease stage. Recurrent patients always have poor prognosis, and therapeutic methods are limited with low response rate[3–5].

At present, the staging system (FIGO stage or TNM stage) together with the “Sedlis Criteria” and other risk factors are generally used to identify cervical cancer patients with a high risk of recurrence according to clinical and pathological characteristics to guide adjuvant treatment [6, 7]. These assessments of recurrence risk undoubtedly provide predictive value; however, they are not as good at determining cancer recurrence in comparison with genetic, epigenetic and molecular biomarkers. Therefore, it is necessary to identify biomarkers for predicting recurrence risk and determining treatment strategies to improve patients’ prognoses [8, 9].

DNA methylation is responsible for regulating gene transcription without changing the DNA sequence. Abnormal hypermethylation of tumor suppressor genes and hypomethylation of oncogenes are vital during the early tumorigenesis process [10, 11] and related epigenetic drugs

have been proved to be effective in the treatment of cervical cancer[12]. Compared to the complex variability in RNA expression between individual patients [13], DNA methylation is a relatively stable and pervasive modification across tumor types that is associated with the pathogenesis and progression of cancer and related to the prognosis and outcome of patients [13]. Researchers have explored the relation between DNA methylation and cervical cancer, including the biological function of genes [14–16], methylation-based tumor classification [17, 18] and radiotherapy-related predictors [19]. However, few studies have investigated a recurrence prediction model of DNA methylation in cervical cancer, which is of clinical significance to help clinicians find high-risk patients and implement active treatment.

In this study, we used DNA methylation and related RNA expression datasets of cervical cancer patients from The Cancer Genome Atlas (TCGA) to explore the epigenetic signature for recurrence risk prediction.

Methods

Sample selection and data processing

DNA methylation datasets, RNA expression datasets and clinical information of cervical cancer patients used for modeling were downloaded from the TCGA (<https://portal.gdc.cancer.gov/>) [20] with the use of the FireBrowse Data Portal (<http://firebrowse.org/>) [21]. Illumina Infinium HumanMethylation450 BeadChip sequencing technology was used to obtain 312 DNA methylation profiles, and Illumina HiSeq2000 was used for 307 mRNA-seq profiles.

All samples in the TCGA were collected from primary tumors of cervical tissue, with tumor purity ranging from 22% to 96% [20]. We excluded samples with missing survival information and included samples according to the recurrence status and disease-free survival time. After

screening and filtration, 85 patients who had been followed-up for more than two years without recurrence were included in the no recurrence group, and 50 patients with evidence of recurrence were included in the recurrence group, as recurrence usually occurs in the first two years after completing treatment [22, 23].

Differential methylation analysis and differential expression analysis

DNA methylation datasets and RNA expression datasets from the TCGA were analyzed in R (version 3.6.0). DNA methylation levels (β values) with a p-value less than 0.01 and with an absolute log fold change larger than 0.025 were considered to be differentially methylated genes (DMGs). RNA expression levels with a p-value less than 0.01 and an absolute log fold change larger than 0.4 were considered to be differentially expressed genes (DEGs). The log fold change thresholds were set to identify genes proximately ranked in the top 1%. DMGs overlapping with DEGs were extracted for the correlation analysis. We used the *lm* function in R to fit the linear model. Those with a significant negative correlation between DMGs and DEGs were further chosen to establish the SVM-based prediction model.

SVM model

SVM is a machine learning algorithm that offers a direct approach to binary classification [26]. The SVM considers data as points in a high-dimensional space and finds an optimal hyperplane to separate data into classes.

The eight selected DMGs from 135 patients in the TCGA database were denoted as 135 pairs $(x_1, y_1), (x_2, y_2), \dots, (x_{135}, y_{135})$. $x_i \in \mathbb{R}^8$: eight-dimensional data represent methylation levels of eight DMGs in the i^{th} sample; $y_i \in \{-1, 1\}$: labels of classification. (x_i, y_i) were mapped into data

points in the eight-dimensional space and were used to train the SVM-based prediction model without cross validation. The model found the optimal linear hyperplane $P0\{x: \phi(x)=0\}$ in the feature space, where ϕ was the linear transformation. Points satisfying $\{x: \phi(x)>0\}$ or $\{x: \phi(x)<0\}$ belonged to the class $y = 1$ or $y = -1$, respectively (the graphical illustration was shown in Figure.3b).

We extracted the ϕ function along with its coefficients from the linear SVM model and generated the SVM score to predict the classification. The SVM score was calculated with the β values of the eight DNA methylation genes and their coefficients in the trained SVM model. A constant bias was also introduced. A sample was classified into the low-risk or high-risk class if the SVM score was higher or lower than the threshold 0. In this work, we used the `e1071` package in R [27], which contained the *svm* function.

Recurrent signature validation

The confusion matrix was constructed to further confirm the prediction of the SVM classifier model. The ROC curve and the area under the curve (AUC) were used to assess the fit of the SVM-based prediction model [28]. A T-test and Fisher's exact test were used to identify clinical features affecting recurrence. The terms with $p\text{-value}<0.05$ were retained for further multivariate Cox regression analysis.

Data visualization using t-SNE

The t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [29], a nonlinear dimensionality reduction algorithm, was used to keep the topological structure in the high dimension and maps each datum onto several different, but related, low-dimensional manifolds,

since an SVM plot can not visualize the eight-dimensional data of the eight DNA methylation genes.

We used the Rtsne package in R [30]. The parameter “perplexity”, balanced attention between local and global aspects of the data and was set to 30, which was an estimate of the approximately 30 closest neighbors each data point had. The number of iterations was 10,000.

External validation of the GEO database

We searched cervical cancer datasets containing both DNA methylation and survival information across the GEO database, and the only dataset GSE30759 (the Illumina Infinium 27k Human Methylation BeadChip for DNA methylation profiles) with 63 samples of primary cervical cancer tissue satisfied such conditions. We performed a similar preprocessing on them as that used on the TCGA datasets. A patient whose recurrence-free survival time was equal to or longer than two years was regarded as a nonrecurrent patient. A patient whose survival time was shorter than the recurrence-free survival time was considered to suffer from the recurrence of the cancer.

The recurrence-related methylation genes in the GEO dataset overlapped with the 8 reference genes obtained from the training model using the TCGA dataset. Three genes, *LOC100132215*, *TIGIT*, and *ZFYVE21*, in the GEO dataset were found to not have probes. We calculated the average DNA methylation levels of these three genes in the TCGA dataset and subtracted them from the SVM score threshold according to (eq. 2), eventually generating the reduced SVM score threshold.

We analyzed the distribution of the methylation levels for the other five genes in both datasets. We found four genes whose distributions in both datasets were similar. We then normalized the levels in the GEO dataset to those in the TCGA datasets to avoid bias. Normalization of the

methylation value of genes from the i^{th} sample in the GEO dataset is as follows:

$$\beta_{Gene,GEO}^i = \frac{\langle \beta_{Gene,TCGA} \rangle}{\langle \beta_{Gene,GEO} \rangle} \times \beta_{Gene,GEO}^i \#(\text{eq. 1})$$

Of the five genes, *MPHOSPH10* was excluded due to the large differences in distribution between two datasets. Similarly, we subtracted the average DNA methylation level of the gene *MPHOSPH10* in the TCGA dataset to further update the reduced SVM score threshold.

Functional enrichment analysis

Gene Ontology (GO) enrichment analysis [24] was performed to identify the function of DMGs, including biological process (BP), cellular component (CC) and molecular function (MF). Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed to identify KEGG pathways. The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used to perform GO enrichment analysis and KEGG enrichment analysis (<https://david.ncifcrf.gov/>) [25]. The enrichment functions and pathways were considered significant if the p-value was < 0.05 .

Results

The study procedures are shown in **Figure 1**.

Identification of DMGs and DEGs

To identify the DMGs and DEGs between the recurrence and no recurrence groups, we screened 24,278 DNA methylation genes and 20,104 RNA expression data from the TCGA database (**Figure 2a**). Volcano plots were employed to show the distribution of DNA methylation or gene

expression fold changes in each group (**Figure 2b, c**). A total of 305 DMGs (p-value <0.01; absolute log fold changes >0.025) and 262 DEGs (p-value <0.01; absolute log fold change value >0.4) were extracted. Eleven hypermethylated genes and 5 hypomethylated genes overlapped with the DEGs, and 8 of them (*ZFYVE21*, *C17orf62*, *PDIA6*, *CUL7*, *TIGIT*, *MPHOSPH10*, *LOC100132215* and *CD3E*) had negative correlations between DNA methylation and mRNA expression levels, with p-values <0.05, as shown in **Figure 2d, e**. These 8 DNA methylation genes were used to further establish the recurrence prediction model.

Figure 1 Flowchart of the whole analysis process.

Figure 2: Selection of DMGs in recurrent and nonrecurrent cervical cancer samples. (a) Venn diagram of overlapping DEGs and DMGs. (b) Volcano plot of DMGs. (c) Volcano plot of DEGs. (d) Coefficient distribution of the gene signatures. (e) The correlation between gene expression and DNA methylation level in cervical cancer.

The SVM-based prediction model of DNA methylation

The β values of the eight selected DNA methylation sites from the 135 patients in the TCGA dataset were visualized by a heat map using the pheatmap package in R (**Figure 3a**). We trained the SVM model according to the recurrence outcomes (**Figure 3b**). The confusion matrix is shown in **Figure 3c**, and the ROC curve (AUC: 0.833) is shown in **Figure 3d**. The SVM score was generated to predict the recurrence rate with the β values of the eight DNA methylation genes, and the formula is as follow:

Figure 3: Establishment of the prediction model. (a) Heat map of 8 methylation-related genes in cervical cancer patients. (b) Schematic of the SVM algorithm. (c) The confusion matrix of the prediction model. Positive: recurrence. (d) ROC curve of the prediction model, AUC=0.833.

SVM score =

$$0.48 \times \beta_{C17orf62} - 0.61 \times \beta_{CD3E} + 0.17 \times \beta_{CUL7} + 0.37 \times \beta_{LOC100132215} + 0.12 \times \beta_{MPHOSPH10} + 0.25 \times \beta_{PDIA6} - 0.49 \times \beta_{TIGIT} + 0.49 \times \beta_{ZFYVE21} - 0.74 \quad \#(eq. 2)$$

An SVM score larger than the threshold 0 is regarded as a sample with low risk, and an SVM score smaller than 0 is regarded as a sample with high risk.

Univariate analysis and multivariate analysis

Clinical features and SVM scores of these patients were analyzed by univariate analysis using a T-test and Fisher's exact test in R (Table 1). SVM score, FIGO stage, N stage, M stage, and pathologic grade were risk factors for recurrence (p-value < 0.05) in univariate analysis.

Multivariate Cox regression analysis was further conducted on these risk factors. SVM score

(hazard ratio (HR) = 0.42; 95% confidence interval (CI): 0.26-0.67, p-value <0.001) and N stage (HR = 3.31; 95% CI: 1.17-9.33, p-value <0.05) were independent risk factors for recurrence (Table 1).

Table 1: Stratification analysis for risk factors for recurrence in patients from the TCGA.

Clinical Characteristic	Mode	Univariate analysis			Cox regression analysis	
		No Recurrence (N=85)	Recurrence (N=50)	P-value	HR (95% CI)	P-value
Age, years ^ζ	mean±SD	48.88±12.37	51.16±13.99	3.31E-01		
SVM score ^ζ	mean±SD	1.32±1.14	-0.21±1.21	2.62E-11	0.418 (0.260-0.670)	3.02E-04
Clinical Characteristic	Group	No Recurrence (N=85)	Recurrence (N=50)	P-value	HR (95% CI)	P-value
FIGO Stage ^η	1	48	24	3.09E-02	0.973 (0.503-1.879)	9.34E-01
	2	23	11			
	3	11	9			
	4	2	5			
	NA	1	1			
Pathologic M Stage ^η	M0	39	11	2.15E-02	0.872 (0.079-9.639)	9.11E-01
	M1	1	4			
	MX	45	35			
Pathologic N Stage ^η	N0	46	10	2.50E-03	3.31 (1.171-9.639)	2.40E-02
	N1	12	13			
	NX	27	27			
Pathologic T Stage ^η	T1	45	16	5.55E-02		
	T2	23	9			
	T3	4	2			
	T4	1	4			
	TX	12	19			

	1	6	3	
Pathologic	2	46	16	
Grade ^a	3	26	28	3.67E-02
	NA	6	2	

ζ : T-test; η: Fisher's exact test.

NA: not available; SD: standard deviation; HR: hazard ratio; CI: confidence interval.

Visualizing the model using t-SNE

The results of our training model with t-SNE are shown in **Figure 4**. The eight-dimensional data were mapped onto a two-dimensional space by the dimensionality reduction method. Samples in each group are represented as a data point in red or blue. The results revealed that samples closer to the bottom left were more likely to be from patients in the recurrence group.

Figure 4: Visualization of the eight-dimensional variables with the t-SNE algorithm. Points in red represent patients with recurrence, and those in blue represent patients without recurrence.

External validation of the GEO database

An independent external GEO cohort (GSE30759) was applied to validate the SVM-based recurrence prediction model of DNA methylation. A flowchart representation of the overall procedure is shown in Figure 5a. Three genes without the probes in the GEO dataset were substituted by the corresponding mean values in the TCGA dataset (*LOC100132215*: 0.91; *TIGIT*: 0.91; and *ZFYVE21*: 0.98) to adjust the threshold of the reduced SVM score. As shown in Figure 5b, the methylation levels of the four genes (b1: *PDIA6*; b2: *CUL7*, b3: *CD3E*, and b4: *C17orf62*) with similar distributions in two datasets formed the reduced SVM score. The mean value of *MPHOSPH10* (b5: TCGA: 0.71±0.11; GEO: 0.02±0.01) in the TCGA dataset was subtracted to further update the reduced SVM score threshold.

As shown in Figure 5c, the initial threshold of the SVM score was 0 on the basis of 8 genes. The cost of missing genes is the range of uncertainty around the threshold, where a sample is at either high or low risk. To improve the identification of high-risk patients, we considered samples within the same region as those at high risk. Therefore, the reduced SVM score formula for the GEO dataset is as follows:

$$TH_{reduced}^{GEO} = 0.74 - 0.37 \times \langle \beta_{LOC100132215}^{TCGA,high} \rangle + 0.49 \times \langle \beta_{TIGIT}^{TCGA,high} \rangle - 0.49 \times \langle \beta_{ZFYVE21}^{TCGA,high} \rangle - 0.12 \times \langle \beta_{MPHOSPH10}^{TCGA,high} \rangle \#(eq. 3)$$

In the above expression, “high” stands for all samples at high risk, and $\langle \dots \rangle$ indicates the average. The reduced SVM score threshold for the four genes in the GEO dataset, $TH_{reduced}^{GEO}$, equals 0.304, instead of the initial threshold 0 for the eight genes in the TCGA dataset. The β values of the four genes (*C17orf62*, *CD3E*, *CUL7*, and *PDIA6*) from the GEO dataset then shifted the mean values to one in the TCGA dataset, avoiding bias.

The recurrence-free survival was significantly longer in the low-risk group ($TH_{reduced}^{GEO} > 0.304$)

than in the high-risk group ($TH_{reduced}^{GEO} \leq 0.304$) in the recurrence curve (p-value=0.039; Figure 5d). ROC curve analysis was also performed on the prediction result (Figure 5e), and the AUC was 0.773.

Figure 5: External validation of the GEO dataset. (a) Flowchart for the external validation. (b) The distribution of the methylation levels for the five genes in both the TCGA (upper) and GEO (lower) datasets. (c) Threshold shift in the procedure. (d) Survival analysis of the GEO dataset using the prediction model. (e) ROC curve of external validation with the AUC.

GO and KEGG analyses

To research the biological function of DMGs, DAVID was used to perform GO analysis for the MFs, BPs and CCs. The MFs (p-value<0.05; **Figure 6a**) were enriched in protein binding, poly(A) RNA binding, protein homodimerization activity, and GTPase activator activity. The BPs (p-value <0.05; **Figure 6b**) were enriched in the regulation of transcription, signal transduction, and positive regulation of GTPase activity immune response. The CCs (p-value <0.05; **Figure 6c**) were enriched in the plasma membrane, cytosol, and nucleoplasm membrane. The results of KEGG analysis (**Figure 6d**) suggested that the DMGs were enriched in human T-lymphotropic virus 1 (HTLV-1) infection, cytokine-cytokine receptor interaction, T-cell receptor signaling pathway, and natural killer cell mediated cytotoxicity.

Figure 6: The results of GO and KEGG pathway analyses of DMGs in cervical cancer. (a) The GO terms in the MF; (b) BP; (c) CC; (d) In the results of KEGG pathway analysis, the x-axis represents the rich factor; different colors represent the $-\log_{10}(\text{p-value})$ and the size of the point represents the number of genes.

Discussion

DNA methylation is a major epigenetic mechanism that plays important roles in various biological processes, such as the regulation of gene expression [31], cell differentiation [32], and inflammation [33]. DNA methylation generally occurs in CpG sites, which are unmethylated or hypomethylated in normal cells [34]. Hypermethylation of these CpG sites may silence tumor suppressor genes and lead to carcinogenesis [35]; therefore, the identification of these abnormally methylated genes contributes to the diagnosis, prediction of prognosis and selection of cancer treatment.

In this study, we performed a comprehensive screen for cervical cancer patients from the TCGA dataset by comparing patients with recurrence and without recurrence. We used the SVM machine learning algorithm along with the generated SVM scores to establish the prediction model. The SVM score has been proven to be an independent risk factor for recurrence by Cox regression analysis. In contrast, FIGO, T and M stage were not found to be independent risk factors for recurrence, which implied that the SVM score was more closely related to tumor recurrence than the traditional tumor staging system. Furthermore, the SVM score is a robust and reliable predictor for recurrence regardless of pathological type and is independent of N stage. The ROC curve showed that the model had a high degree of fit. Independent datasets in GEO further supported the model with survival analysis. In clinical applications, this model allows clinicians to screen patients at high risk of recurrence using eight DNA methylation biomarkers.

The eight DEGs, namely, *ZFYVE21*, *C17orf62*, *PDIA6*, *CUL7*, *CD3E*, *TIGIT*, *MPHOSPH10* and *LOC100132215*, were found to be differentially affected by DNA methylation between patients with and without recurrence. Some of these genes have been reported to be dysregulated in cancer or other diseases; *ZFYVE21* is located on chromosome 14 and is associated with the

metastasis of colorectal carcinoma [36] and malignancy of renal cell carcinoma [37]; *C17orf62* regulates the NADPH oxidase of phagocytes and contributes to the cause of chronic granulomatous disease [38]; *CD3E* is part of the group of immunoreceptors with tyrosine-based activation motif (ITAMs) in the T-cell receptor (TCR) signal-triggering module affecting overall survival (OS) of patients with several cancer types, including cervical cancer [39]; *TIGIT* is expressed on the surface of T and natural killer cells and is associated with the reduction in NK-cell-mediated cytotoxicity and the increase in regulatory T-cell suppression [39]; *CUL7* is an oncogenic gene that promotes cancer cell survival by promoting caspase-8 ubiquitination [39]; *LOC100132215* is correlated with gene expression in cancerous breast tissues [40]; *MPHOSPH10* is involved in ribosomal RNA (rRNA) processing during mitosis [41]; *PDIA* [40] promotes the proliferation and growth of various types of human cancer cells through activating the Wnt/ β -catenin signaling pathway [39]. The role of most of these genes in cervical cancer has not been revealed, and further research is needed to determine their biological functions and mechanisms.

Limitations

There were three main limitations in this study. First, to distinguish cervical cancer patients with and without recurrence, we excluded patients without recurrence who had a short follow-up time, which led to a decrease in the number of samples. As the follow-up time was not long enough, we included the patients who had no recurrence within two years into the no recurrence group, which may cause the deviation to some extent. Second, all the genes used for modeling methylation from the TCGA were not in the external validation datasets from GEO due to the different chips used; therefore, we had to assume the GEO dataset has the same distribution as

the TCGA dataset for the three genes *LOC100132215*, *TIGIT*, and *ZFYVE21*. In addition, the histogram (Figure 5b) shows a large difference in the *MPHOSPH10* gene between the TCGA and GEO datasets. We calculated the average β value in the TCGA dataset and used it in the validation set. Third, further experimental studies on these genes are still needed to clarify their functions and mechanisms in cervical cancer.

Conclusions

In this study, we identified eight recurrence-related DNA methylation biomarkers in cervical cancer through a comprehensive analysis and established a recurrence prediction model using machine learning. This work offers a preliminary but effective approach to predict cancer recurrence with a few biomarkers, which may help clinicians identify high-risk patients and implement active treatment.

Data Availability

All the analyses and preparation codes, along with the related data files, will be collected and attached in the repository <https://github.com/zfeng2019/8dna>. The TCGA Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (TCGA-CESC) data were downloaded on January 28, 2016, from the Broad Institute FireBrowse Data Portal (www.firebrowse.org). The GEO dataset is available at <http://www.ncbi.nlm.nih.gov/geo> with the accession number GSE30759.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Funding

The author(s) received no specific funding for this work.

Authors' contributions

J.-H. M. and Z. F. designed the study; J.-H. M., Y. H. and L.-Y. L. carried out the analysis; J.-H. M. and Z. F. wrote the manuscript. L.-Y. L. and Z. F. visualized the analysis. All authors have read and approved the final manuscript.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394–424.
2. Peters III WA, Liu PY, Barrett RJ, Stock RJ, Monk BJ, Berek JS, et al. Concurrent chemotherapy and pelvic radiation therapy compared with pelvic radiation therapy alone as adjuvant therapy after radical surgery in high-risk early-stage cancer of the cervix. *Obstet Gynecol Surv*. 2000;55:491–2.
3. Huang K, Sun H, Li X, Hu T, Yang R, Wang S, et al. Prognostic risk model development and prospective validation among patients with cervical cancer stage IB2 to IIB submitted to neoadjuvant chemotherapy. *Sci Rep*. 2016;6:27568.
4. Obrzut B, Kusy M, Semczuk A, Obrzut M, Kluska J. Prediction of 5--year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods. *BMC Cancer*. 2017;17:840.
5. McLachlan J, Boussios S, Okines A, Glaessgen D, Bodlar S, Kalaitzaki R, et al. The Impact of Systemic Therapy Beyond First-line Treatment for Advanced Cervical Cancer. *Clin Oncol*. 2017;29:153–60.
doi:<https://doi.org/10.1016/j.clon.2016.10.002>.

6. Soisson AP, Soper JT, Clarke-Pearson DL, Berchuck A, Montana G, Creasman WT. Adjuvant radiotherapy following radical hysterectomy for patients with stage IB and IIA cervical cancer. *Gynecol Oncol.* 1990;37:390–5.
7. Sedlis A, Bundy BN, Rotman MZ, Lentz SS, Muderspach LI, Zaino RJ. A randomized trial of pelvic radiation therapy versus no further therapy in selected patients with stage IB carcinoma of the cervix after radical hysterectomy and pelvic lymphadenectomy: A Gynecologic Oncology Group Study. *Gynecol Oncol.* 1999;73:177–83.
8. Koh W-J, Abu-Rustum NR, Bean S, Bradley K, Campos SM, Cho KR, et al. Cervical cancer, version 3.2019, NCCN clinical practice guidelines in oncology. *J Natl Compr Cancer Netw.* 2019;17:64–84.
9. Moschetta M, Uccello M, Kasenda B, Mak G, McClelland A, Boussios S, et al. Dynamics of Neutrophils-to-Lymphocyte Ratio Predict Outcomes of PD-1/PD-L1 Blockade. *Biomed Res Int.* 2017;2017:1506824. doi:10.1155/2017/1506824.
10. Belinsky SA, Nikula KJ, Palmisano WA, Michels R, Saccomanno G, Gabrielson E, et al. Aberrant methylation of p16INK4a is an early event in lung cancer and a potential biomarker for early diagnosis. *Proc Natl Acad Sci.* 1998;95:11891–6.
11. Alvarez H, Opalinska J, Zhou L, Sohal D, Fazzari MJ, Yu Y, et al. Widespread hypomethylation occurs early and synergizes with gene amplification during esophageal carcinogenesis. *PLoS Genet.* 2011;7.
12. Ali Khan M, Kedhari Sundaram M, Hamza A, Quraishi U, Gunasekera D, Ramesh L, et al. Sulforaphane Reverses the Expression of Various Tumor Suppressor Genes by Targeting DNMT3B and HDAC1 in Human Cervical Cancer Cells. *Evidence-Based Complement Altern Med.* 2015;2015:412149. doi:10.1155/2015/412149.
13. Bao X, Anastasov N, Wang Y, Rosemann M. A novel epigenetic signature for overall survival prediction in patients with breast cancer. *J Transl Med.* 2019;17:380.
14. Löf-Öhlin ZM, Sorbe B, Wingren S, Nilsson TK. Hypermethylation of promoter regions of the APC1A and p16INK4a genes in relation to prognosis and tumor characteristics in cervical cancer patients. *Int J Oncol.* 2011;39:683–8.
15. Abudukadeer A, Bakry R, Goebel G, Mutz-Dehbalaie I, Widschwendter A, Bonn GK, et al. Clinical relevance of CDH1 and CDH13 DNA-methylation in serum of cervical cancer patients. *Int J Mol Sci.* 2012;13:8353–63.

16. Li J-Y, Huang T, Zhang C, Jiang D-J, Hong Q-X, Ji H-H, et al. Association between RASSF1A promoter hypermethylation and oncogenic HPV infection status in invasive cervical cancer: a meta-analysis. *Asian Pac J Cancer Prev.* 2015;16:5749–54.
17. Li X, Cai Y. Methylation-based classification of cervical squamous cell carcinoma into two new subclasses differing in immune-related gene expression. *Int J Mol Sci.* 2018;19:3607.
18. Li C, Ke J, Liu J, Su J. DNA methylation data--based molecular subtype classification related to the prognosis of patients with cervical cancer. *J Cell Biochem.* 2019.
19. Xie F, Dong D, Du N, Guo L, Ni W, Yuan H, et al. An 8-gene signature predicts the prognosis of cervical cancer following radiotherapy. *Mol Med Rep.* 2019;20:2990–3002.
20. Network CGAR. Integrated genomic and molecular characterization of cervical cancer. *Nature.* 2017;543:378–84.
21. Deng M, Brägelmann J, Kryukov I, Saraiva-Agostinho N, Perner S. FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. *Database.* 2017;2017.
22. Pfaendler KS, Tewari KS. Changing paradigms in the systemic treatment of advanced cervical cancer. *Am J Obstet Gynecol.* 2016;214:22–30.
23. Boussios S, Seraj E, Zarkavelis G, Petrakis D, Kollas A, Kafantari A, et al. Management of patients with recurrent/advanced cervical cancer beyond first line platinum regimens: Where do we stand? A literature review. *Crit Rev Oncol Hematol.* 2016;108:164–74.
24. Consortium TGO. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2014;43:D1049–56. doi:10.1093/nar/gku1179.
25. Jiao X, Sherman BT, Huang DW, Stephens R, Baseler MW, Lane HC, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics.* 2012;28:1805–6.
26. Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.
27. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch. and F. <https://CRAN.R-project.org/package=e1071>. 2019.

28. Laszczyńska O, Severo M, Friões F, Lourenço P, Silva S, Bettencourt P, et al. Validity of the Seattle Heart Failure Model for prognosis in a population at low coronary heart disease risk. *J Cardiovasc Med*. 2016;17:653–8.
29. Maaten L van der, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9 Nov:2579–605.
30. Krijthe J, van der Maaten L, Krijthe MJ. Package ‘Rtsne.’ 2018.
31. Thakur C, Chen B, Li L, Zhang Q, Yang Z-Q, Chen F. Loss of mdig expression enhances DNA and histone methylation and metastasis of aggressive breast cancer. *Signal Transduct Target Ther*. 2018;3:1–9.
32. Kulis M, Merkel A, Heath S, Queirós AC, Schuyler RP, Castellano G, et al. Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat Genet*. 2015;47:746.
33. Ai T, Zhang J, Wang X, Zheng X, Qin X, Zhang Q, et al. DNA methylation profile is associated with the osteogenic potential of three distinct human odontogenic stem cells. *Signal Transduct Target Ther*. 2018;3:1–8.
34. Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, et al. DNA methylation markers for diagnosis and prognosis of common cancers. *Proc Natl Acad Sci*. 2017;114:7414–9.
35. Jones PA, Laird PW. Cancer-epigenetics comes of age. *Nat Genet*. 1999;21:163–7.
36. Nagano M, Hoshino D, Koshiba S, Shuo T, Koshikawa N, Tomizawa T, et al. ZF21 protein, a regulator of the disassembly of focal adhesions and cancer metastasis, contains a novel noncanonical pleckstrin homology domain. *J Biol Chem*. 2011;286:31598–609.
37. Takahashi M, Tsukamoto Y, Kai T, Tokunaga A, Nakada C, Hijiya N, et al. Downregulation of WDR 20 due to loss of 14q is involved in the malignant transformation of clear cell renal cell carcinoma. *Cancer Sci*. 2016;107:417–23.
38. Arnadottir GA, Norddahl GL, Gudmundsdottir S, Agustsdottir AB, Sigurdsson S, Jensson BO, et al. A homozygous loss-of-function mutation leading to CYBC1 deficiency causes chronic granulomatous disease. *Nat Commun*. 2018;9:1–9.
39. Wang Q, Li P, Wu W. A systematic analysis of immune genes and overall survival in cancer patients. *BMC Cancer*. 2019;19:1225.

40. Moarii M, Boeva V, Vert J-P, Reyat F. Changes in correlation between promoter methylation and gene expression in cancer. BMC Genomics. 2015;16:873.
41. Westendorf JM, Konstantinov KN, Wormsley S, Shu M-D, Matsumoto-Taniura N, Pirollet F, et al. M phase phosphoprotein 10 is a human U3 small nucleolar ribonucleoprotein component. Mol Biol Cell. 1998;9:437–49.

Supplementary Material --

An 8-DNA methylation signature predicts recurrence risk of cervical cancer

Jing-Hang Ma, Yu Huang, Lu-Yao Liu, Zhen Feng

Statistical analysis of the histopathological types

Squamous cell carcinoma and adenocarcinoma are two major histologic types of cervical cancer [1, 2]. In 135 cervical cancer samples in the TCGA database, 116 cases were squamous cell carcinoma, 18 cases were adenocarcinoma and 1 case was adenosquamous cell carcinoma respectively.

Univariate analysis was performed among histological types. The result showed that there was no statistically significant difference for recurrence rate (0.379 vs 0.313 vs 1, p -value = 0.2882, Fisher's Test) and for SVM score (0.739 ± 1.405 vs 0.783 ± 1.335 vs 0.022, p -value = 0.8999, T-test). Therefore, we combined patient samples with different histologic types, since histologic types of cervical cancer were not associated with the recurrence and SVM score.

We further inspected on the predominated squamous type. Univariate analysis and Cox regression analysis were conducted in **Table.S1**. The result showed that SVM score was still the independent risk factor for recurrence and T stage, instead of the N stage, became the independent risk factor for recurrence.

Statistical analysis of the N stage

The distribution of the N stage in 135 samples is N0: 56; N1: 25; Nx: 54. The unknown N stage (Nx) accounts for nearly half of the total number of samples. To exclude the impact of Nx, we omitted samples with the Nx and performed univariate analysis and Cox regression analysis on the remaining 81 samples in **Table.S2**. The result showed that SVM score and N stage were still independent risk factors for recurrence.

We also deleted N stage and discovered that SVM score became the only independent risk factor for recurrence (**Table.S3**).

References

1. Hopkins MP, Morley GW. A comparison of adenocarcinoma and squamous cell carcinoma of the cervix. *Obstet Gynecol.* 1991;77:912–7. <http://europepmc.org/abstract/MED/2030867>.
2. Green J, Berrington de Gonzalez A, Sweetland S, Beral V, Chilvers C, Crossley B, et al. Risk factors for adenocarcinoma and squamous cell carcinoma of the cervix in women aged 20–44 years: the UK National Case–Control Study of Cervical Cancer. *Br J Cancer.* 2003;89:2078–86. doi:10.1038/sj.bjc.6601296.

Table. S 1 Univariate and Cox regression analysis for squamous cell carcinomas only.

Clinical Char.	Mode	Univariate analysis			Cox regression analysis	
		No Recur. (N=72)	Recur. (N=44)	P-value	HR (95% CI)	P-value
Age, years ^ζ	mean± SD	48.88±12.77	52.27± 13.99	2.00E- 01		
SVM score ^ζ	mean± SD	1.32±1.19	-0.21±1.21	2.35E- 09	0.437 (0.282- 0.675)	1.91E- 04
Clinical Char.	Group	No Recur. (N=72)	Recur. (N=44)	P-value		
FIGO Stage ^η	1	37	20	1.89E- 01		
	2	22	10			
	3	11	9			
	4	1	4			
	NA	1	0			
Pathologic M ^η	M0	36	10	1.17E- 02	1.075 (0.522- 2.216)	8.44E- 01
	M1	0	3			
	MX	25	15			
Pathologic N ^η	N0	35	9	2.63E- 02	0.953 (0.433- 2.098)	9.04E- 01
	N1	11	12			
	NX	16	10			
Pathologic T ^η	T1	34	13	3.17E- 02	1.537(1.018- 2.317)	4.11E- 02
	T2	21	8			
	T3	4	1			
	T4	1	4			
	TX	2	5			
Pathologic Grade ^η	1	4	2	1.88E- 01		
	2	38	15			
	3	24	24			
	NA	5	2			

ζ: T test; η: Fisher's exact test.

NA: not available; SD: standard deviation; HR: hazard ratio; CI: confidence interval.

Table. S 2 Univariate and Cox regression analysis irrespective of samples with the Nx.

Clinical Char.	Mode	Univariate analysis			Cox regression analysis	
		No Recur. (N=58)	Recur. (N=23)	P-value	HR (95% CI)	P-value
Age, years ζ	mean $\pm$ SD	48.04 $\pm$ 10.01	50.53 $\pm$ 13.85	4.39E-01		
SVM score ζ	mean $\pm$ SD	1.31 $\pm$ 1.30	-0.09 $\pm$ 0.98	2.49E-06	0.422 (0.235-0.757)	3.81E-03
Clinical Char.	Group	No Recur. (N=58)	Recur. (N=23)	P-value		
FIGO Stage η	1	38	17			
	2	12	4			
	3	6	0	2.06E-01		
	4	1	2			
	NA	1	0			
Pathologic M η	M0	35	11			
	M1	1	1	4.68E-01		
	MX	22	11			
Pathologic N η	N0	46	10	2.94E-03	4.617 (1.368-15.589)	1.37E-02
	N1	12	13			
	T1	41	14			
Pathologic T η	T2	16	5			
	T3	1	1	2.82E-02	1.997 (0.937-4.255)	7.32E-02
	T4	0	2			
	TX	0	1			
Pathologic Grade η	1	6	1			
	2	32	9	2.83E-01		
	3	18	13			
	NA	2	0			

ζ : T test; η : Fisher's exact test.

NA: not available; SD: standard deviation; HR: hazard ratio; CI: confidence interval.

Table. S 3 Univariate and Cox regression analysis without the N stage.

Clinical	Mode	Univariate analysis			Cox regression analysis	
		No Recur.	Recur.	P-value	HR (95% CI)	P-value

Char.		(N=85)	(N=50)			
Age, years ^ζ	mean±SD	48.88±12.37	51.16±13.99	3.31E-01		
SVM score ^ζ	mean±SD	1.32±1.14	-0.21±1.21	2.62E-11	0.452 (0.302-0.676)	1.11E-04
Clinical Char.	Group	No Recur. (N=85)	Recur. (N=50)	P-value		
FIGO Stage ^η	1	48	24			
	2	23	11			
	3	11	9	3.09E-02	0.777 (0.422-1.433)	4.19E-01
	4	2	5			
	NA	1	1			
Pathologic M ^η	M0	39	11			
	M1	1	4	2.15E-02	0.767 (0.454-1.295)	3.20E-01
	MX	45	35			
Pathologic T ^η	T1	45	16			
	T2	23	9			
	T3	4	2	5.55E-02		
	T4	1	4			
Pathologic Grade ^η	TX	12	19			
	1	6	3			
	2	46	16	3.67E-02		
	3	26	28			
	NA	6	2			

ζ: T test; η: Fisher's exact test.

NA: not available; SD: standard deviation; HR: hazard ratio; CI: confidence interval.