Identification of key pathways and establishment of a seven-gene prognostic signature in cervical cancer

Silu Meng1,2, Ruoqi Ning1,*, Jianwei Zhang1, Shennan Wang2, Xinran Fan1, Bing Xiao1, Shuang, Li1,*

1Department of Obstetrics and Gynecology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, 430030, P.R. China.

2Department of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, 430030, P.R. China.

* Silu Meng and Ruoqi Ning contributed equally to this work.

* Correspondence to:
Shuang Li (shuangli@hust.edu.cn)
Abstract

**Background:** Cervical cancer (CC) remains high morbidity and mortality. We aimed to identify critical pathways underlying cervical carcinogenesis and establish a prognostic signature.

**Methods:** Six datasets from the gene expression omnibus (GEO) database were used to screen the differentially expressed genes (DEGs) between CC and normal tissues. We used the unions of the DEGs to perform functional analysis. The 108 overlapped DEGs were analyzed to determine a prognostic signature by Cox regression and Lasso analysis based on The Cancer Genome Atlas (TCGA) database. Gene Set Enrichment Analysis (GSEA) and Immune Cell Abundance Identifier (ImmuCellAI) were used to determine the relationships between the signature and biological functions.

**Results:** The PI3K-Akt signaling pathway, the Ras signaling pathway, and the viral carcinogenesis pathway may be the critical pathways for CC development. We identified 7 genes to construct a risk score formula. Survival analysis showed that the high-risk group indicated a worse prognosis than the low-risk group (p < 0.0001). The AUC of the prognostic signature was 0.7449, 0.7641, and 0.8146 at 1, 3, and 5 years. We also identified that the signature is an independent prognostic factor. GSEA showed five pathways were relevant to the signature, such as the adherens junction pathway. The abundances of various types of immune cells, such as B cell, CD4+ T cell and CD8+ T cell, were also affected by the signature.

**Conclusion:** In conclusion, the present study explored the key pathways for CC development and the seven-gene signature can effectively make the prognosis evaluation of CC patients.

**Keywords:** cervical cancer, cervical carcinogenesis, prognostic signature, overall survival
Introduction

Cervical cancer (CC) is the fourth most commonly diagnosed cancer and the fourth leading cause of cancer-associated mortality in women worldwide, with 570,000 new cases and 311,000 deaths estimated for 2018. The incidence of CC has tended to be younger in recent years (Bray et al. 2018; Chen et al. 2017; Jiang et al. 2018; Yu et al. 2015). The role of human papillomavirus (HPV) infection has been well established (Schiffman and Wentzensen 2010; zur Hausen 2002). And with the application and development of high-throughput sequencing, many studies have reported many vital genes and pathways in CC carcinogenesis, such as PIK3CA, FBXW7, EP300, MAPK signaling pathway, etc. (Cancer Genome Atlas Research et al. 2017; Huang et al. 2019; Ojesina et al. 2014). However, the detailed mechanism of carcinogenesis of CC is still unveiled and needs to be further studied. In clinical practice, early-stage CC patients are mainly treated with surgery, and late-stage patients are treated with chemoradiotherapy (Marth et al. 2017). However, the rate of recurrence of CC is approximately 20%–25%, and the 5-year survival rate for late-stage CC is less than 50% (Zeng et al. 2015). The International Federation of Gynecology and Obstetrics (FIGO) stage system has been one of the CC’s most important prognostic factors, while the significant differences in survival rate are reported in the same FIGO stage (Heintz et al. 2006). Therefore, it is crucial to identify sensitive and specific biomarkers that could predict the prognosis of CC patients and monitor patient’s outcomes.

Based on the TCGA database and GEO database, we aim to understand the mechanism of CC carcinogenesis better and predict the prognosis of CC patients more precisely. Differentially expressed genes (DEGs) were identified using six datasets in the GEO database, and these DEGs were processed to find the key pathways in CC development. Besides, we found out the genes related to the prognosis of patients from among the DEGs and established a model to predict patients' prognosis. To gain
further insight into the biological pathways and immune cells changes involved in CC pathogenesis related to this prognostic model, Gene Set Enrichment Analysis (GSEA) and immune cells abundance analysis were performed (Supplementary Figure 1).

Methods

Quality control and DEGs identification

We recruited six gene expression profiles of CC (GSE6791, GSE63514, GSE7803, GSE9750, GSE39001, and GSE52903) from the gene expression omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) in this study. The information of datasets is shown in Table 1. We downloaded the normalized data of GSE6791, GSE63514, GSE7803, GSE39001, and GSE52903. However, the data of GSE9750 was not normalized, and we normalized its raw data through the R “affy” package with the RMA algorithm method (Gautier et al. 2004). R “limma” package was used to screen DEGs between normal tissues and CC tissues (Ritchie et al. 2015). Genes with |LogFC| >1 and adjust p-value < 0.05 were considered as DEGs in this study.

Functional analysis based on DEGs

The union of DEGs of the six datasets was processed to Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analysis by R “clusterProfiler” package (Yu et al. 2012). The p-value cutoff and q-value cutoff in this study were 0.01 and 0.05, respectively.

Download and collation of CC data in TCGA Database

The gene expression data of the 304 CC cases (Workflow Type: HTSeq - FPKM-UQ) and corresponding clinical information were downloaded from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/) by R “TCGAbiolinks” package (Colaprico et al. 2016). The
information of distant metastasis, tumor status, vital status, and overall survival (OS), were from the article of Liu et al. published in Cell in 2018 (Liu et al. 2018a).

**Conduction of the risk formula for prognostic prediction**

The risk score formula was constructed using the data of CC of the TCGA database. Firstly, We used the Venn diagram to find the six datasets’ overlapped DEGs and finally got 108 DEGs. By performing univariate Cox regression analysis with R “survival” package, the association between the expression of 108 DEGs and patient’s OS was assessed. 26 DEGs with a p-value of less than 0.05 were included in the subsequent analysis. Secondly, based on the above identification of prognosis-related genes for CC, we further needed to narrow the gene range and establish a prognostic signature. Thus, we performed the Least absolute shrinkage and selection operator (Lasso) analysis by R “glmnet” package, which constructs a more refined model using a penalty function. This method can reduce the model's complexity and reduce the weight of some unimportant indicators to 0 and 13 DEGs were left in this step. Next, we used the 13 DEGs to perform a multivariate Cox regression analysis. In this step, we further optimized the model based on the Akaike's Information Criterion (AIC) value, and finally 7 genes (Procollagen-Lysine,2-Oxoglutarate 5-Dioxygenase 2 (PLOD2), Desmoglein 2 (DSG2), Secreted Phosphoprotein 1 (SPP1), C-X-C Motif Chemokine Ligand 8 (CXCL8), Minichromosome Maintenance Complex Component 5 (MCM5), Helicase Like Transcription Factor (HLTF), Kruppel Like Factor 4 (KLF4)) were left. A risk formula (prognostic signature) was then established based on a linear combination of these DEGs' expression levels, weighted by their regression coefficients derived from the multivariate Cox regression model. The formula was as follows: risk score = [Expression level of PLOD2 * (0.252751)] + [Expression level of DSG2 * (0.381041)] + [Expression level of SPP1 * (0.170055)] + [Expression level of CXCL8 * (0.163175)] + [Expression level of MCM5 *
Finally, risk score was computed for each patient with this formula and patients were classified into high-risk and low-risk group, by taking the median risk score as a cutoff point. To detect the 7 genes function, COXPREdb (https://coxpresdb.jp/) was used to find the top 200 co-expressed genes of the 7 genes and then these co-expressed were processed to KEGG analysis by R “clusterprofiler” package (Yu et al. 2012).

**Assessment of the signature’s effect**

We ranked the each sample’s risk score and counted the number of deaths in the high-risk group and the low-risk group. Kaplan-Meier estimate based on the log-rank test was used to compare the survival difference between the high-risk group and the low-risk group by R “survminer” package. Receiver operating characteristic (ROC) curves were employed to compare the sensitivity and specificity of the survival prediction based on the risk score model by R “survivalROC” package. To test whether the prognostic signature was independent of clinical variables, we performed the univariate Cox regression analysis were performed, and variables with p-value 0.05 were then analyzed by multivariable Cox regression analysis. Then the independent prognostic factor were chosen to construct the nomogram by using R packages “rms” and “forestplot”, and the ROC curves of the predictive nomogram were also performed. All statistical analyses were finished with R version 3.6.3.

**Gene Set Enrichment Analysis (GSEA)**

GSEA (version 4.1.0) was used to explore the signaling pathways related to the risk score model (Subramanian et al. 2005). The phenotypes were determined by the cutoff value of the risk score. The annotated gene set was selected (c2.cp.kegg.v7.1.symbols.gmt) as the reference gene set. Gene set permutations were performed 1,000 times for analysis. The normalized enrichment score (NES),
nominal $p$-value, and false discovery rate (FDR) $q$-value were used to sort the pathways enriched in each group. Pathways with NES > 1, nominal $p$-value < 0.05 and FDR $q$-value < 0.25 were selected out.

**Immune cell component analysis**

Because CXCL8 is an immune-related gene, we used the Immune Cell Abundance Identifier (ImmuCellAI), which estimates the abundance of 24 immune cells, to detect whether the prognostic signature will affect the immune microenvironment (Miao et al. 2020).

**Results**

**Quality control and DEGs identification**

We selected six gene expression datasets of CC (GSE6791, GSE63514, GSE7803, GSE9750, GSE39001, and GSE52903) to assess the DEGs between CC tissues and normal tissues. The quality control results of the six datasets are shown in **Supplementary Figure 2**. The information of the datasets and the number of DEGs of each dataset are shown in **Table 1** and **Figure 1A**.

**Functional analysis based on DEGs**

The union of DEGs of the six datasets was processed to KEGG and GO analysis. In the KEGG analysis, the up-regulated genes enriched in many well-known pathways, such as the cell cycle, the DNA replication, and the nucleotide excision repair (**Figure 1B and Supplementary Table 1a**). And the down-regulated genes are related to many cancer-related pathways, such as the Ras signaling pathway and the PI3K-Akt signaling pathway (**Figure 1B and Supplementary Table 1b**). In the GO analysis, the up-regulated genes mainly participate in the biological process (BP) related to cell cycle, such as the DNA replication, the G2/M transition of mitotic cell cycle, the regulation of mitotic cell cycle phase
transition, etc. (Figure 1C and Supplementary Table 1c). While the down-regulated genes mainly enrich in BP of the cornification, the epidermis development, the peptide cross-linking, etc. (Figure 1D and Supplementary Table 1f). The results of cellular component (CC) and molecular function (MF) are also show in Figure 1C-D and Supplementary Table 1d-e, g-h.

Information of CC in the TCGA database

We included and counted the CC’s clinical information, including age, keratinization, FIGO stage, differentiation, and lymphovascular invasion. And the information of distant metastasis, tumor status, vital status, and OS information was from the article Liu et al. (Table 2) (Liu et al. 2018a). Also, the gene expression data and clinical information were matched which was used for the subsequent survival analysis.

Establishment of a seven-genesis signature for prognosis prediction

The Venn diagram shows the 108 overlapped DEGs of the six datasets, of which 69 were up-regulated genes and 39 were down-regulated genes (Figure 2A). The names of the 108 DEGs are shown in the Supplementary Table 2. The relationship between the 108 DEGs and the patient’s OS was assessed by univariate Cox regression analysis, and 26 DEGs whose parameter p-values were less than 0.05 were chosen for subsequent analysis (Supplementary Table 3). Considering the number of genes and collinearity, Lasso regression analysis was used to narrow the gene range, and only 13 genes were remained in this step (Figure 2B-C). Next, the 13 genes were processed to the multivariate Cox regression model, and we further optimized the model based on the AIC value. Finally, 7 genes (PLOD2, DSG2, SPP1, CXCL8, MCM5, HLT, KLF4) were screened out as the predictor signature and their detailed information including coefficients, HR value, and p value are shown in Figure 2D. Of these, positive coefficients for the PLOD2, DSG2, SPP1, CXCL8 indicated that their up-regulated
levels of expression were associated with shorter survival. The negative coefficient of MCM5, HLTF, KLF4 indicated that up-regulated level of expression was associated with longer survival. A prognostic model based on the coefficients was established and the risk score formula was as follows: risk score = [Expression level of PLOD2 * (0.252751)] + [Expression level of DSG2 * (0.381041)] + [Expression level of SPP1 * (0.170055)] + [Expression level of CXCL8 * (0.163175)] + [Expression level of MCM5 * (-0.514752)] + [Expression level of HLTF * (-0.245823)] + [Expression level of KLF4 * (-0.184198)]. We then calculated the seven-genes signature risk score of each patient in using the above formula. The median risk score (0.3865) was used as the cutoff point to divide the patients into two groups, the high-risk group (N = 152) and the low-risk group (N = 152) (Supplementary Table 4). In addition, to understand the function of the 7 genes, the co-expressed genes were processed to KEGG analysis and they enriched in many cancer-related pathways, such as the PI3K-Akt signaling pathway, the Toll–like receptor signaling pathway, the P53 signaling pathway. And the co-expression genes of PLOD2 and DSG2 also enriched in the Human papillomavirus infection which is inseparable with CC development (Supplementary Figure 3).

Assessment of the signature’s effect

The samples were ranked according to their risk scores (Fig 3A) and the number of deaths increased significantly with the increase of the risk score (Figure 3B). A heatmap was visualized to demonstrate the expression profiles of the 7 genes (Fig 3C). We identified that the OS of the high-risk group is significantly shorter than the low-risk group by Kaplan-Meier method (Figure 3D). To further investigate the discrimination power of the signature, ROC curves were further performed. The area under the curve (AUC) of the signature was 0.7449, 0.7641 and 0.8146 at 1, 3 and 5 years (Figure 3E). To test whether the signature was
independent of clinical variables, univariate and multivariate Cox regression analysis were performed, showing that lymphovascular invasion (HR, 7.0050; 95%CI, 1.3995–35.0632; p = 0.0178), tumor status (HR, 94.2939; 95%CI, 17.2409–515.7118; p < 0.0001), and the signature (HR, 6.5823; 95%CI, 1.9382–22.3537; p = 0.0025) are the independent predictor of poor overall survival (Figure 4A-B). Next, we established a nomogram combining the three independent factor (Figure 4C). The AUC for 1, 3, and 5 year survival using the predictive nomogram reached 0.7549, 0.8062, and 0.8064, respectively (Figure 4D).

Identification of the seven-gene signature altered pathways

To identify potentially altered signaling pathways related to the signature, we performed GSEA analysis based on the risk score classification. We selected out 5 significantly enriched signaling pathways based on the standard NES > 1, nominal p-value < 0.05, and FDR q-value < 0.25, including adherent junction, ECM receptor interaction and focal adhesion, etc. (Figure 5 and Supplementary Table 5).

Immune cell component analysis

ImmuneAI was used to analyze the difference of immune cells between the high-risk and the low-risk group (Supplementary Table 6). The abundance of Tr1 cells, B cells, CD4_T cells, and CD8_T cells were significantly increased in the low-risk group, while the Th17 cells, NKT cells, NK cells and Neutrophil were significantly increased in the high-risk group (Figure 6).

Discussion

The morbidity and mortality of CC remain high. Despite the role of HPV in CC has been well established and many researches have been done in uncovering the potential molecular mechanism of
CC development, the underlying mechanism remains unclear. The surgical and chemoradiotherapy of CC are very mature, while many patients still suffered from recurrence and metastasis. The present research performed an integrated bioinformatical analysis based upon six mRNA expression profile datasets of the GEO database and 304 CC patients of TCGA database to further uncover the mechanism of CC and discover more promising and useful prognosis-related biomarkers in CC.

We used the union of the DEGs of the six datasets to perform KEGG and GO analysis. In the KEGG analysis, the up-regulated genes enriched in the cell cycle, the DNA replication, the nucleotide excision repair, etc., and the DEGs may promote cell proliferation by being involved in these pathways. These pathways are common in various kinds of cancer, such as colorectal cancer and gastric cancer (Liang et al. 2016; Liu et al. 2018b). In addition, the up-regulated genes also enriched in the viral carcinogenesis, which may also prove the role of HPV in the development of CC. The down-regulated genes enriched in the PI3K-Akt signaling pathway and the Ras signaling pathway, which have been reported playing critical roles in CC (Huang et al. 2019). In the GO analysis, the up-regulated genes were mainly related to cell cycle related terms, such as the DNA replication, the G2/M transition of mitotic cell cycle, and the regulation of mitotic cell cycle phase transition. This result is similar to that of KEGG to a large extent. In contrast, the down-regulated genes enriched in many epidermis-related pathways, such as epidermis development, epithelial cell proliferation, positive regulation of epithelial cell proliferation, and structural constituent of epidermis. In summary, the enriched GO terms and KEGG pathways explained the specific molecular mechanisms of CC to some extent.

We identified 108 common DEGs between the CC tissues and normal tissues of the six datasets, including 69 up-regulated genes and 39 down-regulated genes. We further analyzed the relationship between the prognosis of CC patients and the expression levels of the 108 DEGs by univariate Cox
analysis, of which 26 DEGs indicated the significant correlation with OS (p <0.05). Next, we performed both lasso and multivariate Cox analysis to narrow the gene range, and finally we got 7 genes (PLOD2, DSG2, SPP1, CXCL8, MCM5, HLTF, KLF4) to establish a prognostic model which was able to distinguish CC patients into the high-risk group and low-risk group. PLOD2 encodes the key enzyme mediating the formation of the stabilized collagen cross-link, which sometimes can be considered as the “highway” for cancer cell migration and invasion (Du et al. 2017). PLOD2 overexpressed in many cancers, including hepatocellular carcinoma, breast cancer, and sarcoma (Du et al. 2017). Overexpression of PLOD2 can also induce cell migration and invasion in CC and PLOD2 are correlated with the prognosis of CC patients (Wei et al. 2020; Xu et al. 2017). DSG2 is a cell adhesion protein of the cadherin superfamily, which can regulate cell-cell contact. And DSG2 has been reported playing key roles in tumorigenesis (Gupta et al. 2015; Zhou et al. 2017). Qin et al. has reported that DSG2 can promote tumor proliferation and metastasis and is correlated with poor prognosis in early-stage CC (Qin et al. 2020). SPP1 is a secreted glycoprophosphoprotein of the SIBLING family [12]. Deregulation of SPP1 has been identified in various cancers (Choe et al. 2018; Junnila et al. 2010; Zhao et al. 2020). Zhao et al. establish a 5-genes prognostic model for CC, of which SPP1 is among the 5 genes. CXCL8 is a important cytokine that can modulate proliferation, invasion, and migration of tumor cells and can induce tumor immunosuppression [16]. It has been reported that the CXCL8-CXCR1/2 axis is potential to be applied as a cancer therapeutic target (Liu et al. 2016). Yan et al. found that CXCL8 high expression was a poor independent prognostic parameter for CC patients (Yan et al. 2017). MCM5 is a member of the MCM family of chromatin-binding proteins and participates in cell cycle regulation (Gou et al. 2018). MCM5 has been reported as a predictive biomarker for both cervical preinvasive neoplasia and CC (Murphy et al. 2005). HLTF belongs to the
SWI/SNF family of proteins involved in chromatin remodeling and DNA repair, suggesting that it acts as a tumor suppressor gene (Dhont et al. 2016). HLTF expression is altered in cancer through two mechanisms: gene silencing by promoter hypermethylation or expression of truncated proteins that lack functional domains (Dhont et al. 2016). These mechanism has been widely proved in digestive tract cancers (Hibi et al. 2005; Hibi et al. 2003). In CC, Cho et al. showed overexpression of HLTF might confer patients with resistance to radiation (Cho et al. 2011). KLF4 functions both as a tumor suppressor and an oncogene, which is involved in cell differentiation and cell-cycle arrest (Rowland and Peeper 2006). It has shown that KLF4 can regulate cell proliferation, migration, and invasion in multiple cancers, including breast cancer, gastrointestinal cancer, and esophageal cancer (Wei et al. 2006; Yang et al. 2005; Yu et al. 2011). Yang et al. identified the tumor suppressor role of KLF4 in CC and found that promoter hypermethylation of KLF4 can inactive its tumor suppressor function in CC (Yang and Zheng 2012; Yang and Zheng 2014). To further understand the function of the 7 gene we selected, the top 200 co-expressed genes of the 7 genes were processed to KEGG analysis, which shows many critical pathways for CC development, including the PI3K-Akt signaling pathway, the Huam papillomavirus infection, and the p53 signaling pathway.

Based on the prognostic signature, the number of deaths increased significantly with risk score increase. Kaplan–Meier survival analysis showed that the high-risk group indicated a worse prognosis than the low-risk group (p < 0.0001). The AUC of the risk model was 0.7449, 0.7641, and 0.8146 at 1, 3, and 5 years. Next, the univariate and multivariate Cox regression analysis were performed, identifying that lymphovascular invasion, the tumor status, and the signature are independent prognostic factors in CC. The AUC of the predictive nomogram combining the three independent factors for 1, 3, and 5-year survival reached 0.7549, 0.8062, and 0.8064.
To further investigate the model’s underlying mechanism, we performed the GSEA to explore the signaling pathways related to the risk score model, and 5 pathways were selected out. The adherens junction pathway is an element of cell-cell junction, essential for maintaining tissue architecture and cell polarity and can regulate cell movement and proliferation. E-cadherin, and α- and β-catenin are the main components of AJ (Meng and Takeichi 2009). Loss or downregulation of E-cadherin expression is frequently observed in cancers and correlates with the malignancy of the tumors (Birchmeier et al. 1996). Fujimoto et al. has reported that decreased expression of main adhesion molecules may result in invasiveness and metastasis of CC (Fujimoto et al. 1997). Li et al. identified that the genes at breakpoints of HPV integration in CC also enrich in adherens junction pathway (Li et al. 2018). The ECM-receptor interaction pathway controls many cellular activities such as adhesion, migration, differentiation, proliferation, and apoptosis (van der Flier and Sonnenberg 2001). This pathway participates in the progression of various cancers such as breast cancer, prostate cancer, and gastric cancer (Andersen et al. 2018; Bao et al. 2019; Yan et al. 2018). Using two datasets of the GEO database, Wu et al. found that the ECM-receptor interaction pathway is the key pathway during CC development (Wu et al. 2018). Focal adhesion refers to the specialized structures at cell-extracellular matrix contact points, which play essential roles in important biological processes, including cell motility, cell proliferation, cell differentiation (Wozniak et al. 2004). Increased expression and amplification of the focal adhesion kinase gene in human cancer cells are common (Agochiya et al. 1999). Xu et al. showed that PLOD2 could improve the migration and invasion of CC cells by focal adhesion formation (Xu et al. 2017). We also found the other two pathways in this study, the renal cell carcinoma and small cell lung cancer pathways, which may indicate that the underlying molecular mechanism of CC development may coincide with the two kinds of cancer. In addition, we analyzed whether the risk
model affects the immune microenvironment. The abundance of Tr1 cells, B cells, CD4+ T cells, and CD8+ T cells are significantly increased in the low-risk group, while the Th17 cells, NKT cells, NK cells, and Neutrophil are significantly increased in the high-risk group. Wang et al. found that a higher level of activated memory CD4+ T cells was independently associated with favorable OS in CC (Wang et al. 2019). Although we don’t have the result of activated memory CD4+ T cells, the abundance of CD4+ T cells is higher in the low-risk group. It has shown that both the CC patients or cervical precancerous lesion patients have a higher proportion of Th17 cells. Increased Th17 cells were associated with clinical stage, lymph node metastasis, and invasion, and therefore accelerates the disease’s progression (Zhang et al. 2011). And this may partially explain the higher level of Th17 cells in the high-risk group.

In conclusion, we utilized the public online database to find the related pathways underlying CC development and establish a CC’s prognostic model. Firstly, the PI3K-Akt signaling pathway, the Ras signaling pathway, and the viral carcinogenesis pathway may be the critical pathways for CC development. Secondly, we established a seven-gene prognostic signature for CC and validated the effects of the signature. Finally, we explored the possible mechanism underlying the prognostic signature.e p53 signaling pathway.

Data availability statement

Publicly available datasets were analyzed in this study. These can be found in The Cancer Genome Atlas (https://portal.gdc.cancer.gov/); the NCBI Gene Expression Omnibus (GSE6791, GSE63514, GSE7803, GSE9750, GSE39001, GSE52903).
Author contributions

MS, and NR designed the overall idea of this study, analyzed the data, prepared the figures and tables, and wrote the drafts of the paper. ZJ, WS, FX and XB checked the statistical results. LS supervised this study and revised the drafts of the paper.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

Funding

This work was supported by the National Natural Science Foundation of China (Grant number: 81974410).

Reference


Andersen MK et al. (2018) Integrative metabolic and transcriptomic profiling of prostate cancer tissue containing reactive stroma Sci Rep-Uk 8 doi:ARTN 14269 10.1038/s41598-018-32549-1


Liang B, Li CN, Zhao JY (2016) Identification of key pathways and genes in colorectal cancer using
Liu XK et al. (2018b) Identification of Potential Key Genes Associated With the Pathogenesis and Prognosis of Gastric Cancer Based on Integrated Bioinformatics Analysis Front Genet 9 doi:ARTN 265
10.3389/fgene.2018.00265
10.1101/cshperspect.a002899
Qin SH et al. (2020) DSG2 expression is correlated with poor prognosis and promotes early-stage cervical cancer Cancer Cell International 20 doi:10.1186/s12935-020-01292-x


Yan P, He YC, Xie KX, Kong S, Zhao WD (2018) In silico analyses for potential key genes associated with gastric cancer Peer 6 doi:ARTN e6092

10.7717/peerj.6092


10.1371/journal.pone.0088827


10.1155/2020/8347639


Figure legends

Figure 1: The number of DEGs in each dataset and KEGG and GO analysis of DEGs. (A) The number of DEGs in each dataset. (B) KEGG analysis. (C) GO analysis of up-regulated genes. (D) GO analysis of down-regulated genes. DEGs, Differentially expressed genes; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology.

Figure 2: Construction the seven-gene prognostic signature for cervical cancer. (A) The overlapped up-regulated genes; (B) The overlapped down-regulated genes; (C) The lasso analysis identified the most correlated genes with prognosis. (D) The distribution of each lambda and confidence interval. (E) Further narrow of the gene range by multivariate Cox analysis and 7 genes were determined. Lasso, Least absolute shrinkage and selection operator.

Figure 3: Assessment of the prognostic signature. (A) Distribution of risk score in the high-risk group and the low-risk group. (B) Survival status between the high-risk group and the low-risk group. (C) The heatmap of expression profile of 7 genes. Red parts represent up-regulation, blue parts represent down-regulation, and white parts represent no differential expression. (D) Survival analysis showed that the patients in the high-risk group had statistically significantly worse overall survival than those in low-risk group based on TCGA database. (E) ROC analysis was performed to calculate the AUC of 1, 3, 5-year survival for this prognostic signature. TCGA, The Cancer Genome Atlas; ROC, Receiver operating characteristic; AUC, Area under curve.

Figure 4: The Cox regression analysis for evaluating the independent prognostic value of the risk score. (A) The univariate Cox analysis. (B) The multivariate Cox analysis. (C) The nomogram to predict the probabilities 1-year, 3-year, and 5-year OS in patients. (D) ROC curves according to the Nomogram. Age, Continuous variable; Keratinization, Yes vs. No; FIGO, Continuous variable (I, II, III and IV
stage); Grade, ≥G3 vs <G3; Lymphovascular invasion, Yes vs. No; Distant metastasis, Yes vs. No; Tumor status, With tumor vs. Tumor free; ROC, Receiver operating characteristic.

Figure 5: A merged enrichment plot from gene set enrichment analysis (GSEA) including enrichment score and gene sets.

Figure 6: Immune cell abundance analysis between the high-risk and the low-risk group. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$

Supplementary Figure 1: Flow work of this study.

Supplementary Figure 2: Quality control of the six datasets.

Supplementary Figure 3: KEGG analysis of the top 200 coexpressed genes of the 7 genes of the prognostic signature. KEGG, Kyoto Encyclopedia of Genes and Genomes.

Supplementary Table 1: Function analysis of DEGs of the six dataset. (a-b) KEGG of up-regulated and down-regulated genes, respectively. (c-e) BP, CC, and MF of up-regulated genes, respectively. (f-g)BP, CC, and MF of down-regulated genes, respectively. KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, Biological process; CC, Cellular component; MF, Molecular function.

Supplementary Table 2: 108 Common DEGs of the six datasets.

Supplementary Table 3: The associations between overall survival and 108 common DEGs using univariate cox analysis.

Supplementary Table 4: The risk score and risk group of each patient of TCGA database.

Supplementary Table 5: Gene sets enriched in the high-risk group.

Supplementary Table 6: Immune cells abundance analysis of the high-risk group and the low-risk group.
Table 1: Information of the datasets used in this study.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Year</th>
<th>Platform</th>
<th>Normal</th>
<th>Cervical cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE6791</td>
<td>2007</td>
<td>GPL570</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>GSE63514</td>
<td>2015</td>
<td>GPL570</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>GSE7803</td>
<td>2007</td>
<td>GPL96</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>GSE9750</td>
<td>2008</td>
<td>GPL96</td>
<td>24</td>
<td>33</td>
</tr>
<tr>
<td>GSE39001</td>
<td>2013</td>
<td>GPL6244</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>GSE52903</td>
<td>2015</td>
<td>GPL6244</td>
<td>17</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 2: Characteristics of the cervical cancer patients in TCGA database.

<table>
<thead>
<tr>
<th>Clinical characteristics</th>
<th>Subgroup</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>304</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>Range:20-88 (Average:48.2, Median: 46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keratinization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>119</td>
<td>68.4</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>55</td>
<td>31.6</td>
<td></td>
</tr>
<tr>
<td>N1</td>
<td>60</td>
<td>31.1</td>
<td></td>
</tr>
<tr>
<td>FIGO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>162</td>
<td>54.5</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>69</td>
<td>23.2</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>45</td>
<td>15.2</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>21</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>Differentiation grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤G2</td>
<td>153</td>
<td>56.3</td>
<td></td>
</tr>
<tr>
<td>≥G3</td>
<td>119</td>
<td>43.7</td>
<td></td>
</tr>
<tr>
<td>Lymphovascular invasion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>71</td>
<td>47.3</td>
<td></td>
</tr>
<tr>
<td>Present</td>
<td>79</td>
<td>52.7</td>
<td></td>
</tr>
<tr>
<td>Distant metastasis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>273</td>
<td>89.8</td>
<td></td>
</tr>
<tr>
<td>YES</td>
<td>31</td>
<td>10.2</td>
<td></td>
</tr>
<tr>
<td>Tumor status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor free</td>
<td>197</td>
<td>71.1</td>
<td></td>
</tr>
<tr>
<td>With tumor</td>
<td>80</td>
<td>28.9</td>
<td></td>
</tr>
<tr>
<td>Vital status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alive</td>
<td>233</td>
<td>76.6</td>
<td></td>
</tr>
<tr>
<td>Dead</td>
<td>71</td>
<td>23.4</td>
<td></td>
</tr>
</tbody>
</table>
**D**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficients</th>
<th>HR (95% CI)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLOD2</td>
<td>0.2528</td>
<td>1.2876 (1.0301–1.6094)</td>
<td>0.0264</td>
</tr>
<tr>
<td>DSG2</td>
<td>0.3810</td>
<td>1.4638 (1.0924–1.9614)</td>
<td>0.0107</td>
</tr>
<tr>
<td>SPP1</td>
<td>0.1701</td>
<td>1.1853 (1.0408–1.3500)</td>
<td>0.0104</td>
</tr>
<tr>
<td>CXCL8</td>
<td>0.1632</td>
<td>1.1772 (1.0451–1.3260)</td>
<td>0.0072</td>
</tr>
<tr>
<td>MCM5</td>
<td>-0.5148</td>
<td>0.5976 (0.4044–0.8833)</td>
<td>0.0098</td>
</tr>
<tr>
<td>HLTF</td>
<td>-0.2458</td>
<td>0.7821 (0.5654–1.0817)</td>
<td>0.1374</td>
</tr>
<tr>
<td>KLF4</td>
<td>-0.1842</td>
<td>0.8318 (0.6768–1.0222)</td>
<td>0.0799</td>
</tr>
</tbody>
</table>
Survival probability

A

Risk score

B

Survival time (years)

C

Time (Months)

D

Survival probability

E

False positive rate
### A

<table>
<thead>
<tr>
<th>Variable</th>
<th>HR (95% CI)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1.0171 (0.9995−1.0349)</td>
<td>0.0570</td>
</tr>
<tr>
<td>Keratinization</td>
<td>1.6856 (0.8725−3.2566)</td>
<td>0.1202</td>
</tr>
<tr>
<td>FIGO</td>
<td>1.4916 (1.1982−1.8569)</td>
<td>0.0003</td>
</tr>
<tr>
<td>Grade</td>
<td>0.8894 (0.5267−1.5018)</td>
<td>0.6610</td>
</tr>
<tr>
<td>Lymphovascular invasion</td>
<td>10.0413 (2.3613−42.7000)</td>
<td>0.0018</td>
</tr>
<tr>
<td>Distant metastasis</td>
<td>3.1412 (1.8657−5.2886)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Tumor status</td>
<td>21.5956 (11.2868−41.3202)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Risk</td>
<td>4.7053 (2.7001−8.1997)</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

### B

<table>
<thead>
<tr>
<th>Variable</th>
<th>HR (95% CI)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIGO</td>
<td>1.2897 (0.7273−2.2872)</td>
<td>0.384</td>
</tr>
<tr>
<td>Lymphovascular invasion</td>
<td>7.0050 (1.3995−35.0632)</td>
<td>0.0178</td>
</tr>
<tr>
<td>Distant metastasis</td>
<td>0.3384 (0.1139−1.0059)</td>
<td>0.0512</td>
</tr>
<tr>
<td>Tumor status</td>
<td>94.2939 (17.2409−515.7118)</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Risk</td>
<td>6.5823 (1.9382−22.3537)</td>
<td>0.0025</td>
</tr>
</tbody>
</table>

### C

- **Points**
  - **Lymphovascular invasion**: No → Yes
  - **Tumor status**: Tumor free → With tumor
  - **Risk**: Low risk → High risk
  - **Total Points**
    - 0
    - 20
    - 40
    - 60
    - 80
    - 100
    - 120
    - 140
    - 160
    - 180
    - 200

### D

- **1-year survival (AUC = 0.7549)**
- **3-year survival (AUC = 0.8062)**
- **5-year survival (AUC = 0.8064)**

---

**Graph A**

**Graph B**

**Graph C**

**Graph D**

---

**Table A**

**Table B**

**Table C**

**Table D**

---

**Diagram A**

**Diagram B**

**Diagram C**

**Diagram D**