Identification of a Six-gene Signature Predicting Overall Survival for Cervical Cancer

Hengzi Sun

Peking Union Medical College Hospital

Xiaoshuang Zhou

Peking Union Medical College Hospital

Peng Peng

Peking Union Medical College Hospital

Mei Yu

Peking Union Medical College Hospital

Ying Zhang

Peking Union Medical College Hospital

Jiaxin Yang

Peking Union Medical College Hospital

Dongyan Cao

Peking Union Medical College Hospital

Xiao Huo (✉ huoxiaopumc@126.com)

Peking Union Medical College Hospital

Keng Shen

Peking Union Medical College Hospital

Research

Keywords: Cervical cancer, Bioinformatics, Prognostic signature, Gene expression omnibus, Overall survival

Posted Date: August 7th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-52590/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Although the incidence of cervical cancer has decreased in decades with the development of human papillomavirus vaccines and cancer screening, cervical cancer remains one of the leading causes of cancer-related deaths worldwide. Identify the potential biomarker for treatment and prognosis of cervical cancer is necessary.

Methods: Samples with mRNA-seq, copy number variant, single single nucleotide polymorphism data and clinical follow-up information were download from TCGA database, which were randomly divided into training datasets (N=146) and test datasets (N=147). We selected and identified prognostic gene set and genomic mutated gene set and then integrated the two set of data with random survival forest algorithm and constructed a prognostic signature. External validation datasets and immunohistochemical staining were also evaluated.

Results: We obtained 1,416 different expression prognostic-related genes, 624 genes with copy number amplification, 1,038 genes with copy number deletion, and 163 significantly mutated gene. A total of 75 candidate genes were obtained after overlap of the different expression genes and genomic variations. Subsequently we obtained six characteristic genes through random survival forest algorithm. The results showed that high expression of SLC19A3, FURIN, SLC22A3, DPAGT1 and low expression of CCL17, DES were associated with poor prognosis in cervical cancer patients. We constructed a six-gene signature which can separate cervical cancer samples associated with different overall survival and showed robust performance for predicting survival (Training set: $p \leq 0.001$, AUC = 0.82; Testing set: $p \leq 0.01$, AUC = 0.59).

Conclusions: Our study identified a novel six-gene signature and nomogram to predict overall survival of cervical cancer, which may be beneficial to clinical decision making for individual treatment.

Introduction

Cervical cancer (CC) is one of the main causes of cancer death in women and the third leading malignancy among women after breast and colorectal cancers worldwide, with 569,000 new cases each year[1, 2]. Cervical cancer is still a major problem for healthy women worldwide despite significant efforts. Walboomers JM *et al*/ demonstrated that human papillomavirus (HPV) is a necessary cause of cervical squamous cell carcinoma (CSCC) [3]. It is proved that HPV integration induces genomic instability and somatic mutation, playing an important role in the pathogenesis of CC[4]. Although the development of cervical cancer screening has led to a drastic reduction in the incidence of cervical cancer, there are still many challenges for advanced lesions. The 5-year survival rates of distant cervical squamous cell carcinoma patients were only 17% in United States [5]. Therefore, searching for potential biomarkers for treatment and prognosis of CSCC with multi-omics data is necessary. However, few studies were conducted to explore the relation between genomic factor and CSCC prognosis.

At present, biomarkers used for the prognosis of CSCC are mainly divided into two categories. One is the clinical biomarkers for CSCC survival prediction, such as squamous cell antigen (SCC). The other one is

to fine prognosis signatures constructed by several prognostic genes by high-throughput sequencing data analysis. For instance, Shen *et al*/discovered that CD28 and PTEN gene play important roles in the occurrence and development of CSCC through methylated microarray data[6]. Mao *et al*/identified lncRNAs related to the prognosis of CSCC from TCGA database and developed a 15-lncRNA signature risk score to comprehensively assess the prognostic function of lncRNA[7]. Ma *et al*/found the potential miRNA expression signature capable of predicting survival time for CSCC patients[8]. LRIG1 and LRIG3 expression in CSCC is related to HPV status and patient survival[9]. Therefore, it may probably be crucial and helpful to identify genes of prognostic value by bioinformatics analysis. High-throughput multi-omics sequencing data has laid a solid foundation for identifying genes related to cancer prognosis. Multi-omics data analysis can reveal the mechanism of cancer development from multiple perspectives.

In this study, we analyzed the transcriptome data to obtain genes related to the prognosis of CSCC. And then we analyzed the copy number variation (CNV) data and mutation data of CSCC to obtain genes related to the occurrence and development of CSCC. Subsequently, we integrated the results and propose a prognostic gene signature by using the random survival forest algorithm. It is proved that the model can effectively verify the test dataset and external independent dataset, and the signature shows a strong clinical independence. Overall, the prognostic gene signature can effectively predict the prognostic risk of CC and provide a basis for a better understanding of the pathogenesis of CSCC.

Materials And Methods

1. Collection and pre-processing of publicly available data

Cervical cancer RNA sequencing (RNA-seq) FPKM dataset and corresponding clinical follow-up information were downloaded from publicly available TCGA database by UCSC cancer browser (<https://xenabrowser.net/datapages/>), which contains 309 samples and 306 samples respectively. The copy number variation (CNV) data of the SNP 6.0 array contains a total of 297 samples. The mutation annotation file (MAF) downloaded from the GDC client contains a total of 289 samples. The GSE44001 expression profile data and clinical follow-up information were downloaded from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>), which contains a total of 300 samples.

The TCGA transcriptome data were standardized and randomly divided into two groups, one group as the training datasets (n=146), and the other group as the test datasets (n=147). The GSE44001 data were used as independent validation datasets. The age, survival status, grade and tumor stage were expressed in Supplementary S1.

2. Univariate Cox proportional-hazards regression analysis

Univariate Cox proportional-hazards regression analysis was performed for each transcriptome gene in TCGA training datasets, and a gene with a significance level of $p < 0.05$ was selected as a prognostic gene.

3. Copy number variation (CNV) analysis

GISTIC 2.0[10] software was used to identify the genes with significant amplification or deletion. The threshold was set as amplification or deletion length greater than 0.1 and significance level of $p < 0.05$.

4. Gene mutation analysis

MUTSIG 2.0 software was used to identify the genes in MAF with significant mutation with $p < 0.05$.

5. Gene signature construction

At first, the candidate genes significantly related to survival were screened for significant amplification, deletion, and mutation. Then we used random survival forest algorithm[11] to rank the importance of the genes. Monte Carlo iterations are 100, and the number of advances is 5 ($nrep = 100$, $nstep = 5$), Genes with relative importance greater than 0.5 were identified as characteristic genes. Finally, we performed a multivariate Cox regression analysis to build the risk score model:

$$RiskScore = \sum_{k=1}^n Exp_k \times e_k^{HR}$$

where N is the number of prognostic genes, Exp_k is the expression value of prognostic genes, and e_k^{HR} is the estimated regression coefficient of genes in the multivariate Cox regression analysis.

1. Functional enrichment analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed by using the R package *clusterprofiler*[12] for genes, to identify over-represented GO terms in biological processes categories and KEGG pathway. For this analysis, a $FDR < 0.05$ was considered to denote statistical significance.

2. Immunohistochemical staining (IHC)

We collected a total of 160 human cervical tissue samples, 130 of which had accompanying follow-up information, and 30 cancer-adjacent cervical tissue samples from archives of paraffin-embedded tissues between January 2010 and January 2014 at the Department of Pathology of Peking Union Medical College Hospital. The follow-up was performed until December, 2018. The pathological diagnoses were reconfirmed by a pathologist. The project was approved by the Ethical Committee (Peking Union Medical College Hospital), and informed consent was acquired from patients or family members. IHC was performed as previously described[13]. Antibodies against the following were used: CCL17 1:200; DES 1:100; DPAGT1 1:200; FURIN 1:200, SLC19A3 1:500, SLC22A3 1:500. The scoring details have been described previously[14]. Positive staining was defined as follows: five random microscope fields were selected to evaluate semi-quantification staining. The intensity of immunostaining was graded as follows: 1+, weak; 2+, moderate; 3+, strong or 4+, very strong. The area of positive cancer cells in each

microscopic field was categorized as follows: 1+, 0 to 25%; 2+, 25 to 50%; 3+, 50 to 75% or 4+, 75 to 100%. The sum between 5 and 80 was obtained by multiplying the two scores by 5. A sum from 0 to 42 was assigned as “low expression” and that from 43 to 80 as “high expression”. Two pathologists independently reviewed the slides using their individual criterion, and the consensus interpretations were used as the final interpretations. Discrepant interpretations were adjudicated by a third pathologist.

Results

1. Selection and identification of prognostic gene set and genomic mutated gene set

For TCGA training datasets, a total of 1,412 genes were selected as prognostic candidate genes basing on univariate Cox regression analysis ($p < 0.05$). The top 20 most significant prognostic genes with lowest p value were expressed in Table 1. (More details in Supplementary S2)

Meanwhile, genes with significant amplification or deletion were identified by GISTIC 2.0. Figure 1A shows the significantly amplified segments in cervical cancer genome. And the significantly amplified genes on each segment are recorded in Supplementary S3. For instance, ERBB2 was significantly amplified on the 17q12 segment (q value = $5.62E-12$), MYC was significantly amplified on the 8q24.21 segment (q value = $7.70E-09$), and EGFR was significantly amplified on the 7p11.2 segment (q value = 0.01188). In total, 624 genes were identified with significant amplification. Figure 1B shows the significantly deleted segments in cervical cancer genome. And the significantly deleted genes on each segment are recorded in Supplementary S4. For instance, CD3D was significantly deleted in 11q24.2 segment (q value = $9.11E-26$), SMAD4 was significantly deleted in 18q21.2 segment (q value = 0.0010189), and PTEN was significantly deleted in 10q23.31 (q value = 0.0077083). A total of 1,038 genes were identified with significant deletion.

Subsequently, we used MUTSIG 2.0 to identified genes with significant mutation with TCGA MAF data. $P < 0.05$ was set as the threshold. A total of 163 genes were identified with significantly mutation frequency. In Figure 1C, the top 50 genes with the most frequent mutations and the distribution of different mutation forms including synonymous mutations, missense mutations, frame insertions or deletions, frame movements, nonsense mutations, splice sites and other non-synonymous mutations in training datasets were shown. The upper histogram represents the total number of synonymous and nonsynonymous mutations in each patient's 50 genes, and the right histogram represents the mutation proportion in all samples of 50 genes. In these genes, some of which were reported to be closely related to the occurrence and development of cancer in previous studies can be identified, such as FBXW7, PIK3CA, PTEN, KRAS, TP53, etc.

Furthermore, we analyzed the pathways and biological processes involved in copy number variant genes and mutant genes to investigate the potential mechanism. After the integration of copy number variant genes and mutant genes, a total of 1,643 genes were obtained for functional enrichment analysis. As shown in Figure 1D, 1,643 genes were significantly enriched in human T-cell leukemia virus 1 infection,

Apoptosis, FoxO signaling pathway, ErbB signaling pathway, Central carbon metabolism in cancer, endometrial cancer, etc. In Figure 1E, 1,643 genes were significantly in developmental process, protein metabolic process, epidermal cell differentiation.

2. Prognostic signature construction for cervical cancer patients

In this study, we aimed to obtain a prognostic signature for cervical cancer prediction. Therefore, we selected 75 genes with significant copy number variant and mutation in prognostic candidate gene set for stepwise importance rank by random survival forest algorithm. Finally, we identified 6 genes with relative importance greater than 0.5 as the final prognostic signature. Hazard risk (HR), z-score, p value, importance and relative importance of 6 genes are shown in Table 2. Figure 1F shows the relationship between prediction error rates and the number of separate trees, and Figure 1G expresses the out-of-bag importance rank of the top 6 genes. Multivariate Cox regression analysis was conducted to develop a 6-gene signature and the risk model was as follow

$$Risk_6 = -0.6268828 * CCL17 + 0.6038325 * SLC19A3 + 0.4122066 * FURIN - 0.7570359 * DES + 0.278106 * SLC22A3 + 0.426686 * DPAGT1$$

Then the 6-gene signature risk score for each sample was calculated. According to the median value of all the sample, patients were divided into a high-risk group and a low risk group (cutoff=0.02513276). The usage of this 6-gene signature model in TCGA training datasets is shown in Figure 2. Figure 2A shows that 73 patients were divided into the high-risk group and the low-risk group respectively. Figure 2B demonstrates the receiver operating characteristic (ROC) curve, and the Area Under Curve (AUC) of 60-months is 0.82. Figure 2C shows that as the risk score increases, the survival period decreases significantly, and the majority of high-risk group are death samples. We also analyzed the expression level of 6 genes with the increase of risk score. And it is verified that patients in high risk group tend to have higher expression levels of SLC19A3, FURIN, SLC22A3 and DPAGT1, as well as lower expression of CCL17 and DES (Desmin).

1. Detection of the robustness of gene signature

To verify the accuracy of the 6-gene signature, the test datasets were researched with the same model and cut-off value as training datasets. Kaplan-Meier survival curve, ROC curve and AUC of 6-gene signature, the relation between risk score, survival period and gene expression level are shown in Figure 3. Figure 3A demonstrated the significant prognostic difference between high-risk group and low-risk group (log-rank p= 0.009019433). Figure 3B, C shows the consistent conclusion with training datasets that SLC19A3, FURIN, SLC22A3 and DPAGT1 might be risk factors while CCL17 and DES might be protective factors.

Furthermore, we validated the 6-gene signature in external independent datasets downloaded from GEO to confirm the stability with different platforms. The same formula and cut-off value was conducted. In

the consistence with our previous findings, Figure 4 demonstrated the Kaplan-Meier curves of GSE44001 datasets, which further suggested a significant prolonged survival time in the low-risk patients compared to that in the high-risk patients (Figure 4A). The AUC of 60-months is 0.59, and the relation between gene expression level and risk score is also consistent with TCGA datasets (Figure 4B, C). Therefore, the 6-gene signature has robust prognostic capabilities.

2. Independence evaluation of the prognostic signature

To confirm the independence of the 6-gene signature in clinical application, univariate and multivariate Cox regression analysis were conducted with TCGA training datasets, TCGA test datasets and GSE44001 datasets respectively. Variates was analyzed systemically in Table 3 (Supplementary S5) including age, grade, pathologic T stage, pathologic N stage, pathologic M stage, tumor stage and 6-gene risk group. Univariate Cox regression analysis verified that high risk group, pathologic T3/T4, tumor stage III/IV were significantly related to survival in TCGA training datasets, and multivariate Cox regression analysis verified that high risk group (HR= 2.89, 95%CI = 1.20-6.94, p= 0.018), pathologic T3/T4, tumor stage III/IV were independent. Meanwhile, high risk group, pathologic T4 and pathologic N1 were significantly related to survival in TCGA test datasets and high-risk group (HR= 1.444, 95%CI = 1.05-1.97, p= 0.020), pathologic T4 and pathologic N1 were clinically independent. For the GSE44001, high risk group was significantly related to survival and clinically independent (HR= 1.40, 95%CI = 1.09-1.78, p= 0.007).

3. Evaluation of the prognostic signature in cervical cancer by IHC

Supplementary S6 summarizes the characteristics of all patients, including age, disease stage, and tumour grade. The expression of the six gene were evaluated in cervical cancer by IHC. The differences in CCL17, DES, DPAGT1, FURIN, SLC19A3 and SLC22A3 expression between cervical cancer tissues and adjacent normal tissues are shown in Figure 5. Expression of DPAGT1 (53.77 ± 1.84 vs 21.00 ± 3.49 , $p < 0.001$), FURIN (48.55 ± 1.56 vs 28.83 ± 2.39 , $p < 0.001$), SLC19A3 (52.45 ± 1.77 vs 29.67 ± 4.48 , $p < 0.001$) and SLC22A3 (47.16 ± 1.44 vs 25.33 ± 3.32 , $p < 0.001$) was significantly higher in cervical cancer than in adjacent cancer tissue. Conversely, CCL17 (26.73 ± 1.65 vs 49.50 ± 3.79 , $p < 0.001$), DES (30.55 ± 1.50 vs 54.00 ± 2.90 , $p < 0.01$) showed lower expression in cervical cancer tissue. In addition, the correlation between expression of these genes and cervical cancer prognosis is shown in Figure 6. These data reveal that high expression of DPAGT1 (OS, HR = 2.469, 95% CI 1.170-4.586, $p = 0.016$; PFS, HR = 2.126, 95% CI 1.097-3.851, $p = 0.025$), FURIN (OS, HR = 2.291, 95% CI 1.158-4.589, $p = 0.018$; PFS, HR = 1.924, 95% CI 1.035-3.683, $p = 0.040$), SLC19A3 (OS, HR = 1.908, 95% CI 0.979-4.178, $p = 0.058$; PFS, HR = 1.993, 95% CI 1.093-4.184, $p = 0.027$) and SLC22A3 (OS, HR = 2.272, 95% CI 1.147-4.544, $p = 0.019$; PFS, HR = 2.113, 95% CI 1.135-4.030, $p = 0.019$) and low expression of CCL17 (OS, HR = 2.396, 95% CI 1.252-5.180, $p = 0.010$; PFS, HR = 2.102, 95% CI 1.132-4.157, $p = 0.020$) and DES (OS, HR = 2.179, 95% CI 1.079-4.231, $p = 0.030$; PFS, HR = 1.981, 95% CI 1.049-3.687, $p = 0.036$) are associated with poor prognosis in patients with cervical cancer.

Discussion

There were an estimated 569,800 new cervical cancer cases and 311,400 deaths worldwide in 2018[15]. Nearly 90% of cervical cancer deaths occurred in developing parts of the world. In several Western countries, where screening programs have long been established, cervical cancer rates have decreased by as much as 65% over the past 40 years. Whereas in contrast to favorable overall trends, cervical cancer rates are reported to be rising in Uganda and in some countries of Eastern Europe (Estonia, Lithuania, and Bulgaria)[16]. Although increasingly preventable through the use of vaccines, and curable through early cytological detection, women with advanced or recurrent disease face a dismal prognosis with potentially considerable morbidity and mortality[17]. The treatment and prognostic prediction of CC patients remains a big challenge. Heterogeneity is evident in the results for overall and progression-free survival and local recurrence[18]. If the tumor behavior can be reliably predicted at the initial diagnosis, the prognosis of CC would probably be optimally improved. Therefore, it is critical to investigate the molecular mechanism of CC and identify novel biomarkers.

At present, some gene signatures have been used in clinics, such as multigene signature panels (MSPs) to evaluate breast cancer recurrence[19], which reveals that it is a reliable molecular identification method to screen gene prognostic biomarkers by gene expression profiling. Lin *et al*/ constructed a prognostic model for CC patients undergoing radiotherapy and chemotherapy with stage IB-IV based on magnetic resonance imaging (MRI), whole-tumor apparent diffusion coefficient (ADC), human papillomavirus (HPV) genotyping[20]. Mao *et al*/ developed a 26-lncRNAs signature to predict the overall survival of CSCC[21]. In our study, we downloaded RNA-seq data, CNV data and MAF data of CC from TCGA database and divide them into training datasets (N=146) and test datasets (N=147). By analyzing the CNV data and mutation data of the training set, potential genes related to the occurrence and development of cervical cancer were obtained, meanwhile the transcriptome data was used for analysis to screen for biomarkers related to the prognosis of cervical cancer, and subsequently two sets of data were integrated. As a result, the prognostic signature model was determined through random survival forest. The test datasets and external datasets (GSE44001) further verified that 6-gene model was robust. The 6-gene signature has a higher AUC and fewer genes, which might be probably more conducive to clinical application.

In our 6-gene signature, DPAGT1, FURIN, SLC19A3 and SLC22A3 are verified to be risk factors, while CCL17 and DES are verified to be protective factors. Some previous researches have revealed the association between the 6 genes we identified with some other cancer pathogenesis. It was reported that CCL17 played a critical role in preventing tumorigenesis for its ability to engage CCR4+CD8+ T cells [22-24]. Okada et al. used RGD-fiber-mutant adenoviral vector transfected into the tumors of melanoma-bearing mice, and found that recombinant AdRGD-CCL17 induced CD8+CTL cell aggregation to tumor tissue, which effectively inhibiting tumor growth [25]. In addition, the Naoko team also used tumor-bearing animal experiments to confirm that injecting recombinant CCL17-adenovirus into the tumor of mice with colorectal cancer will induce infiltration of activated CD8+CTL and exert significant anti-tumor immune function [26]. Zhan et al. also found that the CCL17 loaded drug-eluting stents were inoculated into pancreatic cancer-bearing mice, and the locally released CCL17 recruited CCR4+CD8+ T cells to pancreatic tumor tissues, thereby inhibiting the occurrence and metastasis of pancreatic tumors [27].

Interestingly, many studies also showed that CCL17 can also stimulate cancer cell proliferation and migration [28, 29] [30, 31]. However, in present study, the expression of CLL17 was evaluated in the cervical cancer tissue and showed lower expression was associated with poor prognosis of cervical cancer. DES encodes a muscle-specific class III intermediate filament which plays a crucial role in maintaining the structure of sarcomeres, inter-connecting the Z-disks and forming the myofibrils [32]. Study revealed that the expression of DES was significantly downregulated in gall bladder carcinogenesis according to correlated promoter hypermethylation, which indicated its potential as a candidate biomarker for gall bladder carcinogenesis [33]. DES was demonstrated that it can inhibit telomerase in prostate cancer cells [34]. Moreover, previous studies also reported that DES is a prognostic predictor and therapeutic target for colorectal cancer [35, 36]. However, there was rare DES study in cervical cancer. At present study, we also demonstrated that DES was significantly down-regulated in cervical cancer which associated with poor prognosis. DPAGT1 is an upstream regulator of E-cadherin N-glycosylation status and adheres junction's composition, and dysregulation of DPAGT1 may causes disturbances in intercellular adhesion in oral cancer[37, 38]. In 2019, the data also demonstrated that overexpression of DPAGT1 was significantly associated with a poor chemotherapeutic response in hepatitis B virus - positive hepatocellular carcinoma patients [39]. FURIN encodes a member of the subtilisin-like proprotein convertase family, which includes proteases that process protein and peptide precursors trafficking through regulated or constitutive branches of the secretory pathway. Studies revealed that FURIN is associated with the prognosis of pancreatic cancer[40], laryngeal cancer[41], and lung cancer[42]. More importantly, Zhao et al. recently proposed a five-mRNA-signature to predict recurrence and survival in cervical cancer also included the DPAGT1 and FURIN [43]. However, the difference is that they only analyzed the gene expression in cervical cancer and we integrated analysis of mRNA, CNV and Gene mutation data which were significantly related to HPV integration. In present study, we also evaluated the expression of signature genes in more cervical cancer tissue by IHC. SLC19A3 is a ubiquitously expressed transmembrane thiamine transporter. It was demonstrated that down-regulation of SLC19A3 can inhibit the breast cancer metastasis[44]. Another research revealed that aberrant SLC19A3 promoter hypermethylation in plasma may be a novel biomarker for breast and gastric cancer diagnosis[45]. SLC22A3 is a plasma integral membrane protein which located in a cluster on chromosome 6. Xin- Guan et al. reported that the expression of SLC22A3 was associated with the early development and progression of familial esophageal cancer [46]. The rs2504938 SNP in SLC22A3 significantly influenced the overall survival of pancreatic cancer patients. Moreover, SLC22A3 played an important role in the colorectal cancer development and progression[47].

We integrated multi-omics data including transcriptome, CNV data and mutation data, eventually constructed a 6-gene signature for CC prognosis. The signature was verified to be robust and stable in different databases. Furthermore, the signature was shown to be optimally independent even with influence of many clinical factors. But it is not completely reliable since we only draw the conclusion with bioinformatics analysis alone. Therefore, the further experimental verification researches would be necessary.

Conclusion

In present study, we constructed a 6-gene signature for cervical cancer prognosis based on multi-omics data. The predictive value of the signature was further tested in training datasets and external independent validation datasets by applying Kaplan–Meier survival curves. And the signature shows strong robustness and clinical independence. Therefore, the 6-gene signature can be used as a prognosis assessment for cervical cancer patients, and it may provide some guidance for clinical treatment.

Abbreviations

CC: Cervical cancer, HPV: human papillomavirus, C5CC: cervical squamous cell carcinoma, SCC: squamous cell antigen, CNV: copy number variation, MAF: mutation annotation file, TCGA: The Cancer Genome Atlas TCGA, MAF: mutation annotation file, GO: Gene Ontology, KEGG: Kyoto Encyclopedia of Genes and Genomes, IHC: Immunohistochemical staining, HR: Hazard risk, ROC: receiver operating characteristic, AUC: Area Under Curve, MRI: magnetic resonance imaging

Declarations

Ethics approval and consent to participate:

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments. (The name and affiliation of the ethics committee that approved this study: The institutional ethics committee of Peking Union Medical College Hospital, CAMS Chinese Academy of Medical Sciences, No. S-183 2019)

Consent for publication:

Not applicable

Availability of data and materials:

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests:

The authors declare that they have no competing interests

Funding:

This study was funded by the National Natural Science Foundation of China [81402140] and [CAMS Innovation Fund for Medical Sciences (CIFMS) CAMS-2017-I2M-1-002] (Shen Keng).

Authors' contributions:

HX Study design and data analysis, SHZ: IHC Experiment, Data analysis, Manuscript writing, ZXS: Data analysis, PP, YM, ZY, YJX, CDY: Clinical data provider. SK: Study design, Funding. All authors have read, edited and approved the final version of the manuscript.

Acknowledgements

Not applicable

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: A Cancer Journal for Clinicians* 2018, **68**(6):394-424.
2. Johnson CA, James D, Marzan A, Armaos M: **Cervical Cancer: An Overview of Pathophysiology and Management.** *Semin Oncol Nurs* 2019, **35**(2):166-174.
3. Walboomers. JMM, Jacobs. MV, Manos. MM, Bosch. FX, Kummer. JA, Shan. KV, Snijders. PJF, Peto. J, Mrijer. CJLM, Munoz. N: **Human Papillomavirus Is a necessary cause of invasive cervical cancer worldwide.** *journal of pathology* 1999, **189**:12-19.
4. **Integrated genomic and molecular characterization of cervical cancer.** *Nature* 2017, **543**(7645):378-384.
5. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2020.** *CA Cancer J Clin* 2020, **70**(1):7-30.
6. Shen F, Zheng H, Zhou L, Li W, Liu J, Xu X: **Identification of CD28 and PTEN as novel prognostic markers for cervical cancer.** *J Cell Physiol* 2019, **234**(5):7004-7011.
7. Mao X, Qin X, Li L, Zhou J, Zhou M, Li X, Xu Y, Yuan L, Liu Q-N, Xing H: **A 15-long non-coding RNA signature to improve prognosis prediction of cervical squamous cell carcinoma.** *Gynecologic Oncology* 2018, **149**(1):181-187.
8. Wei Q, Ma C, Zhang W, Wu Q, Liu Y, Wang C, Lao G, Yang L, Liu P: **Identification of a microRNA signature associated with survivability in cervical squamous cell carcinoma.** *Plos One* 2018, **13**(3):e0193625.
9. Muller S, Lindquist D, Kanter L, Flores-Staino C, Henriksson R, Hedman H, Andersson S: **Expression of LRIG1 and LRIG3 correlates with human papillomavirus status and patient survival in cervical adenocarcinoma.** *International Journal of Oncology* 2013, **42**(1):247-252.
10. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome Biology* 2011, **12**(4).
11. Jeremy M. G. Taylor P: **Random Survival Forests.** *Journal of Thoracic Oncology* 2011, **6**:1974-1975.

12. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *OMICS* 2012, **16**(5):284-287.
13. Li YL, Ye F, Cheng XD, Hu Y, Zhou CY, Lu WG, Xie X: **Identification of glia maturation factor beta as an independent prognostic predictor for serous ovarian cancer.** *European journal of cancer (Oxford, England : 1990)* 2010, **46**(11):2104-2118.
14. Zhang SF, Wang XY, Fu ZQ, Peng QH, Zhang JY, Ye F, Fu YF, Zhou CY, Lu WG, Cheng XD *et al*: **TXNDC17 promotes paclitaxel resistance via inducing autophagy in ovarian cancer.** *Autophagy* 2015, **11**(2):225-238.
15. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Pineros M, Znaor A, Bray F: **Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods.** *Int J Cancer* 2019, **144**(8):1941-1953.
16. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A: **Global cancer statistics, 2012.** *CA Cancer J Clin* 2015, **65**(2):87-108.
17. Menderes G, Black J, Schwab CL, Santin AD: **Immunotherapy and targeted therapy for cervical cancer: an update.** *Expert Rev Anticancer Ther* 2016, **16**(1):83-98.
18. Green JA, Kirwan JM, Tierney JF, Symonds P, Fresco L, Collingwood M, Williams CJ: **Survival and recurrence after concomitant chemotherapy and radiotherapy for cancer of the uterine cervix: a systematic review and meta-analysis.** *The Lancet* 2001, **358**(9284):781-786.
19. Bhutiani N, Egger ME, Ajkay N, Scoggins CR, Martin RCG, McMasters KM: **Multigene Signature Panels and Breast Cancer Therapy: Patterns of Use and Impact on Clinical Decision Making.** *Journal of the American College of Surgeons* 2018, **226**(4):406-412.e401.
20. Lin G, Yang L-Y, Lin Y-C, Huang Y-T, Liu F-Y, Wang C-C, Lu H-Y, Chiang H-J, Chen Y-R, Wu R-C *et al*: **Prognostic model based on magnetic resonance imaging, whole-tumour apparent diffusion coefficient values and HPV genotyping for stage IB-IV cervical cancer patients following chemoradiotherapy.** *European Radiology* 2018, **29**(2):556-565.
21. Mao Y, Fu Z, Dong L, Zheng Y, Dong J, Li X: **Identification of a 26-lncRNAs Risk Model for Predicting Overall Survival of Cervical Squamous Cell Carcinoma Based on Integrated Bioinformatics Analysis.** *DNA Cell Biol* 2019, **38**(4):322-332.
22. Inoue H, Iga M, Xin M, Asahi S, Nakamura T, Kurita R, Nakayama M, Nakazaki Y, Takayama K, Nakanishi Y *et al*: **TARC and RANTES enhance antitumor immunity induced by the GM-CSF-transduced tumor vaccine in a mouse tumor model.** *Cancer immunology, immunotherapy : CII* 2008, **57**(9):1399-1411.
23. Hirata A, Hashimoto H, Shibasaki C, Narumi K, Aoki K: **Intratumoral IFN- α gene delivery reduces tumor-infiltrating regulatory T cells through the downregulation of tumor CCL17 expression.** *Cancer gene therapy* 2019, **26**(9-10):334-343.
24. Semmling V, Lukacs-Kornek V, Thaïss CA, Quast T, Hochheiser K, Panzer U, Rossjohn J, Perlmutter P, Cao J, Godfrey DI *et al*: **Alternative cross-priming through CCL17-CCR4-mediated attraction of CTLs toward NKT cell-licensed DCs.** *Nature immunology* 2010, **11**(4):313-320.

25. Okada N, Sasaki A, Niwa M, Okada Y, Hatanaka Y, Tani Y, Mizuguchi H, Nakagawa S, Fujita T, Yamamoto A: **Tumor suppressive efficacy through augmentation of tumor-infiltrating immune cells by intratumoral injection of chemokine-expressing adenoviral vector.** *Cancer gene therapy* 2006, **13**(4):393-405.
26. Nishioka M, Tanemura A, Yang L, Tanaka A, Arase N, Katayama I: **Possible involvement of CCR4+ CD8+ T cells and elevated plasma CCL22 and CCL17 in patients with rhododenol-induced leukoderma.** *Journal of dermatological science* 2015, **77**(3):188-190.
27. Zhan Q, Shen B, Fang Y, Deng X, Chen H, Jin J, Peng C, Li H: **Drug-eluting scaffold inhibited in vivo pancreatic tumorigenesis by engaging murine CCR4(+)CD8(+) T cells.** *Colloids and surfaces B, Biointerfaces* 2017, **158**:469-473.
28. Liu LB, Xie F, Chang KK, Shang WQ, Meng YH, Yu JJ, Li H, Sun Q, Yuan MM, Jin LP *et al*: **Chemokine CCL17 induced by hypoxia promotes the proliferation of cervical cancer cell.** *Am J Cancer Res* 2015, **5**(10):3072-3084.
29. Li J, Yang C, Yang J, Zou L: **Down-regulation of CCL17 in cancer-associated fibroblasts inhibits cell migration and invasion of breast cancer through ERK1/2 pathway.** *Cancer management and research* 2019, **11**:7439-7453.
30. Al-haidari AA, Syk I, Jirström K, Thorlacius H: **CCR4 mediates CCL17 (TARC)-induced migration of human colon cancer cells via RhoA/Rho-kinase signaling.** *International Journal of Colorectal Disease* 2013, **28**(11):1479-1487.
31. Al-Haidari AA, Syk I, Thorlacius H: **HMG-CoA reductase regulates CCL17-induced colon cancer cell migration via geranylgeranylation and RhoA activation.** *Biochem Biophys Res Commun* 2014, **446**(1):68-72.
32. Hnia K, Ramspacher C, Vermot J, Laporte J: **Desmin in muscle and associated diseases: beyond the structural function.** *Cell Tissue Res* 2015, **360**(3):591-608.
33. Bhunia S, Barbhuiya MA, Gupta S, Shrivastava BR, Tiwari PK: **Epigenetic downregulation of desmin in gall bladder cancer reveals its potential role in disease progression.** *The Indian journal of medical research* 2020, **151**(4):311-318.
34. Geier R, Adler S, Rashid G, Klein A: **The synthetic estrogen diethylstilbestrol (DES) inhibits the telomerase activity and gene expression of prostate cancer cells.** *The Prostate* 2010:n/a-n/a.
35. Arentz G, Chataway T, Price TJ, Izwan Z, Hardi G, Cummins AG, Hardingham JE: **Desmin expression in colorectal cancer stroma correlates with advanced stage disease and marks angiogenic microvessels.** *Clinical proteomics* 2011, **8**(1):16.
36. Y W, Y L, Z C, T W, J G, X W, Y Y, M W, Z P: **The Evaluation of Colorectal Cancer Risk in Serum by anti-DESMIN-conjugated CdTe/CdS Quantum Dots.** *Clinical laboratory* 2017, **63**(3):579-586.
37. Nita-Lazar M, Noonan V, Rebustini I, Walker J, Menko AS, Kukuruzinska MA: **Overexpression of DPAGT1 Leads to Aberrant N-Glycosylation of E-Cadherin and Cellular Discohesion in Oral Cancer.** *Cancer Research* 2009, **69**(14):5673-5680.

38. Jamal B, Sengupta PK, Gao ZN, Nita-Lazar M, Amin B, Jalisi S, Bouchie MP, Kukuruzinska MA: **Aberrant amplification of the crosstalk between canonical Wnt signaling and N-glycosylation gene DPAGT1 promotes oral cancer.** *Oral Oncol* 2012, **48**(6):523-529.
39. Li R, Xu T, Wang H, Wu N, Liu F, Jia X, Mi J, Lv J, Gao H: **Dysregulation of the miR-325-3p/DPAGT1 axis supports HBV-positive HCC chemoresistance.** *Biochem Biophys Res Commun* 2019, **519**(2):358-365.
40. ZHOU. YZM, WEI. H, ZHOU. H, HE. J, LU. Y: **Furin promotes epithelial-mesenchymal transition in pancreatic cancer cells via Hippo-YAP pathway.** *INTERNATIONAL JOURNAL OF ONCOLOGY* 2017, **50**:1352-1362.
41. Lee M, Ryu CH, Chang HW, Kim GC, Kim SW, Kim SY: **Radiotherapy-associated Furin Expression and Tumor Invasiveness in Recurrent Laryngeal Cancer.** *Anticancer Research* 2016, **36**(10):5117-5126.
42. Bassi DE, Zhang J, Renner C, Klein-Szanto AJ: **Targeting proprotein convertases in furin-rich lung cancer cells results in decreased in vitro and in vivo growth.** *Molecular Carcinogenesis* 2017, **56**(3):1182-1188.
43. Ju M, Qi A, Bi J, Zhao L, Jiang L, Zhang Q, Wei Q, Guan Q, Li X, Wang L *et al*: **A five-mRNA signature associated with post-translational modifications can better predict recurrence and survival in cervical cancer.** *J Cell Mol Med* 2020, **24**(11):6283-6297.
44. Shin. IWCVY, Siu. MT, Tsang. JY, Ho. JC, Chen. J, Tse. GM, Wang. X, Kwong. A: **Association of EP2 receptor and SLC19A3 in regulating breast cancer metastasis.** *Am J Cancer Res* 2015, **5**(11):3389-3399.
45. Ng EK, Leung CP, Shin VY, Wong CL, Ma ES, Jin HC, Chu KM, Kwong A: **Quantitative analysis and diagnostic significance of methylated SLC19A3 DNA in the plasma of breast and gastric cancer patients.** *PLoS One* 2011, **6**(7):e22233.
46. Fu L, Qin YR, Ming XY, Zuo XB, Diao YW, Zhang LY, Ai J, Liu BL, Huang TX, Cao TT *et al*: **RNA editing of SLC22A3 drives early tumor invasion and metastasis in familial esophageal cancer.** *Proceedings of the National Academy of Sciences of the United States of America* 2017, **114**(23):E4631-e4640.
47. Ren A, Sun S, Li S, Chen T, Shu Y, Du M, Zhu L: **Genetic variants in SLC22A3 contribute to the susceptibility to colorectal cancer.** *International Journal of Cancer* 2019, **145**(1):154-163.

Tables

Table 1. Top 20 most significant prognostic genes

gene	HR	coefficient	z socre	p value
ENSG00000135917	1.748677971	0.558860057	4.436721804	9.13E-06
ENSG00000163421	1.682368397	0.520202561	4.376242475	1.21E-05
ENSG000000000005	1.513401732	0.41435992	4.361855389	1.29E-05
ENSG00000162604	1.501643857	0.406560412	4.099799021	4.14E-05
ENSG00000196468	1.673950164	0.515186201	3.960270954	7.49E-05
ENSG00000164588	1.537245105	0.429991922	3.940550485	8.13E-05
ENSG00000117335	2.46589333	0.902554147	3.865969034	0.000110649
ENSG00000172023	1.529686885	0.425063064	3.797280694	0.000146292
ENSG00000113327	1.513257567	0.414264656	3.784384898	0.000154089
ENSG00000226490	1.444495797	0.367760331	3.772939226	0.000161336
ENSG00000185730	1.973737202	0.679928802	3.734920178	0.000187775
ENSG00000214686	1.424691673	0.35395542	3.729554952	0.000191818
ENSG00000162736	2.603303699	0.956781292	3.723143696	0.000196757
ENSG00000204007	1.472298714	0.386824931	3.696985142	0.000218175
ENSG00000214107	1.557528566	0.443100312	3.690028732	0.000224229
ENSG00000118702	11.76746014	2.465338107	3.641055676	0.000271522
ENSG00000204979	1.544184328	0.434495828	3.61125427	0.00030472
ENSG00000197275	1.563705433	0.447058282	3.602161044	0.000315583
ENSG00000168843	1.379657366	0.321835183	3.599920612	0.000318314
ENSG00000180708	1.388555784	0.328264203	3.560948718	0.000369517

Table 2. 6-genes significantly associated with the overall survival in the training-set patients

Ensembl Gene ID	Symbol	HR	Z-score	P value	Importance	Relative Importance
ENSG00000102970	CCL17	0.44	-3.100495	1.93E-03	0.0193	1
ENSG00000135917	SLC19A3	1.75	4.436722	9.13E-06	0.0185	0.9583
ENSG00000140564	FURIN	1.53	2.163624	3.05E-02	0.0117	0.6042
ENSG00000175084	DES	0.52	-2.476295	1.33E-02	0.0105	0.5417
ENSG00000146477	SLC22A3	1.59	2.498738	1.25E-02	0.0101	0.5208
ENSG00000172269	DPAGT1	1.65	2.65869	7.84E-03	0.0097	0.5

Table 3. univariate and multivariate Cox regression analysis to verify the independence of 6-gene signature in test datasets.

Variables	Univariate analysis			Multivariable analysis		
	HR	95%CI of HR	P value	HR	95%CI of HR	P value
Validation cohort,TCGA test datasets, GSE44001						
TCGA test datasets						
6-gene risk score						
Low risk group	1(reference)			1(reference)		
High risk group	1.25	1.04-1.49	0.016	1.444	1.05-1.97	0.020
Age	1.01	0.98-1.03	0.463	0.988	0.94-1.02	0.560
Grade 1	1(reference)			1(reference)		
Grade 2	1.37	0.31-5.89	0.674	1.340	0.14-12.11	0.794
Grade 3	1.51	0.34-6.67	0.59	1.12	0.12-9.88	0.919
Pathologic T 1	1(reference)			1(reference)		
Pathologic T 2	0.97	0.37-2.49	0.943	0.17	0.017-1.72	0.135
Pathologic T 3	0.78	0.17-3.40	0.742	1.150	0.20-6.58	0.875
Pathologic T 4	11.68	3.20-42.48	1.92E-04	231.567	7.16-7483.09	0.002
Pathologic N 0	1(reference)			1(reference)		
Pathologic N 1	3.48	1.5- 8.06	0.004	10.266	3.30-31.88	0.000
Pathologic M 0	1(reference)			1(reference)		
Pathologic M 1/ M X	1.90	0.91-3.94	8.50E-02	1.341	0.48-3.69	0.570
Tumor stage I	1(reference)			1(reference)		
Tumor stage II	0.73	0.29-1.801	0.498	7.57	0.49-114.99	0.145
Tumor stage III/IV	1.31	0.60-2.81	0.49	0.446	0.05-3.92	0.467
GSE44001						
6-gene risk score						
Low risk group	1(reference)			1(reference)		
High risk group	1.40	1.09-1.79	0.007	1.40	1.09-1.78	0.007
Tumor stage I	1(reference)			1(reference)		

Figures

Figure 1

A. significant amplified segments in cervical cancer genome. B. significant deleted segments in cervical cancer genome. C. the top 50 genes with most significant p value. The upper histogram represents the total number of synonymous and nonsynonymous mutations in each patient's 50 genes, and the right histogram represents the mutation proportion in all samples of 50 genes. D. 839 genes with significant copy number variation and mutation involved in the KEGG pathway. E. 839 genes with significant copy number variation and mutation involved in biological process (GO bp). F. the relationship between prediction error rates and the number of separate trees. G. the out-of-bag importance rank of the top 6 genes.

Figure 2

A. Kaplan-Meier survival curve of the 6-gene signature in TCGA training datasets. B. ROC curve and AUC of 6-gene signature. C. risk score, survival period and gene expression level.

Figure 3

A. Kaplan-Meier survival curve of the 6-gene signature in TCGA test datasets. B. ROC curve and AUC of 6-gene signature. C. risk score, survival period and gene expression level.

Figure 4

A. Kaplan-Meier survival curve of the 6-gene signature in independent validation datasets. B. ROC curve and AUC of 6-gene signature. C. risk score, survival period and gene expression level.

Figure 5

Immunohistochemistry for CCL17, DES, DPAGT1, FURIN, SLC19A3 and SLC22A3. Samples of cervical cancer-adjacent tissue (N=30) and cervical cancer (N=130). Cancer-adjacent cervical tissue samples with weak immunostaining scores for DPAGT1 (G), FURIN (J), SLC19A3 (M), SLC22A3 (P) and strong immunostaining scores for CCL17 (A), DES (D). Cervical cancer sample of weak and strong immunostaining score for CCL17 (B, C), DES (E, F), DPAGT1 (H, I), FURIN (K, L), SLC19A3 (N, O) and

SLC22A3 (Q, R), respectively. Expression of the CCL17, DES, DPAGT1, FURIN, SLC19A3 and SLC22A3 genes were depicted in S slides. (X 100)

Figure 6

Overall (OS) and disease-free (DFS) survival curves for cervical cancer (N=130) according to CCL17 (A, B), DES (C, D), DPAGT1 (E, F), FURIN (G,H), SLC19A3 (I, J) and SLC22A3 (K, L) gene expression status (low or high). Gene expression status was divided according to their median values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryS6.docx](#)
- [SupplementaryS5.docx](#)
- [SupplementaryS4.xlsx](#)
- [SupplementaryS3.xlsx](#)
- [SupplementaryS2.xlsx](#)
- [SupplementaryS1.docx](#)