Identification of a Microbial-Related Gene Signature Indicative of Disease Prognosis in Cervical Cancer

Yue Qi  
Tianjin Medical University

Xuanyan Li  
Tianjin Central Hospital Gynecology Obstetrics

Wenwen Zhang  
Tianjin Central Hospital Gynecology Obstetrics

Xinyu Li  
Tianjin Medical University

Pengpeng Qu (✉ qu.pengpeng@hotmail.com)  
Tianjin Medical University

Research Article

Keywords: Microorganisms, Cervical cancer, SALL3, GABRP, Prognostic model

Posted Date: April 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2786491/v1

License: Ⓒ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Cervical cancer (CC) is the fourth most common female malignancy and the fourth leading cause of cancer death worldwide. There is an urgent need to study the underlying mechanisms of the malignant biological behavior of CC, identify new prognostic markers, and develop individualized treatment strategies to improve the survival rate of patients with CC. Much research in recent years suggested that cervical cancer was directly linked to the makeup of the vaginal microbiota and HPV infection. Firstly, we examined the relationship between microbes and the transcriptome in the development of CC. Secondly, we compared and contrasted the microbiota of cancer patients with cervical squamous cell carcinoma and adenocarcinoma (CESC) to determine their similarities and differences. Thirdly, we used correlation analysis to verify the correlation between key genes and microbes and construct a prognostic model. Finally, GO and KEGG enrichment analyses were conducted to reveal potential mechanisms and verified the accuracy by GSE4001. We found that Achromobacter, Natronomonas, and Nafulsella were identified in early and late CC patients. In addition, 16 microbial-related genes were obtained by correlation analysis. The results showed that SALL3 and GABRP had a significant correlation between the three microbes and affect overall survival in CC. The prognostic model constructed by key genes was verified accurately. Achromobacter, Natronomonas, and Nafulsella may play an important role in CC progression. Besides, SALL3 and GABRP may influence oxygen transport and metabolic pathways, which affect tumor outcome and prognosis. We hope our study could provide a theoretical reference for further research on the mechanism of the microbial influence on transcriptome genes in the progression of CC.

1. Background

Cervical cancer (CC) is one of the most prevalent tumor types among gynecological cancers. Global cancer statistics showed that cervical cancer accounted for an estimated 604,000 new cases and 342,000 deaths worldwide in 2020(1). Infection with high-risk human papillomavirus (hrHPV) is a main risk factor linked to the development of cervical cancer. Although the incidence rates have displayed a declining trend recently due to improved screening strategies and HPV vaccination, CC continues to be the main cause of cancer-related mortality among women in poor living conditions (2). For effective treatments such as surgery, the 5-year survival rate of primary CC can reach approximately 75–85%, while the 5-year survival rate of recurrent, persistent, and metastatic CC is only 15%(3). Due to the lack of effective molecular markers, it is difficult to effectively monitor the progression of CC, so there is an urgent need for molecular labelling to monitor progression and improve the survival rate. According to recent studies, the vaginal microbiota may play a significant role in modifying high-risk HPV infection, and a variety of microbiota species may act as sensors for monitoring changes in the cervical microenvironment brought about by hrHPV infection(4). Therefore, there is an urgent need to identify new prognostic markers between the vaginal microbiota and HPV that may provide a theoretical basis for the subsequent treatment of CC.

The human body contains more than 300 million microorganisms, including bacteria, viruses, and fungi(5). Furthermore, next-generation sequencing (NGS) revealed that previously thought to be sterile
organs and tissues, such as the lung, prostate, bladder, breast, liver, and pancreas, may actually retain low-biomass microbial populations(6). These organisms keep our body functioning normally by offering a nutrient-rich microenvironment in exchange for assistance with digestion and metabolism. In addition, studies have shown that specific microorganisms can have pleiotropic interactions in the process of tumor development. Microorganisms have a significant influence on tumorigenesis by their expression and secretion of virulence factors, physical binding-induced signaling, and the recruitment of immune cells(7). Recently, an increasing number of studies have demonstrated that CC is directly connected with the vaginal microbiota and HPV infection (8, 9). In addition, studies have shown that Lactobacillus iners is also a common vaginal microorganism associated with viral infections(10). Studies have demonstrated that the abundances of the vaginal microbes L. gasseri and Atopobium are correlated with HPV clearance and persistence, respectively(11). Different microorganisms may have different effects on the clearance or persistence of HPV infection and the occurrence and development of cervical cancer. In our study, by using multiple bioinformatics methods, we aimed to explore the new prognostic markers and underlying molecular mechanisms of microorganisms in the pathogenesis and clinical prognosis of CC. In addition, we aimed to identify the key microorganisms in the progression of CC and to identify the key genes associated with CC prognosis.

2. Results

2.1 Microbiome analysis of the early- and late-stage groups

In this study, we distinguished between the “late-stage group” and the “early-stage group” based on the extent of the tumor’s spread beyond the cervix. In Fig. 1A, we display the community composition of the microbiota at the phylum level between the early-stage and late-stage groups. The most common microorganisms in the two groups were the same, namely, Proteobacteria at the phylum level. In addition, Actinobacteria, Firmicutes, and Bacteroidetes were the predominant phyla in tumor specimens. Furthermore, we analyzed the main microbial community at the genus level (Fig. 1B). The early-stage group specimens were rich in Terrabacter, Bacteroides, Proteus, and Shigella. Late-stage group specimens had an abundance of Terrabacter, Neisseria, Bacteroides, Proteus, and Shigella. Because we focused on tumor tissue, the main microbial species within a genus were very similar, but the abundances were significantly different. Therefore, we conducted alpha diversity analysis based on the genus level, and the results revealed that the difference in the Shannon index was not statistically significant (P = 0.078); the difference in the Simpson index was also not statistically significant (P = 0.1) (Fig. 1C). To further evaluate the different microbiomes in early- and late-stage CC patients, we examined the β diversity and presented the results of the principal coordinate analysis (PCoA) as a plot. The results revealed significant differences between the cervical microbiomes of the early-stage and late-stage groups (P = 0.019) (Fig. 1D). To verify the accuracy of the key bacteria, we used three methods (LEfSe analysis, random forest classification, and classical univariate statistical comparisons) and took the intersection. Figure 1E displays the LDA score coupled with LEfSe between the early-stage and late-stage groups. In the LEfSe analysis, the Kruskal–Wallis test was used to test the hypothesis that a p value < 0.5 was
considered significant. When the LDA score was > 3, this bacteria was considered an important contributor. A total of seventeen microorganisms were identified that had a significant association with early- and late-stage patients. Thirteen microorganisms that contributed most significantly to the grouping were identified by random forest classification (Fig. 1F). We displayed the most significant microorganisms in Supplemental Table 3. Based on the results, we conducted classical univariate statistical comparison analysis and found that 11 microorganisms had remarkable significance (Fig. 1G). Next, we identified three key microorganisms, that is, Achromobacter, Natronomonas, and Nafulsella, which may play an important role in CC development (Fig. 1H).

2.2 Correlation between key bacteria and DEGs

To investigate the DEGs between the early-stage and late-stage groups, a volcano map was constructed that revealed 26 DEGs, including 11 upregulated and 15 downregulated genes (Fig. 2A). The heatmap displays the DEGs identified by cluster analysis (Fig. 2B). In the supplementary plot (Figure S1), boxplots of distinct gene expression are shown. Sixteen of 26 DEGs were found to be associated with microbes by Pearson’s correlation-based analysis with the three key bacteria. Figure 2C displays the correlation between the three key bacteria and the 26 DEGs. The results showed that eight genes had a significant association with those bacteria (**P < 0.01). Furthermore, GO and KEGG functional enrichment analyses were performed to determine the biological features of these 16 DEGs. GO functional enrichment analysis revealed 14 terms (FDR < 0.05) across the BP, CC, and MF categories (Fig. 2D). DEGs (correlation-bacteria) were markedly enriched in oxygen transport and hydrogen peroxide catabolic process and bicarbonate transport in the BP category. In the cellular component category, haemoglobin complex and endocytic vesicle lumen were enriched. Therefore, we conclude that these 16 genes may play an important role in tumor development, especially in oxygen metabolism processes. In addition, several genes were mainly enriched in the vesicle and secretory granule terms. In addition, these genes were involved in cancer metabolism-associated pathways, including arachidonic acid metabolism, the glycolysis/gluconeogenesis signaling pathway, pyruvate metabolism, threonine metabolism, and tyrosine metabolism (Fig. 2E). ABC transporters, which are the largest of all the protein families, have a variety of physiological functions. Overall, the DEGs were enriched in the metabolism pathway and certain cancer-related pathways. Metabolic pathways and cancer-related pathways may influence tumor outcome and prognosis. According to the above results, we inferred that these DEGs may play an important role in metabolic pathways and then affect the occurrence and development of tumors.

2.3 The identification of microbial-related gene biomarkers

To explore the microbial-related genes, we next constructed a microbial-related gene signature through log-rank test survival analysis based on the expression of those 16 genes (DSG1, KIF1A, HBA2, HBB, HBA1, SALL3, AGR2, C10orf81, NWD1, ABCA13, CYP2B7P1, ADH1C, ALOX1S, GABRP, LOXL4, and PGR) (Figure S2). Next, we analyzed the relationship between DEGs and prognosis and found that SALL3 and GABRP expression had a noticeable impact on prognosis (P < 0.05) (Fig. 3A). Table 4 shows the coefficient value of each gene and the risk score of each sample. Then, we determined the risk scores of each sample and patients who were categorized into two groups with different risks in accordance with
the median value (Fig. 3B). Interestingly, we discovered that patients in the low-risk group had a much higher overall survival rate than that of patients in the high-risk group (Fig. 3C). ROC analysis revealed that the areas under the curves (AUCs) used to predict the 3- and 5-year survival rates of patients with CESC were 0.7 and 0.63, respectively (Fig. 3D). We assessed the predictive capabilities of the microbial-related gene signature in the validation sets. In the test set, we performed the same analysis. In a similar manner, the KM curve examined during the validation produced outcomes that matched those of the test set. The prognosis was better for the low-risk patients (Fig. 3E). Finally, we used the GEPIA database to assess the expression levels of the two prognostic genes (Fig. 3F). The expression of SALL3 varied between the different stages (P < 0.05). The expression trend of GABRP was consistent with our training set, even though there was no statistically significant difference between the various phases.

2.4 Construction of the prediction model for clinical prognosis

Some clinical features were used for univariate survival analysis, such as age, weight, and stage (Figure S3). The outcomes demonstrated that survival time could be considerably influenced by weight and FIGO stage. Furthermore, the multivariate Cox analysis results demonstrated that both the FIGO stage [hazard ratio (HR): 4.3093; 95% confidence interval (CI): 1.5545–11.946] and risk score (HR: 2.5903; 95% CI: 1.2901–5.201) were independent prognostic factors for the OS of CESC patients (Fig. 4A). We attempted to create a nomogram by merging all microbial-related gene signatures and clinical traits in the full set to better utilize the microbial-related gene signature (Fig. 4B). The performance of the nomogram was assessed using the AUC index of the ROC analysis and calibration curve. The AUCs of the nomogram model for predicting 1-, 3-, and 5-year overall survival rates were 0.83, 0.73, and 0.63, respectively (Fig. 4C). The calibration curve for the probability of 3- and 5-year OS demonstrated good consistency between the actual observed survival and the nomogram-based prediction (Fig. 4D).

3. Discussion

Our body harbours an estimated three trillion bacterial members that orchestrate a comprehensive interplay of physiological processes and disease susceptibilities (12) (13). Particularly in colorectal malignancies, these microbial organisms exercise their actions mostly through indirect channels (including metabolites and the immune system) on distant or proximal tumor tissues (14). Gynecological cancers were also reported to be associated with microbiome formation (15).

The microbiome of our study was predominantly composed of Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes at the phylum level in tumor specimens. This result can be explained by the female gut-vaginal microbiome “cross-talk” (16). Lactobacillus sp. is considered to be the principal cervical-vaginal microbiome in healthy conditions (17). Based on our results, we discovered that the diversity of the cervical flora steadily increased as the cervical lesions progressed, whereas the Lactobacillus abundance gradually declined. However, the alpha diversity analysis of the samples was not statistically significant. The tumor microenvironment may play a role by providing a space for similar microbial
populations to invade and flourish even in different tissue types (18). The β diversity was statistically distinct between the early-stage and late-stage groups. The results demonstrated that the variety and number of microbes in the various groups differed significantly from one another. Next, we uncovered three microbiota (Achromobacter, Natronomonas, and Nafulsella) from the intersection of those three approaches that may play an important role in CC clinical development. Although these three microbiota were not the most numerous, they made the largest contribution to our classification.

Achromobacter is a genus of bacteria in the family Alcaligenaceae and the order Burkholderiales(19). They have been identified in persons with specific immunosuppressive diseases, such as cystic fibrosis, cancer, and kidney failure, as opportunistic human pathogens(20). In the Human Body Map, which is a database of bacteria found in the human body and their relationships to illnesses and health, Achromobacter was connected to papillomavirus infections, endometrial neoplasms, and ovarian neoplasms when they first developed in the cervix. This genus was connected to colorectal neoplasms when it was discovered in the intestine. Achromobacter xylosoxidans is the main species of this genus. It was thought that this bacterium could cause sepsis in people with impaired immune systems. This pathogen releases toxins through three separate secretion mechanisms to infect its target (21). This is also consistent with the secreted proteins and transporter functions of our related genes. Natronomonas pharaonis is an aerobic, extremely haloalkaliphilic bacterium. Additionally, in the Human Body Map, it is prevalent in the fallopian tubes and vagina and is connected to vaginitis. It is not present in the fallopian tubes of healthy women. Nafulsella is a gram-negative, rod-shaped, gliding, pale-pink-pigmented bacterium. There is currently less research on this bacterium in the disease. A limitation of this study was that it was not based on the results of metagenomic sequencing. Therefore, strain-level gene and functional annotation, comparative genomic analysis, evolutionary analysis, and single bacterial assembly could not be conducted. There are few published studies related to these three microorganisms and cervical cancer, but through our review of the Human Body Map, we found that these three microorganisms are related to HPV infection. Based on the above results, we suspect that these microorganisms may play a role in the progression of cervical cancer by secreting substances that affect the host’s metabolism transcriptome genes and immunity. Certainly, more research is needed to explore the mechanism in the future.

More proof that bacteria and transcriptome genes do indeed regulate each other available. Next, we looked for prognostic genes from the perspective of microorganisms by Pearson correlation coefficient and log-rank survival analysis in this study. Pearson correlation analysis proved that these genes were also related to microbial changes. Those genes are based on early-/late-stage differential genes, which could also indirectly prove that they play a role in cancer progression. The patient’s prognosis was then linked to two genes. Spalt-like transcription factor 3 (SALL3) encodes a C2H2-type zinc-finger protein in a family of evolutionarily conserved genes present in species(22). SALL3 mRNA levels were lower in HPV-positive cervical cancer tissues than in HPV-negative cervical cancer tissues, and HPV infection was positively correlated with hypermethylation of the SALL3 promoter(23). Gamma-aminobutyric acid (GABA) operates as the main inhibitory neurotransmitter in the mature mammalian central nervous system through the activation of GABA receptors(24). In some cancers, GABRP is linked to cancer
migration and proliferation\(^{(25)},(26)\). Analysis of those tumor microbiome genes revealed that they were highly enriched in cell secretion, oxygen transport, and metabolic pathways. Microorganisms require the control of oxygen transport since they exist in an anaerobic environment. Some genes are abundant in endocytic vesicles and function to secrete factors, which is consistent with the manner in which we examined Achromobacter activity. To fulfill the increased bioenergetic and biosynthetic demand as well as to reduce the oxidative stress necessary for cancer cell growth and survival, cancer cells independently modify their flow through several metabolic pathways\(^{(27)}\). The KEGG gene functions are enriched in retinol metabolism, eicosanoid acid remodelling, and fatty acid breakdown. Inhibiting tumor growth and proliferation, retinol metabolism and distribution play a significant role in the development and evolution of tumors \(^{(28)}\). Eicosanoid acid reshapes the tumor microenvironment. Large amounts of ATP are produced during fatty acid breakdown, and this energy source is crucial for activating key tumor immune cells such as CD8 T lymphocytes. It is believed that fatty acid breakdown will strengthen the immunological memory impact \(^{(28)}\). The results of the above studies on gene function are consistent with the mechanism of microbial action, which also proves that our Pearson correlation analysis between microbes and genes is appropriate.

The results of gene validation in the GEPIA website showed that the expression level was related to the stage, and the trend was the same as in our study. This could also demonstrate that these genes play a key role in cervical cancer progression. Therefore, it is likely that the microbes we discovered secreted chemicals that influenced the methylation or expression of transcriptome genes, controlling metabolism and encouraging tumor growth. Here, we hypothesize that microbes may play a role in cancer progression by influencing these two genes or that cancer progression affects the microbiome by altering the status of these genes. We constructed and validated a prognostic model of microbial-associated genes, which is a preindependent prognostic factor for cervical cancer. This opens up new avenues for future research into the mechanisms and prognosis of cervical cancer progression.

### 4. Conclusions

Our study examined the relationship between microbes and the transcriptome in the development of cervical cancer by bioinformatics analysis. We found that different microbes varied in the CC progression process and identified three major microbiota that were associated with CC. Some microbial-related genes were obtained by correlation analysis. The KEGG/GO results revealed potential mechanisms between bacteria and tumor genes, which can also be used in our follow-up study. Second, two microbial-related genes were used to construct a prognostic model. Because the cervix is identical to the outside world, it is easier to obtain secretions or tissue samples for sequencing to identify relevant genes and microorganisms. Therefore, the prognostic model we constructed can provide a basis for the preoperative prognostic risk of patients achieving individualized treatment for different prognostic risks. The causal relationship between the microbes and cancer progression will be our follow-up research target.

### 5. Methods
5.1 Data download and processing——Cervical squamous cell carcinoma and adenocarcinoma (CESC) dataset

The Cancer Genome Atlas cervical cancer microbiome (TCGA-CESC-microbiome) and clinical data were downloaded from cBioPortal (https://www.cbioportal.org/). A summary of the patient clinical characteristics is shown in Table 1. Whole-transcriptome sequencing analysis was performed on 291 CESC samples, which included 1074 genera (Table 2). The RNA sequencing datasets were downloaded from cBioPortal. By using the GDC website, we processed the raw gene expression, converted the probe IDs into gene symbols from the corresponding platform annotation profile and normalized the raw matrix data with log2 conversion.

5.2 Microorganism analysis of the early-stage and late-stage groups

The samples were divided into the early-stage and late-stage groups by the TNM stage (2017 AJCC Eighth edition). Patients at the T1N0M0 stage comprised the early-stage group. The remaining patients comprised the late-stage group (T2NanyMany). Alpha diversity, which refers to the species diversity in a specific area or ecosystem, is frequently assessed using Shannon's and Simpson's indices. Shannon's index is correlated with diversity, while Simpson's index is inversely proportional to diversity. Beta diversity analysis is a comparative analysis between groups of species diversity among different ecosystems or microbial communities to obtain similarities or differences in community composition among different grouped samples. To identify the important microorganisms, we combined our three methods for finding key microorganisms (the linear discriminant analysis effect size (LEfSe), random forest and univariate “classification” approaches). The common microorganisms obtained by the three methods were taken as the key microorganisms. Key microorganisms were used for subsequent analysis. Microbiome analysis was conducted by the microbiomeanalyst.ca website (https://www.microbiomeanalyst.ca).

5.3 DEG identification

After the data were standardized, the “limma” package was used to identify differentially expressed genes (DEGs) between the early-stage and late-stage groups in R statistical software (version 3.6.3). DEGs with \( P < 0.05 \) and \(| \log_2 FC | > 1\) were considered statistically significant. The volcano plots and heatmaps were drawn using the “ggplot2” and “Pheatmap” packages. Independent Student's \( t \) test was applied to evaluate the difference in \( \log_2 FC \) values calculated by GFOLD between the early-stage and late-stage groups.

5.4 Correlation analysis between key microbes and DEGs

The Pearson correlation test was used to analyse the correlation between the key bacteria and DEGs by the “VGAM” package. A total of 3/26 bacteria-gene pairs were calculated. The significance of each bacteria-gene pair was determined based on an adjusted \( P \) value < 0.05.

5.5 Functional enrichment analysis
The “ClusterProfiler” package was used for Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. The reference gene list was c2.cp.Reactome.v7.0.symbols. GMT was downloaded from the Molecular Signatures Database (MSigDB). GO analysis included the biological process (BP), cellular component (CC), and molecular function (MF) categories. The significance threshold of GO analysis was set at an adjusted P value < 0.05, and KEGG analysis was set as a P value < 0.05.

5.6 Microbial-related gene biomarker identification

To identify key prognostic genes associated with microbes, we performed batch log-rank survival analysis to estimate the relationship between microbial-related genes individually and the overall survival (OS) of patients (P < 0.05). By using gene expression values and the related Cox coefficient, a microbial-related gene signature was established. The algorithm of the risk score value for one patient with CESC was as follows: \( \text{Risk score} = \sum \text{exp} \cdot \text{cof} \). The risk scores of patients with CESC in the entire set were calculated using the aforementioned algorithm. Patients with CESC in the entire set were divided into the high-risk and low-risk groups based on the optimal threshold of risk scores. Ultimately, Kaplan–Meier (KM) and time-dependent ROC curves were utilized to measure the reliability of the microbial-related gene signature. The test set GSE4001, including 300 samples, was used to validate the stability of the risk signature. To standardize the RNA sequencing data from the TCGA and GTEx datasets, we used the GEPIA website (http://gepia.cancer-pku.cn/). The expression levels of two prognostic genes in GEPIA have been verified at various stages [stage I, stage II, stage III, and stage IV].

5.7 Correlation between the prediction model and clinical characteristics

Univariate survival analysis was performed for clinical characteristics, including age, stage, and weight. To verify the independent factors to predict patient prognosis, we combined clinical data (age, stage, and weight) for analysis in both univariate and multivariate Cox regression models. A nomogram including age, stage, and risk score was used to calculate the total score. We assessed the 1-, 3-, and 5-year survival probabilities based on the nomogram. The area under the curve (AUC) was determined to assess the performance of the nomogram.

5.8 Statistical methods

Data analysis was performed based on R version 4.1.1. The log-rank test was used to test the difference in overall survival between the early-stage and late-stage groups. A hazard ratio (HR) and a 95% confidence interval (CI) were evaluated by univariable and multivariate Cox regression models. A statistical value < 0.05 indicates that the difference is statistically significant (∗ value < 0.05, ∗∗ value < 0.01, ∗∗∗ value < 0.001).

Abbreviations
<table>
<thead>
<tr>
<th>Full name</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cervical cancer</td>
<td>CC</td>
</tr>
<tr>
<td>Human papilloma virus</td>
<td>HPV</td>
</tr>
<tr>
<td>Cervical squamous cell carcinoma and adenocarcinoma</td>
<td>CESC</td>
</tr>
<tr>
<td>Gene ontology</td>
<td>GO</td>
</tr>
<tr>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
<td>KEGG</td>
</tr>
<tr>
<td>High-risk human papillomavirus</td>
<td>hrHPV</td>
</tr>
<tr>
<td>Next-generation sequencing</td>
<td>NGS</td>
</tr>
<tr>
<td>The cancer genome atlas</td>
<td>TCGA</td>
</tr>
<tr>
<td>Genomic data commons</td>
<td>GDC</td>
</tr>
<tr>
<td>Tumor node metastasis</td>
<td>TNM</td>
</tr>
<tr>
<td>Linear discriminant analysis effect size</td>
<td>LEfSe</td>
</tr>
<tr>
<td>Differential gene expression</td>
<td>DEGs</td>
</tr>
<tr>
<td>Biological process</td>
<td>BP</td>
</tr>
<tr>
<td>Cellular component</td>
<td>CC</td>
</tr>
<tr>
<td>Molecular function</td>
<td>MF</td>
</tr>
<tr>
<td>Overall survival</td>
<td>OS</td>
</tr>
<tr>
<td>Kaplan–Meier</td>
<td>KM</td>
</tr>
<tr>
<td>Receiver operating characteristic</td>
<td>ROC</td>
</tr>
<tr>
<td>Area under the curve</td>
<td>AUC</td>
</tr>
<tr>
<td>Hazard ratio</td>
<td>HR</td>
</tr>
<tr>
<td>Federation International of Gynecology and Obstetrics</td>
<td>FIGO</td>
</tr>
</tbody>
</table>

**Declarations**

**Ethics approval and consent to participate**

The TCGA and GEO datasets are publicly available, no ethical approval is required.

**Consent for publication**

Not applicable.
Availability of data and materials

All datasets generated for this study are included in the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the Tianjin Science and Technology Plan 15 Project (No.19YFZCSY00600) and the Beijing-Tianjin-Hebei Basic Research 16 Cooperation Project (No. J200009 20JCZXJC00010).

Authors' contributions

Y.Q.: original draft writing. XY.L.: draft review and editing. WW.Z.: picture presentation. XY.L.: conceptualization. PP.Q.: supervision. All authors reviewed the manuscript.

Acknowledgements

We would like to thank TCGA network and GEO dataset. We also thank Dr. Piao from Nankai University school of medicine, Nankai University, Tianjin, China, for idea of this manuscript.

Authors' information

YQ,XYL,PPQ:
Clinical School of Obstetrics and Gynecology Center, Tianjin Medical University, Tianjin, China

YQ,, XYL, WWZ,XYL,PPQ:
Department of Gynecological Oncology, Tianjin Central Hospital Gynecology Obstetrics, Tianjin, China

XYL, PPQ:
Nankai University school of medicine, Nankai University, Tianjin, China

Correspondence: PPQ

References


Tables

Tables 1 to 4 are available in the Supplementary Files section.

Figures
Figure 1

**Microorganisms analysis of the early- and late groups of cervical cancer.**

(A) Composition of the microbiota at the phylum level in patients with different groups (B) core microbiota at the genus level in patients with different groups. (C) Alpha diversity analysis of samples in the early and late dataset. (D) PCoA based on unweighted UniFrac distances between the early-stage and
late-stage groups (p < 0.05). (E) Bar plot shows taxa with LDA score > 3.0 from the order to the genus level. (F) A random forest "classification" approach was used to find key bacteria associated with groupings. (G) 11 key microbiota's boxplot from Classical Univariate Statistical. (H) A Venn diagram illustrates that 3 microbiota out of 41 microbes were shared between the Lefse green circle, Random-forest (blue circle), and Classical Univariate Statistical Comparisons (pink circle).
**Correlation between key bacteria and differently expressed genes**

(A) A volcano plot of the DEGs between early-stage and late-stage groups of CESC samples. (B) The heatmap of intersection DEGs. (C) Pearson correlation coefficients of 21 genes and 3 key bacteria. P value evaluated by Student's t-test for comparing the difference in log2FC values calculated by GFOLD between early-stage and late-stage groups. Red, positive; Blue, negative. (D-E) GO and KEGG pathway enrichment analysis Microbial-Related Genes.
Figure 3

**Microbial-Related Gene biomarker identification**

(A) Paired sample expression analysis and survival analysis for the 2 key genes. (B) The distribution and median value of the risk scores of patients from the training cohort. (C) KM investigation of the risk scores for the significant gene signatures. (D) Three- and five-year ROC curves of the training cohort. (E) Survival status distribution in the test cohort. (F) Amount of biomarker gene expression in the GEPIA.
Figure 4

Prediction model construction for clinical prognosis

(A) The forest plot was composed of risk score and clinical features. (B) The nomogram was constructed to predict the probability of patient mortality. (C) 1-3- and 5-year ROC curves of the CESC TCGA dataset. (D) The calibration plot of nomograms between predicted and observed 3-year and 5-year outcomes.
Supplementary Files

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

- S1.tif
- S2.tif
- S3.tif
- table1.xlsx
- table2.xlsx
- table3.xlsx
- Table4.xlsx