

RESEARCH

Open Access



Developing a 5-gene prognostic signature for cervical cancer by integrating mRNA and copy number variations

Wenxin Liu^{1*}, Qiuying Jiang^{2†}, Chao Sun^{3†}, ShiHao Liu⁴, Zhikun Zhao³ and Dongfang Wu³

Abstract

Background: Cervical cancer is frequently detected gynecological cancer all over the world. This study was designed to develop a prognostic signature for an effective prediction of cervical cancer prognosis.

Methods: Differentially expressed genes (DEGs) were identified based on copy number variation (CNV) data and expression profiles from different databases. A prognostic model was constructed and further optimized by stepwise Akaike information criterion (stepAIC). The model was then evaluated in three groups (training group, test group and validation group). Functional analysis and immune analysis were used to assess the difference between high-risk and low-risk groups.

Results: The study developed a 5-gene prognostic model that could accurately classify cervical cancer samples into high-risk and low-risk groups with distinctly different prognosis. Low-risk group exhibited more favorable prognosis and higher immune infiltration than high-risk group. Both univariate and multivariate Cox regression analysis showed that the risk score was an independent risk factor for cervical cancer.

Conclusions: The 5-gene prognostic signature could serve as a predictor for identifying high-risk cervical cancer patients, and provided potential direction for studying the mechanism or drug targets of cervical cancer. The integrated analysis of CNV and mRNA expanded a new perspective for exploring prognostic signatures in cervical cancer.

Keywords: Cervical cancer, Copy number variations, Differential expressed genes, Prognostic signature, Bioinformatics

Background

Human papillomavirus (HPV) vaccine uptake could prevent the incidence of cervical cancer, but according to global cancer statistics, in 2020 there were 13.3 per 100 000 women suffering from cervical cancer, and 604 127

new cases were diagnosed [1]. Surgery and chemotherapy are the main strategies for treating cervical cancer, and the International Federation of Gynecology and Obstetrics (FIGO) has also developed a staging system for personalized therapy [2]. Patients classified as having a low risk recurrence by FIGO can still develop metastasis [3], which will inevitably increase the difficulties of treatment. Therefore, discovery of efficient predictor may help predict the prognosis and guide personalized treatment of cervical cancer.

In the recent years, various biomarkers, such as immune genes [4], long non-coding RNAs [5, 6], microRNAs [7, 8] and histone genes [9], have been discovered to

*Correspondence: tj_liuwenxin@163.com

†Wenxin Liu, Qiuying Jiang and Chao Sun contributed equally to this work.

¹ Department of Gynecological Oncology, Tianjin Medical Cancer Institute and Hospital, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin, Tianjin's Clinical Research Center for Cancer, West Huan-Hu Rd, Ti Yuan Bei, Hexi District, 300060 Tianjin, China

Full list of author information is available at the end of the article



evaluate the prognosis of cervical cancer patients. Up to now, prognostic signature based on copy number variations (CNVs) has not been investigated before. A number of studies have demonstrated that CNVs are involved in tumorigenesis in many cancer types, such as lung cancer [10], leukaemia [11] and breast cancer [12]. In a pan-cancer research, Shao et al. revealed a close relation between CNVs and gene expression enriched in oncogenic pathways [13]. Advances in gene microarray technology enable us to detect duplications or deletions from focal to chromosomal associated with cancer development using various databases.

Tumor microenvironment (TME) plays a critical role in cancer cell proliferation, metastasis and immune escape. Particularly, to a large extent, the efficacy of immunotherapy is determined by TME [14]. Immunotherapy is a potentially effective strategy for cervical cancer patients with metastasis. Immune checkpoint blockade such as programmed death receptor-1 (PD-1) and CTLA-4 inhibitors has been seen as re-activators for T cell activation [15]. Currently, there are ongoing clinical trials exploring immune checkpoint inhibitors for aggressive cervical cancer.

In this study, cervical cancer samples were obtained from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases, where TCGA and GSE44001 datasets containing expression data and sequencing data were downloaded. We constructed a prognostic signature for cervical cancer patients according to combined data of CNVs and mRNAs. The effectiveness of the prognostic signature was validated using TCGA and GSE44001 datasets. The signature was robust in dividing patients into high-risk and low-risk groups, which showed distinctly different overall survival (OS). A nomogram was proposed based on the prognostic signature to satisfy a convenient clinical use. In addition, TME of high-risk and low-risk groups was described for understanding immune infiltration in the two groups. The current prognostic signature manifested a robust performance through comparison with previously reported signatures.

Methods

Data source

The workflow of constructing a prognostic model for cervical cancer was shown in Fig. 1. The data of cervical

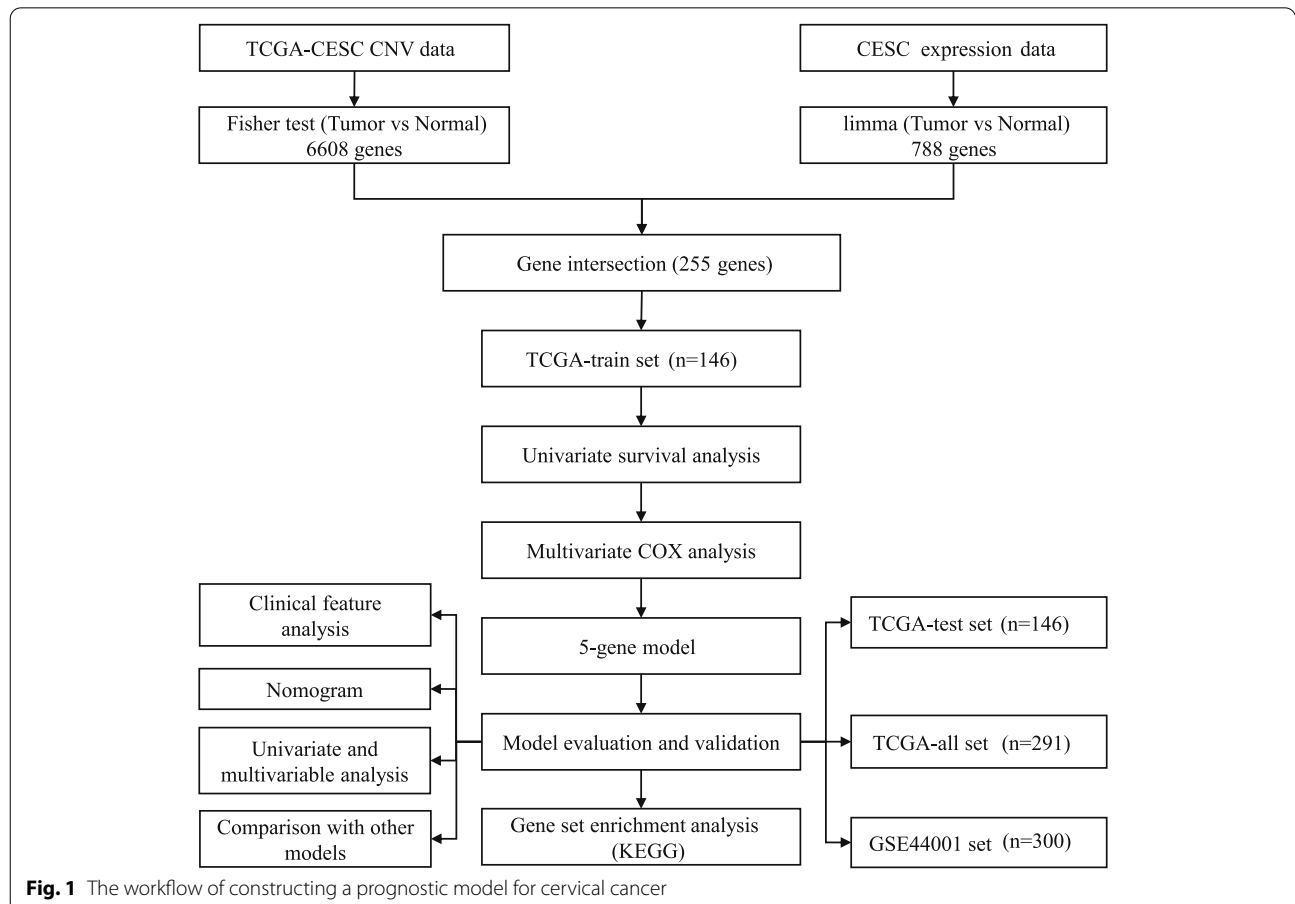


Fig. 1 The workflow of constructing a prognostic model for cervical cancer

cancer samples were downloaded from TCGA (<https://portal.gdc.cancer.gov/>) database and GEO (<https://www.ncbi.nlm.nih.gov/geo/>) database on August 30, 2021. TCGA dataset included RNA sequencing (RNA-seq) data, CNV data and clinical information. GSE44001 from GEO included expression profiles. 10 normal samples of cervix uteri containing expression profiles were downloaded from GTEx database (<https://www.gtexportal.org/>) on August 30, 2021.

Data preprocessing

For TCGA dataset, samples without clinical information, survival status (dead and alive) or survival time were excluded. Primary solid tumor and normal solid tissue were retained. “RemoveBatchEffect” function in limma R package [16] was used to remove batch effects between TCGA and GTEx datasets (Supplementary Figure S1). For GSE44001 dataset, samples without survival status were excluded. Probe ID was converted to gene symbol. One probe containing multiple genes was excluded. When one gene had multiple probes, averaged expression value of these probes was selected. Finally, 291 samples were remained in TCGA dataset, which were grouped by survival status (220 alive and 71 dead), T stage (T1-137, T2-67, T3-16, T4-10, TX-61), N stage (N0-128, N1-55, NX-108), M stage (M0-107, M1-10, MX-174), stage (I-159, III -64, III-41, IV-21, X-6), grade (G1-18, G2-129, G3-116, G4-1, GX-27), age (139 samples \leq 45 years and 152 samples $>$ 45 years) and HPV status (positive-9, negative-167, NA-115). In GSE44001 dataset, 300 samples were remained with 262 alive and 38 dead samples. The clinical information of these samples was displayed (Supplementary Table S1).

Identification of differentially expressed genes (DEGs)

DEGs were obtained from differential CNVs and differential genes. For CNV data, bedtoolsr R package [17] was applied to transform CNV segments to genes, and CNVs in tumor samples and normal samples were calculated. The differential CNVs were identified by *Chi*-square test with $P < 0.05$. For expression profiles, limma R package was used to identify DEGs between normal samples in GTEx dataset and tumor samples in TCGA dataset ($P < 0.01$, |fold change (FC)| $>$ 1.5). Finally, the intersection of genes in differential CNVs and DEGs in the expression profiles was the gene set of DEGs.

Analysis of GO function and KEGG pathways

WebGestaltR package [18] was applied to analyze Gene Ontology (GO) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. WebGestaltR is a popular tool supporting various functional categories and databases when performing enrichment analysis.

GO function includes molecular function, cellular component and biological process. Only the top 10 enriched terms or pathways were visualized.

Construction of a prognostic model

TCGA dataset was randomly divided into training group and test group at a ratio of 1:1. The suitable division was selected under the conditions of similar distribution of two groups in ages, genders, survival status, follow-up time and similar number of samples of binary classification for expression profiles. Finally, 146 samples in training group and 145 samples in test group were determined. *Chi*-square test showed no significant difference between two groups ($P > 0.05$, Supplementary Table S2). Moreover, on the univariate Cox regression analysis of genes in training and test groups, a similar distribution of their *P* values was displayed, indicating that the classification of two groups was reliable (Supplementary Figure S2). GSE44001 was set as an independent validation group.

Survival coxph function in survival R package was used to perform univariate Cox regression analysis in training group, and $P < 0.05$ was set to screen DEGs. StepAIC in MASS package [19] was applied for model optimization. The AIC of the model decreased by decreasing variables one by one, with a lower AIC reflecting a more optimized model. The prognostic model was defined as risk score = coefficient 1 * gene expression 1 + coefficient 2 * gene expression 2 + ... + coefficient n * gene expression n, where coefficients were obtained from the result of univariate Cox regression analysis. Risk score was converted to z-score to classify samples into high-risk and low-risk groups through z-score = 0. TimeROC R package [20] was used to show receiver operating characteristic (ROC) curves, and area under ROC curve (AUC) was calculated to evaluate the prediction of the prognostic model. Kaplan–Meier survival analysis was employed to analyze the survival, the differences of which in the two risk group were analyzed by log-rank test.

Estimation of STromal and Immune cells in Malignant Tumours using Expression data (ESTIMATE)

ESTIMATE is a method that can evaluate the fraction of stromal and immune cells based on gene expression signatures through single sample gene set enrichment analysis (ssGSEA) [21]. The method calculates three enrichment scores, that is, stromal score, immune score and ESTIMATE score, where ESTIMATE score is the combined score of stromal score and immune score.

Microenvironment Cell Populations-counter (MCP-counter)

MCP-counter calculates the enrichment score of 10 immune-related cells (CD3 T cells, CD8 T cells, cytotoxic

lymphocytes, B lymphocytes, NK cells, monocytic lineage, myeloid dendritic cells, neutrophils, endothelial cells and fibroblasts) across mRNA mixtures in tumor tissue [22]. This tool enables to estimate relative abundance of immune-related cells based on a series of cell markers in a complex tumor microenvironment.

Single sample gene set enrichment analysis

GSVA R package was used to conduct ssGSEA for determining enrichment score of a gene set in one sample [23]. With this method, the abundance of gene expression can be calculated and compared between different groups. TME was analyzed using ssGSEA to obtain the enrichment of 28 immune cells.

Construction of a nomogram

Visualization of a nomogram allows a direct prediction of overall survival based on a series of risk factors. We included the risk factors with hazard ratio (HR) > 1 ($P < 0.05$) from multivariate Cox regression analysis. Each risk factor was assigned with a score, and total points of risk factors corresponded to survival chance of 1-year, 3-year and 5-year period.

Decision curve analysis (DCA)

To objectively compare the different factors in survival prediction, DCA, which enables a standard comparison for evaluating performance of predictive factors in clinical decision based on net benefit, was introduced here to evaluate the cost performance of nomogram, risk score and other clinical features. The methodology is commonly used in evaluating predictive models for clinical use [24, 25].

Statistical analysis

R (version 3.4.2) software was used to conduct all statistical analysis and bioinformatics analysis. $P < 0.05$ was considered as significant. Statistical methods were presented in the corresponding figure legends. All parameters were defined as default if there was no specific descriptions.

Results

Identification of DEGs based on CNV data and expression profiles

For CNV data in TCGA dataset, we analyzed the CNV of each sample by comparing with normal samples, and 6608 differential CNVs containing 6608 genes were screened ($P < 0.05$). 788 DEGs including 268 up-regulated and 520 down-regulated genes were identified ($P < 0.01$ and $|FC| > 1.5$, Fig. 2). A Venn diagram was plotted for discovering the common genes between 6608 genes from differential CNVs and 788 DEGs (Supplementary Figure S3). Functional analysis

on 255 intersected DEGs annotated 282 terms of biological processes and 86 terms of cellular components ($P < 0.05$, Fig. 3A and B, Supplementary Table S3), but no molecular function was significantly enriched. We found that cell cycle-related terms, such as chromosome segregation and nuclear division, were significantly enriched. In addition, 12 KEGG pathways were significantly annotated, and the top 10 enriched pathways were shown ($P < 0.05$, Fig. 3C). Several pathways related to cell signaling and cell proliferation, such as cell cycle, gap junction, leukocyte transendothelial migration and cell adhesion molecules, were enriched.

Construction of a prognostic model based on 255 DEGs

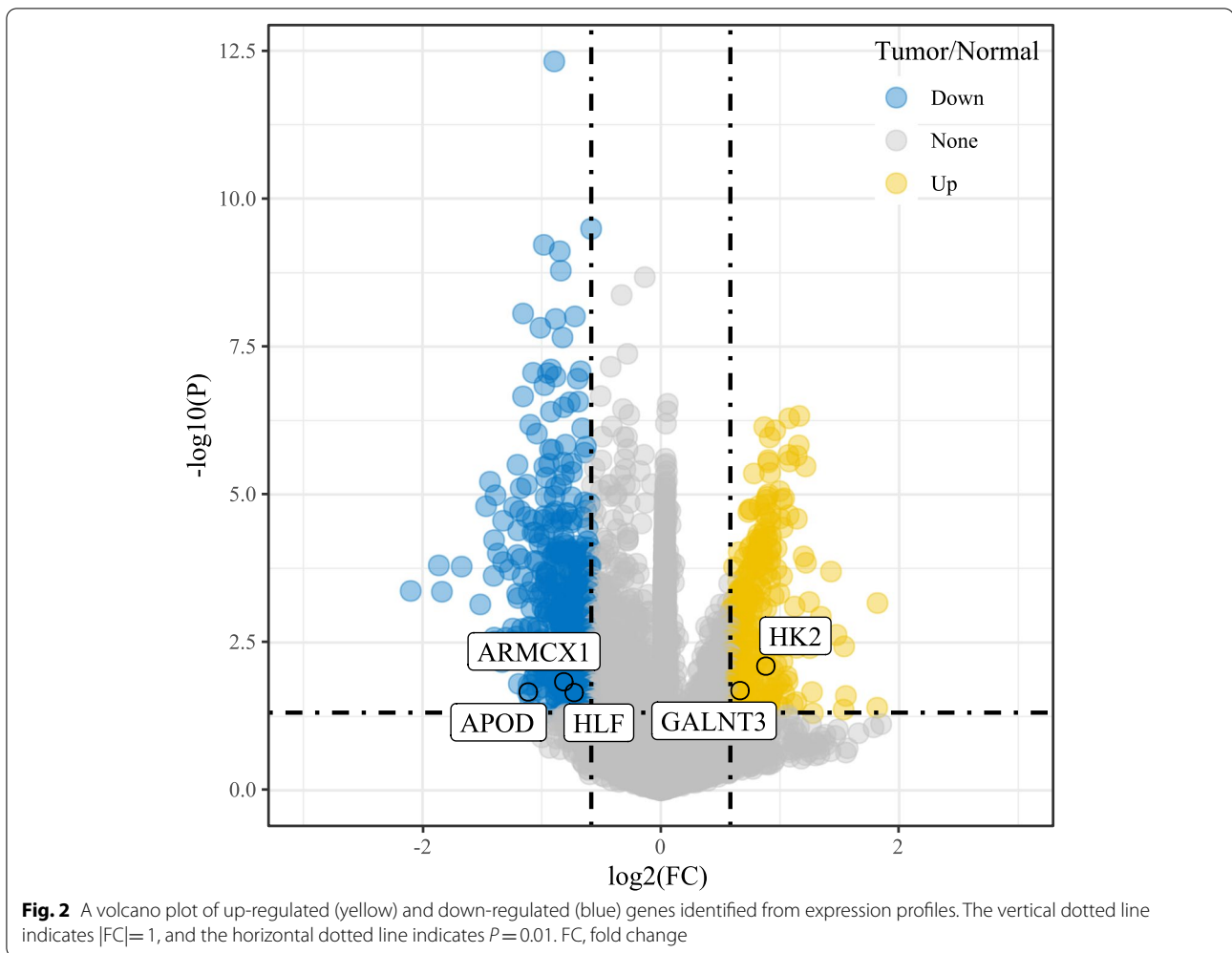
Based on 255 DEGs identified from CNV data and mRNA data, we attempted to develop a prognostic model for cervical cancer. Univariate Cox regression analysis was conducted to detect genes significantly associated with prognosis. Totally 11 genes were screened (Supplementary Table S4), and then stepAIC was used to reduce the number of genes for constructing an optimal model. Finally, 5 genes (*APOD*, *ARMCX1*, *GALNT3*, *HK2* and *HLF*) remained, and the prognostic model was defined as follow:

$$\text{Risk score} = -0.238 * \text{APOD} + 0.462 * \text{ARMCX1} + 0.503 * \text{GALNT3} + 0.406 * \text{HK2} - 0.407 * \text{HLF}$$

The distribution of these 5 genes were reflected in a genome map, with *APOD* in chromosome 3, *ARMCX1* in chromosome X, *GALNT3* and *HK2* in chromosome 2, and *HLF* in chromosome 17 (Supplementary Figure S4). The expression of the 5 genes in TCGA samples was significantly changed compared with normal samples, specifically, *APOD*, *ARMCX1* and *HLF* were down-regulated, and *GALNT3* and *HK2* were up-regulated (Supplementary Figure S5). We also found an obvious correlation between the expression of three genes (*ARMCX1*, *GALNT3* and *HK2*) and CNV (Supplementary Figure S6).

Evaluation of the 5-gene prognostic model

Firstly, we randomly divided 291 samples from TCGA dataset into two groups, with 146 samples as a training group and 145 samples as a test group (Supplementary Table S2). The risk score for each sample was calculated according to the expression of 5 prognostic genes. Risk score was converted to z-score, and z-score = 0 was set as a cut-off to classify samples into high-risk and low-risk groups in the training group (Fig. 4A). It was observed that samples of dead status were more enriched in high-risk group, and that the expression level of 5 genes were significantly distinct between two risk groups. ROC analysis revealed a high AUC of 1-year (0.81), 3-year (0.76) and 5-year (0.74) OS prediction (Fig. 4B), indicating the



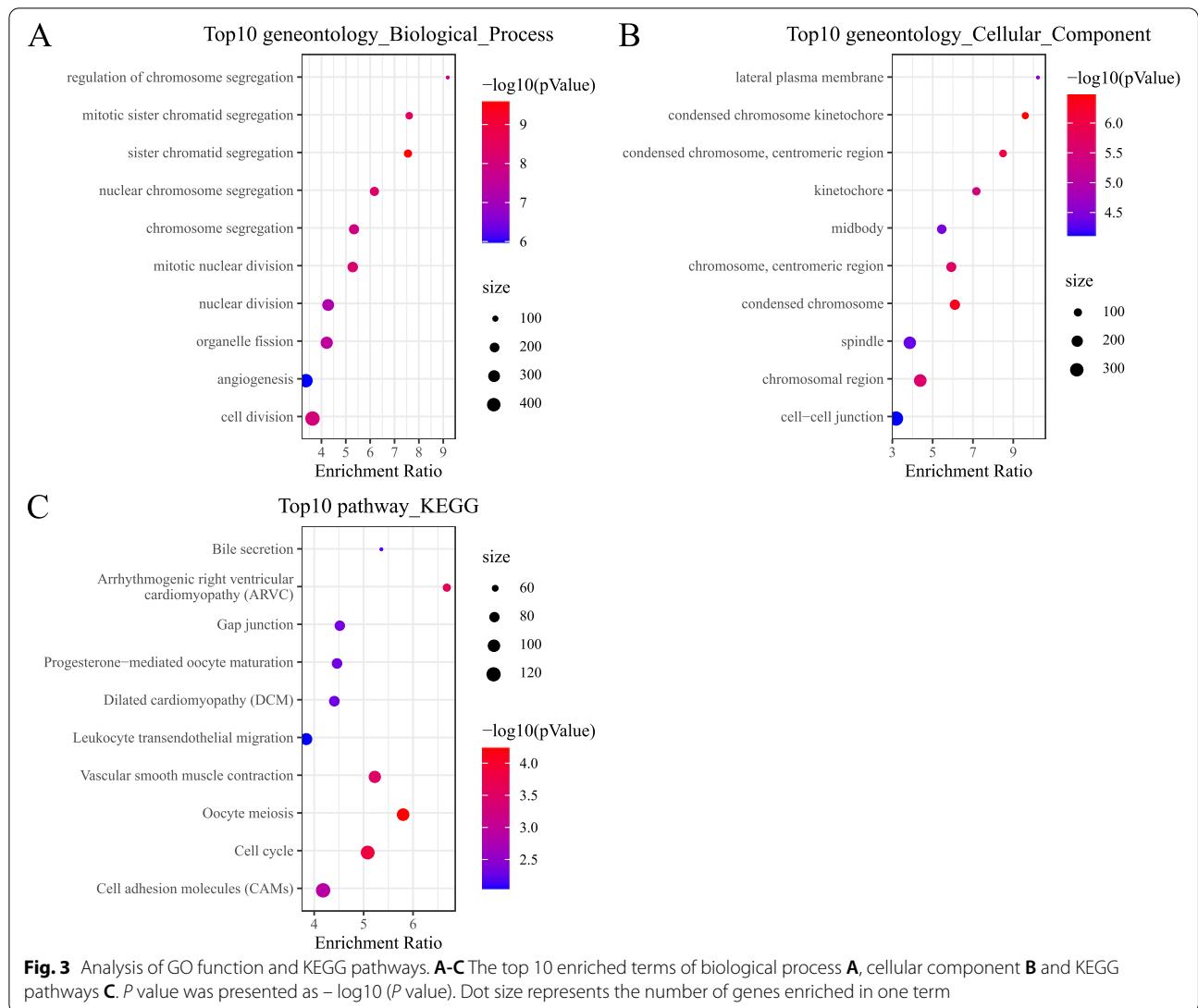
effectiveness of the prognostic model. Kaplan–Meier survival analysis also showed a significant classification of high-risk group with 66 samples and low-risk group with 80 samples ($P=0.0041$, $HR=2.98$, $95\%CI=2.01-4.41$, Fig. 4C).

The prediction of the prognostic model was further evaluated in the test group, and the model manifested the similar results when compared with the training group (Supplementary Figure S7). 145 samples were neatly classified into high-risk and low-risk groups, with a favorable AUC of 1-year (0.69), 3-year (0.68) and 5-year (0.76). Survival analysis showed that the risk score could classify patients into two groups with distinct OS ($P=0.0044$, Supplementary Figure S7C). The distribution of risk score for total samples in TCGA dataset was shown (Fig. 5). 143 samples and 148 samples were classified into high-risk and low-risk groups, respectively, with differential OS ($P=0.00013$, $HR=2.06$, $95\%CI=1.61-2.63$, Fig. 5C). Furthermore, we included an independent dataset (GSE44001) to validate the robustness of the

5-gene prognostic model, and obtained similar results that the samples were significantly divided into high-risk and low-risk groups with distinctly different prognosis ($P=0.0059$, Supplementary Figure S8). Therefore, the 5-gene prognostic model was effective in distinguishing high-risk and low-risk for cervical cancer patients.

Risk score is associated with clinical features

To assess the relation between risk score and clinical features, we evaluated the distribution of risk score in different clinical features (Supplementary Figure S9). There was a higher proportion of alive samples in low-risk group ($P<0.05$, Supplementary Figure S9A). Although no significant difference was observed among different stages, a higher proportion of samples was observed in advanced cancer stages (Supplementary Figure S9B-E). In different ages, genders and HPV status, there was no differential distribution of risk score (Supplementary Figure S9F and G).



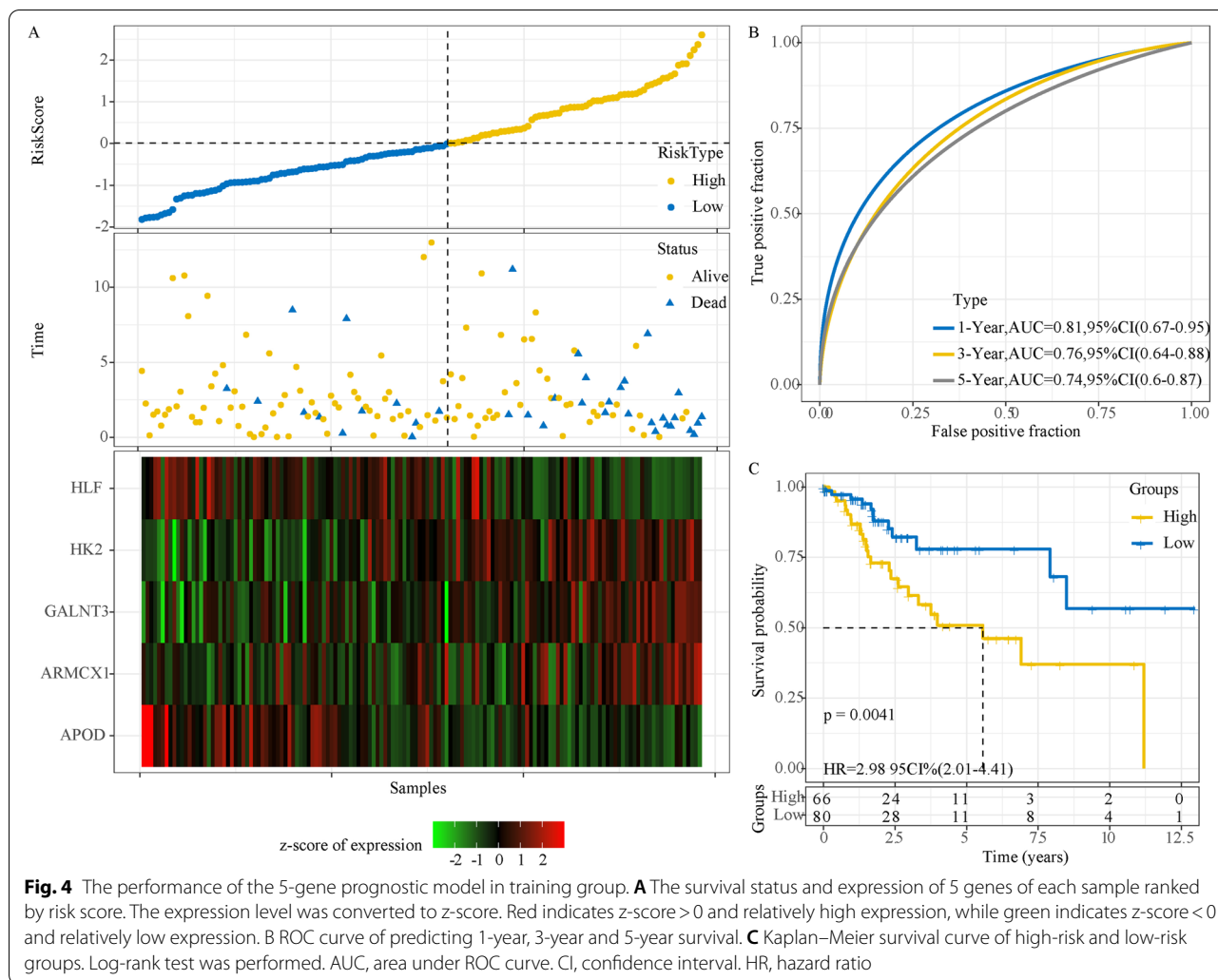
Moreover, we analyzed if risk score could classify patients into high-risk and low-risk groups with different clinical features. The results manifested that risk score was also robust in sample classification into high-risk and low-risk groups in different clinical features, including age >45 and age ≤45, T1 and T2 stages, N0 and N1 stages, M0 stage, stage I and II, stage III and IV, grade 1 and 2, grade 3 and 4, and HPV-positive ($P < 0.05$, Fig. 6).

To assess the independence of the 5-gene prognostic signature in clinical use, we applied univariate and multivariate Cox regression analysis on TCGA dataset. Univariate Cox regression analysis revealed that T stage, N stage, M stage, stage and risk type were the risk factors for cervical cancer patients ($P < 0.05$, Fig. 7A). From the data of multivariate Cox regression analysis, T stage, N stage and risk score were considered as risk factors

(Fig. 7B). High HR of risk type was presented in the univariate and multivariate Cox regression analysis, with $HR = 2.58$ ($P < 0.0001$, $95\%CI = 1.56-4.28$) and $HR = 6.21$ ($P = 0.044$, $95\%CI = 1.05-36.7$), respectively.

Constructing a nomogram based on risk score

To satisfy a convenient clinical use, we developed a nomogram that can directly exhibit the prognostic model to predict prognosis. Three risk factors (T stage, N stage and risk score) selected based on the results of multivariate Cox regression analysis and TCGA samples were included to construct a nomogram (Fig. 8A). Each risk factor can obtain a score, and total points indicates the death rate of 1-year, 3-year and 5-year survival. Predicted OS was corrected then by the observed OS (Fig. 8B).



Furthermore, DCA was used to evaluate the efficiency of the nomogram, and we found that the nomogram could more accurately assist clinical decision than other predictors in prognosis prediction (Fig. 8C).

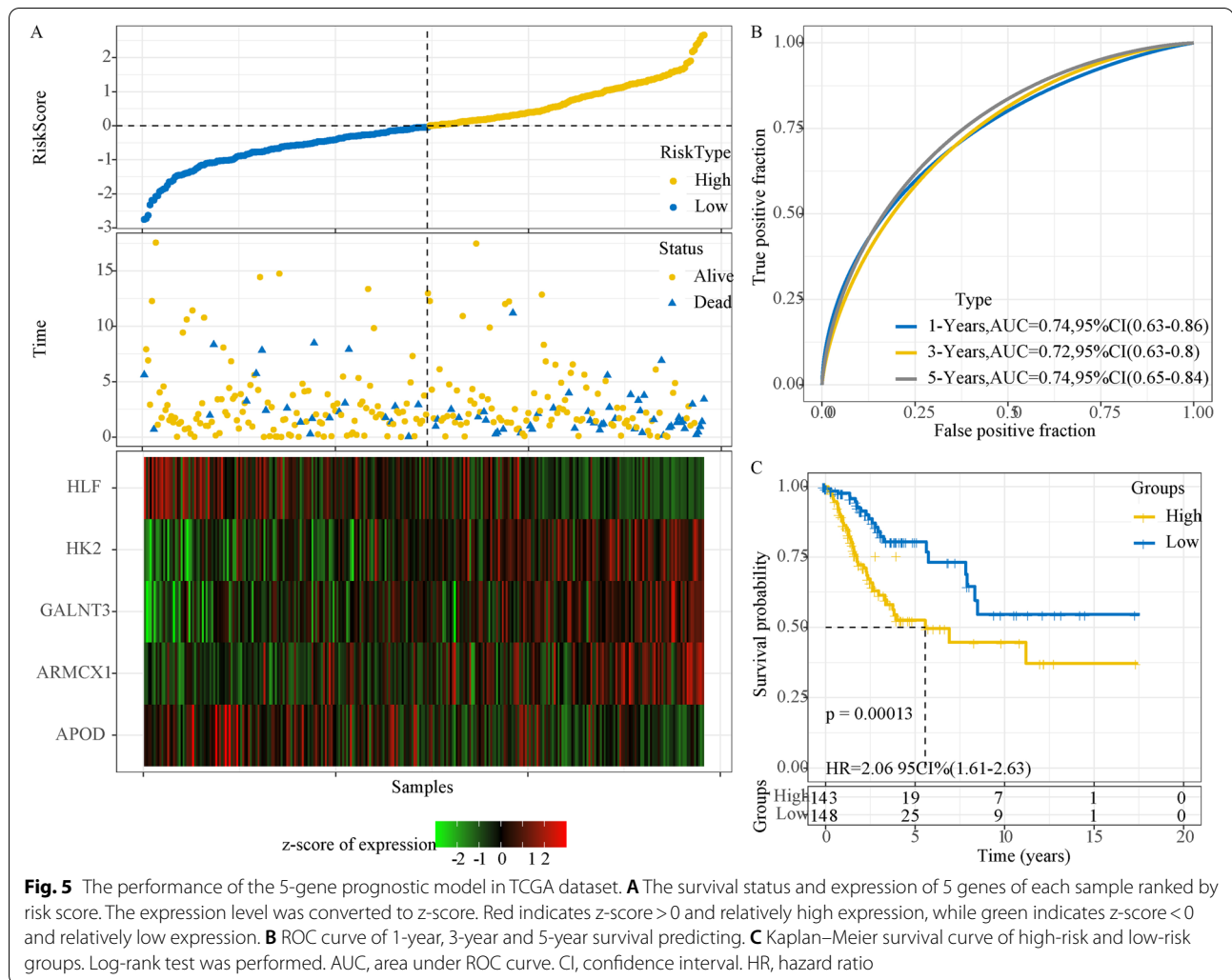
Differential immune features between high-risk and low-risk groups

Tumor microenvironment plays an important role in cancer development and immune escape, therefore we analyzed the relation between immune infiltration and risk score through ESTIMATE, MCP-counter and ssGSEA. ESTIMATE revealed that low-risk group had higher stromal score and immune score than high-risk group, indicating high immune infiltration in low-risk group ($P < 0.01$, Fig. 9A). The MCP-counter results showed that T cells, B lineage and myeloid dendritic cells contributed to a higher enrichment in low-risk group, while neutrophils and endothelial cells were higher-enriched in high-risk group ($P < 0.05$, Fig. 9B).

Additionally, we evaluated 28 types of immune cells obtained from a previous study through ssGSEA [26]. Activated B cells, immature B cells, activated CD8 T cells, effector memory CD8 T cells and myeloid-derived suppressor cells (MDSC) had significantly high enrichment score in low-risk groups ($P < 0.05$, Fig. 9C), which suggested a more activated immune response in low-risk group than high-risk group. Moreover, we assessed the expression of 47 immune checkpoints obtained from Danilova et al. [27]. 24 of 47 immune checkpoints exhibited differential expression between high-risk and low-risk groups ($P < 0.05$, Fig. 9D), meaning that the differential expression of these immune checkpoints may result in differential immune response.

Functional pathways related to risk score

To evaluate the relation between risk score and enrichment of functional pathways, we used ssGSEA to calculate the enrichment score of KEGG pathways for each



sample in TCGA dataset. Pearson correlation analysis revealed that 17 KEGG pathways were enriched and correlated with risk score ($P < 0.05$, correlation coefficient > 0.25), and most of them were related to metabolism (Fig. 10). 8 KEGG pathways, such as galactose metabolism, focal adhesion, ERBB signaling pathway and adherens junction, were positively correlated with risk score, and 9 KEGG pathways, such as oxidative phosphorylation, drug metabolism cytochrome p450 and tyrosine metabolism, were negatively correlated with risk score (Fig. 10).

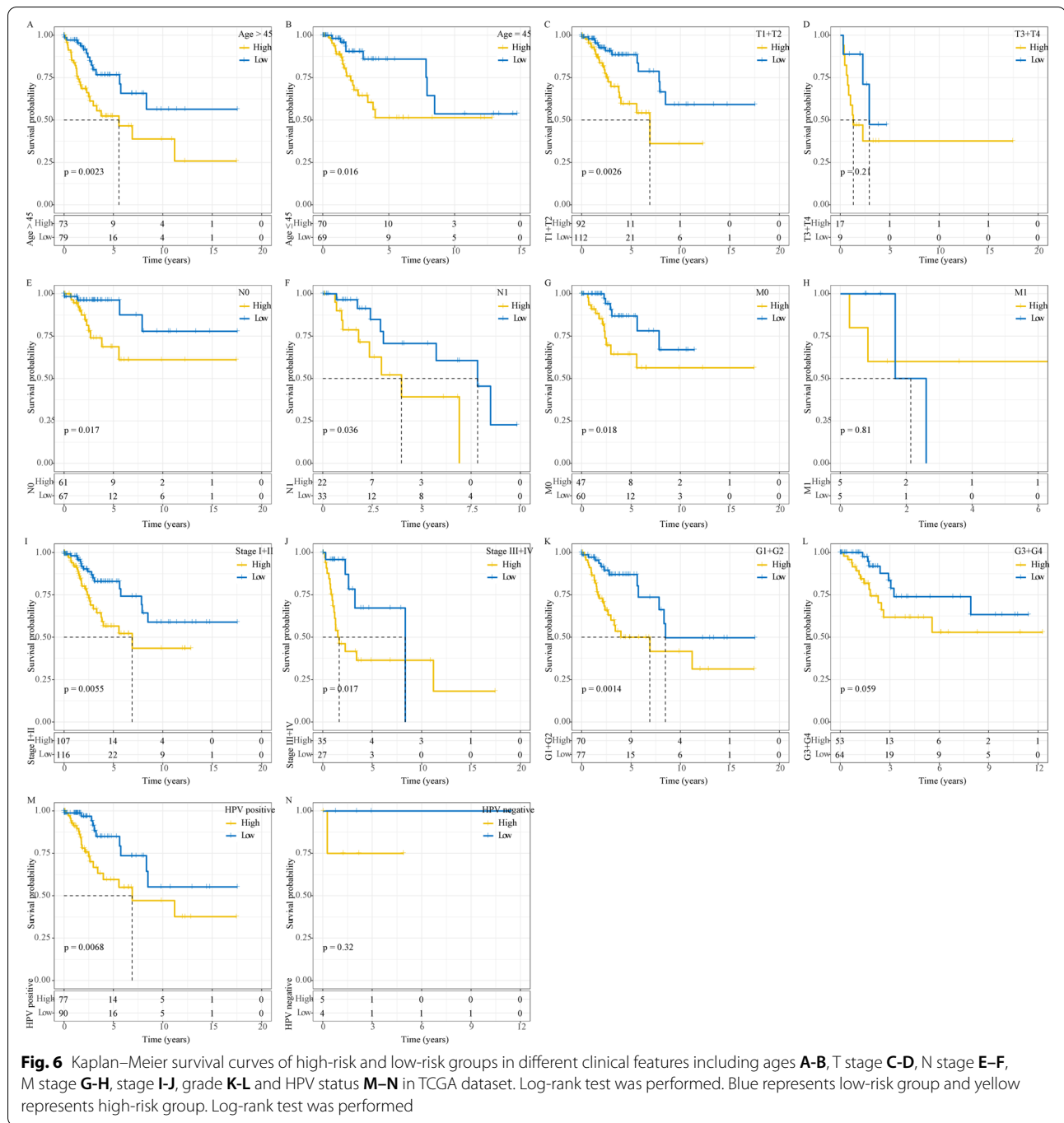
Comparison with other prognostic models

Previous studies have proposed a series of prognostic signatures for cervical cancer, we finally included four other prognostic models that used the same TCGA dataset and whose number of genes were close to our signature. To ensure a comparable standard, the method used in the current research was applied to calculate risk score

for each TCGA sample using the four prognostic models. Kaplan–Meier survival curves and ROC curves of four models were plotted (Fig. 11). These four prognostic models could clearly divide samples into high-risk and low risk groups with distinctly different prognosis ($P < 0.05$). Compared with our prognostic signature, a 5-gene signature by Ju et al. had the highest AUC (0.75, 95%CI = 0.65–0.84) of 5-year prognosis (Fig. 11A and B), and an 8-gene signature by Xie et al. had the highest AUC (0.77, 95%AUC = 0.68–0.86) of 1-year prognosis (Fig. 11G and H). However, our 5-gene signature was relatively more accurate in predicting 1-year, 3-year and 5-year prognosis, with an AUC of 0.74, 0.72 and 0.74 respectively (Fig. 5B and C).

Discussion

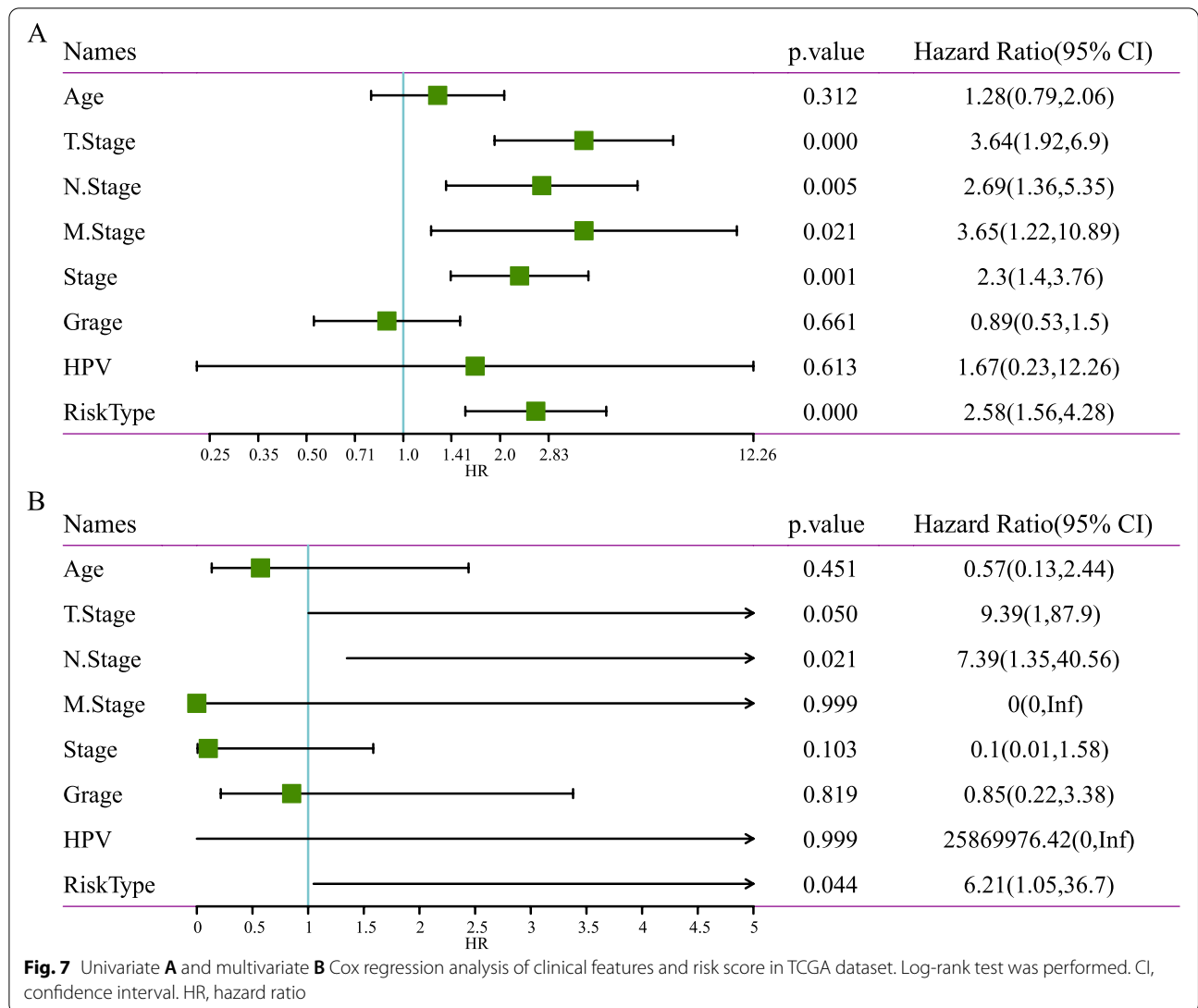
In the present study, we proposed a 5-gene prognostic signature based on integrated analysis of CNV data and expression profiles. The 5-gene signature



can effectively classify cervical cancer patients into high-risk and low-risk groups with distinctly different prognosis, and it showed a robust performance in both TCGA and GSE44001 datasets. This work was the first attempt to report a prognostic signature based on combined data of CNVs and mRNAs. Differential gene expression and CNV were both present in the 5 prognostic genes, therefore, expression level and CNV

detection may both serve as detective means for cervical cancer theoretically. CNVs are more convenient to be detected compared with mRNA expression, therefore, detecting CNVs of the 5 prognostic genes could serve as a preliminary screen of high-risk cervical cancer patients.

Risk score can be calculated according to the expression of 5 prognostic genes (*APOD*, *ARMCX1*, *GALNT3*,

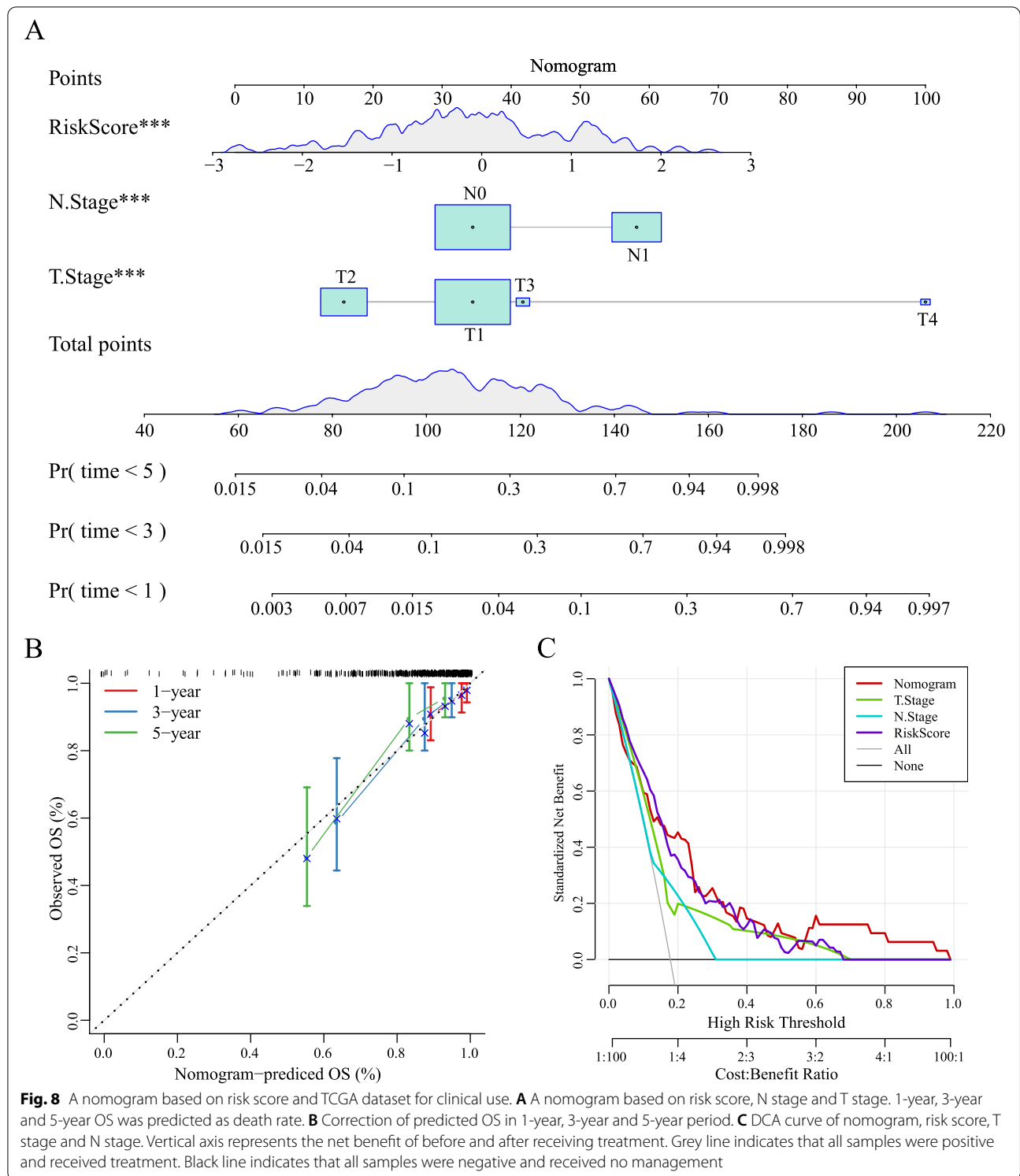


HK2 and *HLF*), which emerged as an independent risk factor in Cox regression analysis. For a convenient use in clinical practice, we constructed a nomogram based on risk score for predicting patients' survival chance. DCA evaluation showed a better performance of the nomogram than risk score. Thus, the nomogram was recommended as a predictive measurement for prognosis prediction of cervical cancer patients.

We further studied the 5 prognostic genes, and they were found to be associated with cancers reported in the previous research. APOD is one of apolipoproteins (APOs) that bind lipids and transport them to various tissues during lipid metabolism. A series of APOs have been illustrated to be associated with cancer development [28]. For example, APOA family, especially APOA1, has been considered as an independent

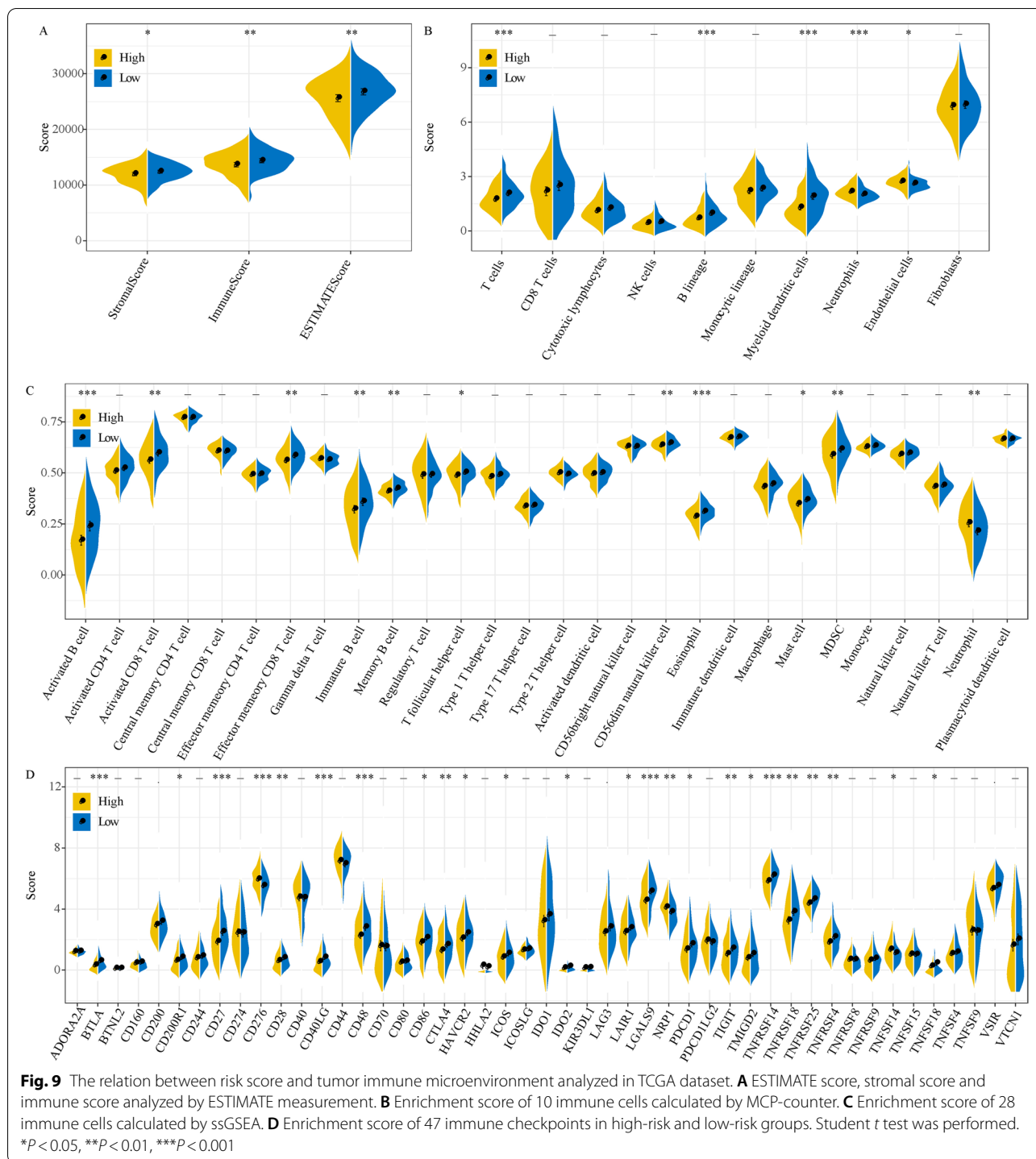
predictor for progressing cancers such as non-small cell lung adenocarcinoma, ovarian cancer, colorectal cancer and prostate cancer [28]. APOD is associated with high density lipoprotein (HDL), and is identified as a predictive biomarker hepatocellular carcinoma [29]. Low expression level of APOD is predictive of unfavorable prognosis in many cancer types, for instance, colorectal cancer [30], ovarian cancer [31] and breast cancer [32]. However, APOD is high-expressed in other cancers such as melanoma [33] and renal cell cancer [34]. In our study, APOD was lower-expressed in high-risk group as compared with low-risk group, suggesting that low-expressed APOD was associated with unfavorable prognosis of cervical cancer.

ARMCX1 (also known as ALEX1) belongs to ARM CX family clustered in X chromosome, which



could regulate protein–protein interaction. Evidence has shown that ARMCX family is involved in tumorigenesis and tumor progression through oncogenic pathways such as WNT signaling pathway [35, 36].

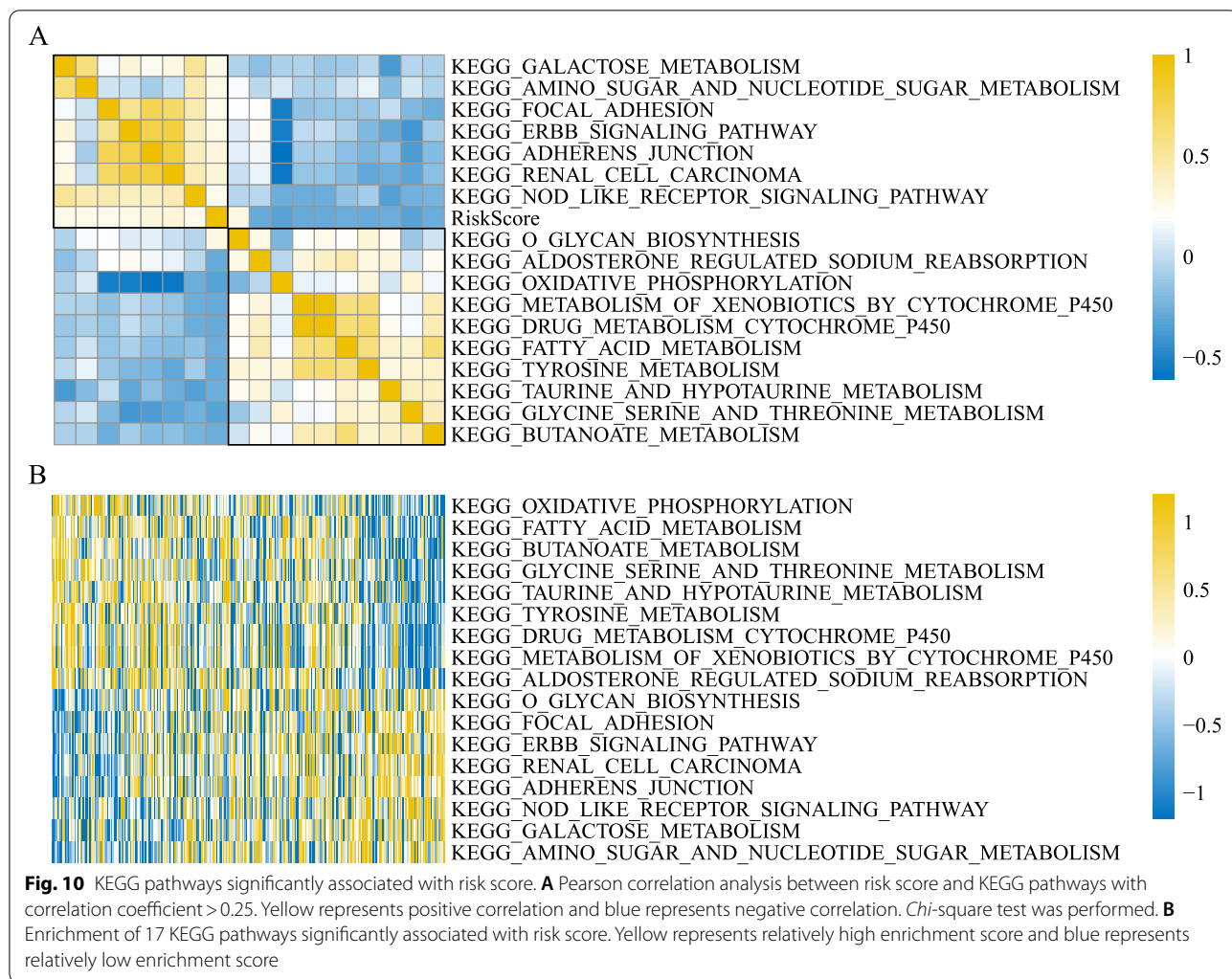
ARMCX1 is identified as a prognostic biomarker in ovarian cancer [37], colorectal cancer [38] and also in cervical cancer [39]. Zeng et al. observed a higher expression of ARMCX1 in cervical cancer tissues than



normal cervical tissues [39]. Similarly, our study found that ARMCX1 was high- expressed in high-risk group with poor prognosis.

GALNT3 is an enzyme for O-glycosylation, and down-regulation of GALNT3 has been reported to be associated with poor prognosis of lung adenocarcinoma

[40], colorectal cancer [41], and pancreatic cancer [42, 43]. Conversely, high expression of GALNT3 is detected in oral squamous cell carcinomas [44] and ovarian cancer [45]. Here in cervical cancer, we observed up-regulated GALNT3 compared with normal cervical tissues, particularly in high-risk group. Wang et al. discovered

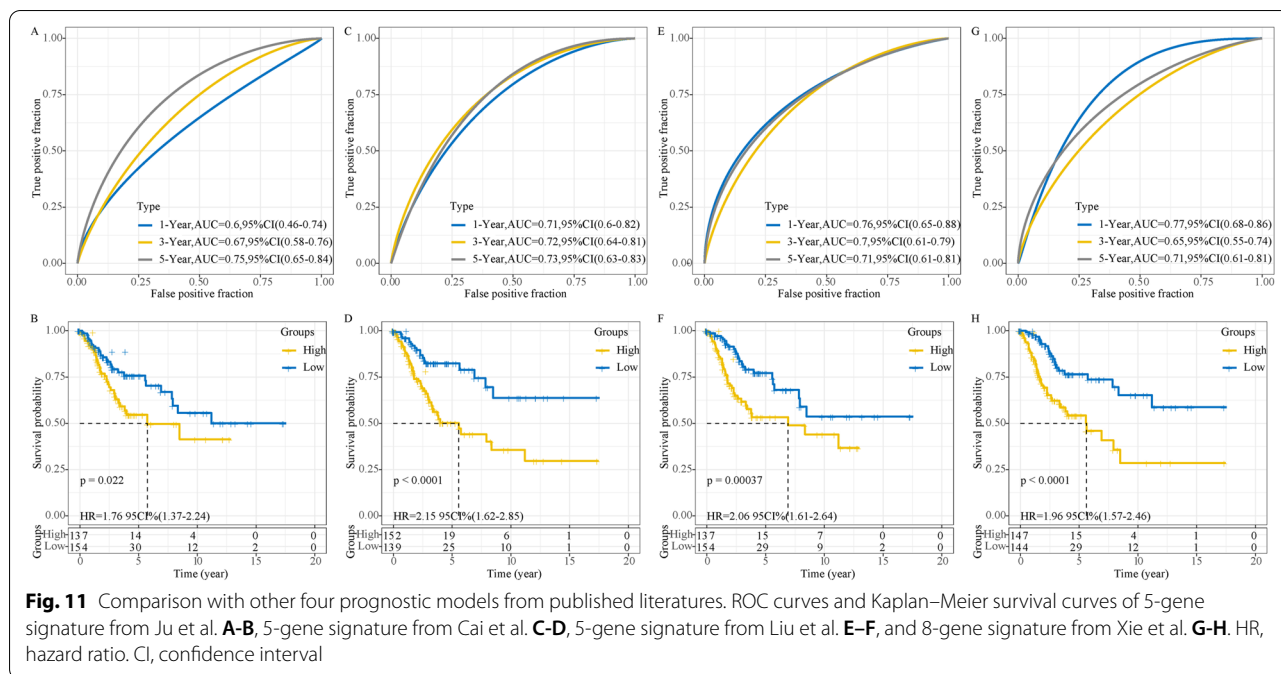


that knockdown of GALNT3 is correlated with cell adhesion molecules β -catenin and E-cadherin in ovarian cancer, which supported the invasion of epithelial ovarian cancer.

HK2 is a hexokinase with a critical role in glycolysis. Studies have found that HK2 is overexpressed in many cancer types, and that the inhibition of HK2 expression can inhibit cancer cell proliferation [46, 47]. Consistently, in this study, HK2 was also overexpressed in cervical cancer tissues and its high expression was associated with poor OS. Previously, knockdown of HK2 in cervical cancer cells has been observed to inhibit expression of AKT and mTOR, which are involved in cancer progression [48]. Such a finding supported that HK2-related pathway can serve as a potential target for treating cervical cancer. Hepatic leukemia factor (HLF) is considered as an oncogenic transcript factor, and its high expression correlates with favorable prognosis

in many cancers such as glioma [49] and non-small cell lung cancer [50]. This study also found that high expression of HLF was associated with a better prognosis of cervical cancer.

Overall, we proposed a 5-gene prognostic signature in which each component gene was closely related to various cancers. In addition, we assessed the TME of high-risk and low-risk groups, and detected a significant difference between them, indicating that these genes may participate in the modulation of immune infiltration. Compared with high-risk group, low-risk group manifested higher immune infiltration, especially activated B cells and CD8 T cells, suggesting that low-risk group had a more activated anti-tumor response and therefore may result in a favorable prognosis. In comparison with other prognostic signatures calculated by the same methodology, our 5-gene signature showed the highest AUC. However, the signature established based



on bioinformatics analysis requires further validation in clinical practice.

Conclusions

In conclusion, this study developed a novel prognostic signature with a robust performance in different datasets. The study provided the first comprehensive assessment of prognostic genes according to the relation between CNVs and cervical cancer. In addition to predictive ability for cervical cancer prognosis, the 5 prognostic genes emerged as new targets for further understanding the mechanisms of cervical cancer development.

Abbreviations

AIC: Akaike information criterion; APOs: Apolipoproteins; AUC: Area under ROC curve; CI: Confidence interval; CNV: Copy number variation; DCA: Decision curve analysis; DEGs: Differentially expressed genes; ESTIMATE: Estimation of STromal and Immune cells in Malignant Tumours using Expression data; FIGO: The International Federation of Gynecology and Obstetrics; GEO: Gene Expression Omnibus; GO: Gene ontology; HLF: Hepatic leukemia factor; HPV: Human papillomavirus; HR: Hazard ratio; KEGG: Kyoto Encyclopedia of Genes and Genomes; MCP-counter: Microenvironment Cell Populations-counter; MDSC: Myeloid-derived suppressor cells; OS: Overall survival; PD-1: Programmed death receptor-1; ROC: Receiver operating characteristic; ssGSEA: Single sample gene set enrichment analysis; TCGA: The Cancer Genome Atlas; TME: Tumor microenvironment.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12885-022-09291-z>.

Additional file 1: Supplementary Figure S1. The PCA plot before (A) and after (B) removing batch effects of TCGA and GTEx datasets. PCA, principle component analysis.

Additional file 2: Supplementary Figure S2. The density plot of *P* values on univariate Cox regression between genes in training and test groups. Log-rank test was performed.

Additional file 3: Supplementary Figure S3. The intersection between 6608 genes from differential CNVs and 788 DEGs.

Additional file 4: Supplementary Figure S4. The location of 5 prognostic genes in genome.

Additional file 5: Supplementary Figure S5. The expression of 5 prognostic genes in normal and cancer samples. EXP, expression.

Additional file 6: Supplementary Figure S6. Pearson correlation analysis between CNV and the expression of 5 prognostic genes.

Additional file 7: Supplementary Figure S7. The performance of 5-gene prognostic model in test group. (A) The survival status and expression of 5 genes of each sample ranking by risk score. (B) ROC curve of predicting 1-year, 3-year and 5-year survival. (C) Kaplan–Meier survival curve of high-risk and low-risk groups. Log-rank test was performed. AUC, area under ROC curve. CI, confidence interval. HR, hazard ratio.

Additional file 8: Supplementary Figure S8. The performance of 5-gene prognostic model in GSE44001 dataset. (A) The survival status and expression of 5 genes of each sample ranking by risk score. (B) ROC curve of predicting 1-year, 3-year and 5-year survival. (C) Kaplan–Meier survival curve of high-risk and low-risk groups. Log-rank test was performed. AUC, area under ROC curve. CI, confidence interval. HR, hazard ratio.

Additional file 9: Supplementary Figure S9. The distribution of risk score in different clinical features including survival status (A), T stage (B), N stage (C), M stage (D), stage (E), age (F), HPV status (G) and grade (H). ANOVA was performed. **P* < 0.05.

Additional file 10: Supplementary Table S1. The clinical information of TCGA, GTEx and GSE44001 datasets.

Additional file 11: Supplementary Table S2. The clinical information of training group and test group.

Additional file 12: Supplementary Table S3. The list of 255 common genes between genes in CNVs and DEGs.

Additional file 13: Supplementary Table S4. The results of univariate Cox regression analysis for screening prognostic genes significantly associated with prognosis.

Acknowledgements

Not applicable.

Authors' contributions

WXL and QYJ designed the study. CS and SHL contributed to the literature research, analyzed and interpreted the data. ZKZ wrote the initial draft of the manuscript. DFW reviewed and edited the manuscript. All authors read and approved the manuscript.

Funding

This study was supported by Natural Science Foundation of Tianjin, China (NO. 20JCZJC00100).

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the [GSE44001] repository in [<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44001>].

Declarations

Ethics approval and consent to participate

A local ethics committee ruled that no formal ethics approval was required in this particular case. The consent was waived as only the public data and retrospective research were applied in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Gynecological Oncology, Tianjin Medical Cancer Institute and Hospital, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin, Tianjin's Clinical Research Center for Cancer, West Huan-Hu Rd, Ti Yuan Bei, Hexi District, 300060 Tianjin, China. ²Department of Internal Medicine, Second Affiliated College of Harbin Medical University, 246 Xuefu Road, Nangang District, Harbin City 230100, Heilongjiang Province, China. ³YuceBio Technology Co., Ltd, 4th floor, phase I, dabaihui center, no.2002, Shenyang Road, Haishan street, Yantian District, Shenzhen City 440300, Guangdong Province, China. ⁴Department of Obstetrics and Gynecology, The Fourth Hospital of Hebei Medical University, NO.12, JianKang Road, Shijiazhuang 130100, Hebei Province, China.

Received: 3 June 2021 Accepted: 4 February 2022

Published online: 21 February 2022

References

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*. 2021;71(3):209–49.
- Benedet JL, Bender H, Jones H 3rd, Ngan HY, Pecorelli S. FIGO staging classifications and clinical practice guidelines in the management of gynecologic cancers. FIGO Committee on Gynecologic Oncology. *Int J Gynaecol Obstet*. 2000;70(2):209–62.
- Peiretti M, Zapardiel I, Zanagnolo V, Landoni F, Morrow CP, Maggioni A. Management of recurrent cervical cancer: a review of the literature. *Surg Oncol*. 2012;21(2):e59–66.
- Chen Y, Lin H, Pi YN, Chen XX, Zhou H, Tian Y, et al. Development and Validation of a Prognostic Signature Based on Immune Genes in Cervical Cancer. *Frontiers in oncology*. 2021;11:616530.
- Mao Y, Dong L, Zheng Y, Dong J, Li X. Prediction of recurrence in cervical cancer using a Nine-lncRNA Signature. *Front Genet*. 2019;10:284.
- Xia L, Wang H, Cai S, Su X, Shen J, Meng Q, et al. Integrated analysis of a competing endogenous RNA network revealing a prognostic signature for cervical cancer. *Front Oncol*. 2018;8:368.
- Liang B, Li Y, Wang T. A three miRNAs signature predicts survival in cervical cancer using bioinformatics analysis. *Sci Rep*. 2017;7(1):5624.
- How C, Pintiie M, Bruce JP, Hui AB, Clarke BA, Wong P, et al. Developing a prognostic micro-RNA signature for human cervical carcinoma. *PloS One*. 2015;10(4):e0123946.
- Li X, Tian R, Gao H, Yang Y, Williams BRG, Gantier MP, et al. Identification of a histone family gene signature for predicting the prognosis of cervical cancer patients. *Sci Rep*. 2017;7(1):16495.
- Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukheim R, et al. Characterizing the cancer genome in lung adenocarcinoma. *Nature*. 2007;450(7171):893–8.
- Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007;446(7137):758–64.
- Cybulski C, Wokolorczyk D, Huzarski T, Byrski T, Gronwald J, Górski B, et al. A deletion in CHEK2 of 5,395 bp predisposes to breast cancer in Poland. *Breast Cancer Res Treat*. 2007;102(1):119–22.
- Shao X, Lv N, Liao J, Long J, Xue R, Ai N, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med Genet*. 2019;20(1):175.
- Piersma SJ. Immunosuppressive tumor microenvironment in cervical cancer patients. *Cancer Microenviron*. 2011;4(3):361–75.
- Eskander RN, Tewari KS. Immunotherapy: an evolving paradigm in the treatment of advanced cervical cancer. *Clin Ther*. 2015;37(1):20–38.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
- Patwardhan MN, Wenger CD, Davis ES, Phanstiel DH. Bedtools: An R package for genomic data analysis and manipulation. *J Open Source Softw*. 2019;4(44):1742.
- Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47(W1):W199–w205.
- Zhang Z. Variable selection with stepwise and best subset approaches. *Ann Transl Med*. 2016;4(7):136.
- Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*. 2013;32(30):5381–97.
- Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612.
- Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol*. 2016;17(1):218.
- Hänzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7.
- Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol*. 2018;74(6):796–804.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–74.
- Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep*. 2017;18(1):248–62.
- Danilova L, Ho WJ, Zhu Q, Vithayathil T, De Jesus-Acosta A, Azad NS, et al. Programmed Cell Death Ligand-1 (PD-L1) and CD8 Expression Profiling Identify an Immunologic Subtype of Pancreatic Ductal Adenocarcinomas with Favorable Survival. *Cancer Immunol Res*. 2019;7(6):886–95.

28. Ren L, Yi J, Li W, Zheng X, Liu J, Wang J, et al. Apolipoproteins and cancer. *Cancer Med.* 2019;8(16):7032–43.
29. Utsunomiya T, Ogawa K, Yoshinaga K, Ohta M, Yamashita K, Mimori K, et al. Clinicopathologic and prognostic values of apolipoprotein D alterations in hepatocellular carcinoma. *Int J Cancer.* 2005;116(1):105–9.
30. Ogawa K, Utsunomiya T, Mimori K, Yamashita K, Okamoto M, Tanaka F, et al. Genomic screens for genes upregulated by demethylation in colorectal cancer: possible usefulness for clinical application. *Int J Oncol.* 2005;27(2):417–26.
31. Vázquez J, González L, Merino A, Vizoso F. Expression and clinical significance of apolipoprotein D in epithelial ovarian carcinomas. *Gynecol Oncol.* 2000;76(3):340–7.
32. Díez-Itza I, Vizoso F, Merino AM, Sánchez LM, Tolivia J, Fernández J, et al. Expression and prognostic significance of apolipoprotein D in breast cancer. *Am J Pathol.* 1994;144(2):310–20.
33. Miranda E, Vizoso F, Martín A, Quintela I, Corte MD, Seguí ME, et al. Apolipoprotein D expression in cutaneous malignant melanoma. *J Surg Oncol.* 2003;83(2):99–105.
34. Sandim V, Pereira Dde A, Kalume DE, Oliveira-Carvalho AL, Ornellas AA, Soares MR, et al. Proteomic analysis reveals differentially secreted proteins in the urine from patients with clear cell renal cell carcinoma. *Urol Oncol.* 2016;34(1):5.e11–25.
35. Iseki H, Takeda A, Andoh T, Takahashi N, Kurochkin IV, Yarmishyn A, et al. Human Arm protein lost in epithelial cancers, on chromosome X 1 (ALEX1) gene is transcriptionally regulated by CREB and Wnt/beta-catenin signaling. *Cancer Sci.* 2010;101(6):1361–6.
36. Wang T, Zhong H, Qin Y, Wei W, Li Z, Huang M, et al. ARMCX Family Gene Expression Analysis and Potential Prognostic Biomarkers for Prediction of Clinical Outcome in Patients with Gastric Carcinoma. *Biomed Res Int.* 2020;2020:3575038.
37. Montavon C, Gloss BS, Warton K, Barton CA, Statham AL, Scurry JP, et al. Prognostic and diagnostic significance of DNA methylation patterns in high grade serous ovarian cancer. *Gynecol Oncol.* 2012;124(3):582–8.
38. Iseki H, Takeda A, Andoh T, Kuwabara K, Takahashi N, Kurochkin IV, et al. ALEX1 suppresses colony formation ability of human colorectal carcinoma cell lines. *Cancer Sci.* 2012;103(7):1267–71.
39. Zeng F, Liao K, Wu J, Gao Y, Li H, Fan J, et al. ALEX1 may be a novel biomarker for human cervical squamous cell carcinoma. *Int J Clin Exp Pathol.* 2015;8(8):9434–9.
40. Dosaka-Akita H, Kinoshita I, Yamazaki K, Izumi H, Itoh T, Katoh H, et al. N-acetylgalactosaminyl transferase-3 is a potential new marker for non-small cell lung cancers. *Br J Cancer.* 2002;87(7):751–5.
41. Shibao K, Izumi H, Nakayama Y, Ohta R, Nagata N, Nomoto M, et al. Expression of UDP-N-acetyl-alpha-D-galactosamine-polypeptide galNAc N-acetylgalactosaminyl transferase-3 in relation to differentiation and prognosis in patients with colorectal carcinoma. *Cancer.* 2002;94(7):1939–46.
42. Li Z, Yamada S, Inenaga S, Imamura T, Wu Y, Wang KY, et al. Polypeptide N-acetylgalactosaminyltransferase 6 expression in pancreatic cancer is an independent prognostic factor indicating better overall survival. *Br J Cancer.* 2011;104(12):1882–9.
43. Chugh S, Meza J, Sheinin YM, Ponnusamy MP, Batra SK. Loss of N-acetylgalactosaminyltransferase 3 in poorly differentiated pancreatic cancer: augmented aggressiveness and aberrant ErbB family glycosylation. *Br J Cancer.* 2016;114(12):1376–86.
44. Harada Y, Izumi H, Noguchi H, Kuma A, Kawatsu Y, Kimura T, et al. Strong expression of polypeptide N-acetylgalactosaminyltransferase 3 independently predicts shortened disease-free survival in patients with early stage oral squamous cell carcinoma. *Tumour Biol.* 2016;37(1):1357–68.
45. Wang ZQ, Bachvarova M, Morin C, Plante M, Gregoire J, Renaud MC, et al. Role of the polypeptide N-acetylgalactosaminyltransferase 3 in ovarian cancer progression: possible implications in abnormal mucin O-glycosylation. *Oncotarget.* 2014;5(2):544–60.
46. Garcia SN, Guedes R, Marques MM. Unlocking the Potential of HK2 in Cancer Metabolism and Therapeutics. *Curr Med Chem.* 2019;26(41):7285–322.
47. Lis P, Dyląg M, Niedźwiecka K, Ko YH, Pedersen PL, Goffeau A, et al. The HK2 Dependent "Warburg Effect" and Mitochondrial Oxidative Phosphorylation in Cancer: Targets for Effective Therapy with 3-Bromopyruvate. *Molecules (Basel, Switzerland).* 2016;21(12):1730.
48. Liu C, Wang X, Zhang Y. The Roles of HK2 on tumorigenesis of cervical cancer. *Technol Cancer Res Treat.* 2019;18:1533033819871306.
49. Liu Q, Ge H, Liu P, Li Y. High Hepatic leukemia factor expression indicates a favorable survival in glioma patients. *Medicine.* 2021;100(6):e23980.
50. Chen J, Liu A, Lin Z, Wang B, Chai X, Chen S, et al. Downregulation of the circadian rhythm regulator HLF promotes multiple-organ distant metastases in non-small cell lung cancer through PPAR/NF-kb signaling. *Cancer Lett.* 2020;482:56–71.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

