Submitted: 23 May, 2023 Accepted: 12 July, 2023 Published: 15 October, 2023 DOI:10.22514/ejgo.2023.079

European Journal of Gynaecological Oncology

ORIGINAL RESEARCH

Development of a novel transcription factor signature for accurate cervical cancer prognosis

Siru Chen[1,2,†], Xuerou Li[3,†], Tingting He[4], Xin Chen[1], Xiaoyu Tang[1], Qin Lu[5], Minmin Yu[6,*], Changsong Lin[3,*]

[1] Nanjing University of Chinese Medicine, 210023 Nanjing, Jiangsu, China
[2] Department of Traditional Chinese Medicine, The Affiliated Huai'an Hospital of Xuzhou Medical University and The Second People's Hospital of Huai'an, 223002 Huai'an, Jiangsu, China
[3] Department of Bioinformatics, Nanjing Medical University, 211166 Nanjing, Jiangsu, China
[4] Department of Gynecology, Nanjing Hospital of Chinese Medicine Affiliated to Nanjing University of Chinese Medicine, 210022 Nanjing, Jiangsu, China
[5] Department of Ultrasound Medicine, The Affiliated Huai'an Hospital of Xuzhou Medical University and The Second People's Hospital of Huai'an, 223002 Huai'an, Jiangsu, China
[6] Department of Gynecology, Nanjing Hospital Affiliated to Nanjing University of Chinese Medicine, 210003 Nanjing, Jiangsu, China

***Correspondence**
lcs04bio@njmu.edu.cn
(Changsong Lin);
njyy022@njucm.edu.cn
(Minmin Yu)

† These authors contributed equally.

Abstract

Cervical cancer (CC) is a leading cause of cancer-related deaths in women. During tumor development, transcriptional factors regulate the transcription of proto-oncogenes and tumor suppressor genes. We examined the possibility of using transcription factors as prognostic biomarkers for patients with cervical cancer. Single-cell RNA-sequencing data were downloaded from the Gene Expression Omnibus database to identify specific activated transcription factors in different types of cells from CC. Publicly available bulk RNA-sequencing and clinical data of CC were obtained to identify associated prognostic transcription factors using survival analysis and the random survival forest methods. Accuracy and effectiveness of the established transcription factor-related predictive random survival forest model were verified using training and test datasets. We identified specific activated transcription factors in tissue cells of cervical cancer. A 3-transcription factors (*PBX4* (*PBX Homeobox 4*), *EBF2* (*EBF Transcription Factor 2*) and *ZNF696* (*Zinc Finger Protein 696*)) prognostic signature for patients with cervical cancer was constructed showing good survival prediction. Gene function enrichment analysis indicated a correlation between the prognostic characteristics and different signaling pathways associated with cancer. Using the random survival forest model based on the 3-transcription factor signature, patients with cervical cancer were stratified into low- and high-risk groups with significant variations in overall survival ($p < 0.001$). The area under the curve of the time-dependent receiver operator characteristic revealed a strong predictive accuracy for training and test datasets of the corresponding signature. CC has cellular heterogeneity of transcriptional activation. Our analyses provide a novel transcription factor-associated prognostic model for CC. These transcription factors could be used as effective prognostic biomarkers and potential therapeutic targets for patients with cervical cancer.

Keywords

Cervical cancer; Cancer prognosis; Single-cell RNA-sequencing; Transcription factors; Overall survival

1. Introduction

Cervical cancer (CC), the most common malignancy of female reproductive organs, is the second leading cause of cancer deaths in women [1]. Persistent infections with oncogenic human papillomavirus (HPV) contribute to the vast majority of high-risk CC cases, increasing disease morbidity [2, 3]. The 5-year survival rate with and without early intervention is 74% and 40% on average, respectively, thus highlighting the importance of early CC diagnosis and therapy. Treatment differs according to the clinical stage of the tumor. Presently, the most common clinical treatments for recurrent and metastatic CC are immunotherapy and targeted therapy [4]. Prognostic biomarkers can predict the course of disease and provide a basis for targeted therapy. Unfortunately, biomarkers that are currently able to predict the survival outcome of patients with CC show insufficient sensitivity and/or specificity. Therefore, it is imperative to explore and identify new prognostic indicators and potential therapeutic targets to enhance the survival of patients with CC.

Transcription factors (TFs) are capable of binding to specific DNA sequences and subsequently affect the transcription or regulation of gene expression [5]. They can be identified *via* single-cell sequencing technology, a powerful tool for studying gene expression and functional changes at the single-cell level. This technology has the capacity to reveal differences between cells, providing information on tumor heterogeneity. Considering the crucial role of TFs in cell cycle regulation, using TFs in therapeutic interventions for cancer has become of interest and may serve as novel prognostic indicators of CC.

Recently, studies have focused on the development of prediction models of cancer prognostic genes using public tumor databases, such as The Cancer Genome Atlas (TCGA). This resulted in the identification of a series of biomarkers for tumor diagnosis and prognosis [6–8]. The random survival forest (RSF) model, a machine learning model based on survival trees, is suitable for building prognostic models with survival follow-up data [9].

We aimed to identify new targets for CC treatment to provide insights into methods of overall survival (OS) prediction in CC patients. We hypothesized that single-cell RNA sequencing (scRNA-seq) analyses can identify TF activation in single cells of CC tumors. Any potential TFs were then integrated in an RSF model with combined training and test datasets, validating its accuracy and effectiveness in predicting CC prognosis.

2. Materials and methods

2.1 Dataset preparation

We collected scRNA-seq data of normal adjacent and CC tissues from the Gene Expression Omnibus database (GEO; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE168652). Bulk RNA-seq and clinical data were obtained from TCGA (https://portal.gdc.cancer.gov/). The CC cohort comprised 306 samples and three adjacent normal tissues. Corresponding patient clinical data included sex, age, survival time, survival status and the tumor node metastasis stage. Probe IDs were transformed into gene symbols in these datasets according to their ensemble gene IDs.

2.2 Single-cell RNA-sequencing data analysis

The Seurat package (v4.0) [10] in R software (v4.0.3, R Project for Statistical Computing, Vienna, Austria) was used for scRNA-seq data analysis. Data did not contain genes of mitochondrial, ribosomal or hemoglobin origin. Following standardization and normalization of the datasets, cell dimensional reduction and visual analysis were performed using the UMAP (Uniform Manifold Approximation and Projection) function. The cell cluster consisted of endothelial cells, smooth muscle cells, T-cells, monocytes and tumor cells. It was annotated separately using SingleR (v2.2.0, https://github.com/dviraran/SingleR) based on expression profiles of the genes [11]. The copy number variation in all cells was analyzed using inferCNV (v1.16.0, Trinity CTAT Project, https://github.com/broadinstitute/inferCNV) to identify tumor cells. Next, the main transcriptional regulators were identified and a regulon specificity score calculated according to the Jensen-Shannon divergence. Finally, pySCENIC (v0.11.2, https://packages.guix.gnu.org/packages/pyscenic/0.11.2/) software was used to identify the activation of TFs among the cells from CC and normal adjacent tissues.

2.3 Identification of differentially expressed transcription factors from TCGA data

TF expression data were extracted from the mRNA expression profiles. We screened differentially expressed TFs (DETFs) between CC and normal samples using the limma R package (v3.50.0) [12] with $p < 0.05$ and $|\text{log2-fold change}| > 1$ as cut-off values. To identify TFs that may be associated with CC, we ran functional enrichment analyses. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed with the DETFs using the clusterProfiler R package (v4.2.0) [13], where a $p < 0.05$ was considered statistically significant. The enriched signaling pathways were then analyzed.

2.4 Protein--protein interaction network construction

DETF genes were uploaded to the STRING database (http://www.string-db.org/) and the protein-protein interaction (PPI) network regulation relationships of these genes were analyzed and visualized using Cytoscape software (v3.7.2). Hub genes were screened using the cytoHubba plugin of the Maximal Clique Centrality method. We selected important modules in the PPI network using the MCODE (Molecular Complex Detection) plugin and constructed three subnetworks. These were used to conduct further functional enrichment analyses to investigate the biological functions and mechanisms associated with the DETFs.

2.5 Prognostic model construction and evaluation

DETF expression data from TCGA were combined with the clinical survival data. The R survival package (v3.2-13) was used to perform univariate Cox regression analysis. TFs closely related to CC prognosis were filtered using $p < 0.001$. Original data were divided into training and test datasets in a 6:4 ratio using the bootstrap method. Subsequently, variable genes were used to construct an RSF model for the training datasets. Within the in-bag, out-of-bag and test training datasets, i genes were randomly selected from the variable genes for j model tests. The c-index was used to select models that performed well in the out-of-bag and test datasets.

To assess survival differences between high- and low-risk groups, samples were divided according to median risk scores. Observed differences were evaluated by plotting Kaplan-Meier survival curves of patients with CC in the two groups using the R survival package. Accuracy and validity of the model were verified by calculating the area under the curve (AUC) for 1-, 3- and 5-year survival, and running receiver operator characteristic (ROC) analyses with the survivalROC R package (v1.0.3) [14]. An AUC value >0.60 was considered an acceptable predictive value of the risk score for the three time-dependent outcomes.

Finally, to evaluate the predictive capability of this prognostic model and confirm its reproducibility, the test dataset was used as the validation cohort. Patients were classified into high- and low-risk groups depending on the median risk score of the training dataset. Survival analysis was repeated to compare the results.

3. Results

3.1 Specific activated transcription factors in CC

In this article, we explore the role of transcription factors in cervical cancer from the perspectives of single cell sequencing and bulk RNA sequencing (Fig. 1A). First, we used Seurat software to process cervical cancer data. The SingleR package was utilized to annotate cell types, and the InferCNV software package was used to identify tumor cells. Then the cell classification results were displayed based on UMAP dimensionality reduction. The results showed that the cell types in cervical cancer tissue consisted of epithelial cells (Tumor cells), smooth muscle cells, endothelial Cells, monocyte and T cells (Fig. 1B). In order to explore the cellular heterogeneity of transcriptional activation in CC, we then applied pySCENIC to infer regulon activation underlying each cluster. Analysis revealed that TFs had obviously distinct activation in the CC. We then identified and labeled five up-regulated TFs in each kind of cell. In the malignant cell, TFs with the highest regulon specificity scores were TP63 (Tumor Protein P63), CEBPE (CCAAT Enhancer Binding Protein Epsilon), BARX2 (BarH-like homeobox 2), PRRX2 (Paired Related Homeobox 2), EN1 (Engrailed Homeobox 1) (Fig. 1C–G). Also, we can clearly see that the activated TFs have cell specificity among the different types of cells in CC (Fig. 1H). Dimensional reduction and clustering analyses according to regulon activation were visualized, with the cells showing separation from each other, indicating that regulon has tissue specificity in CC (Fig. 1I).

3.2 Differentially expressed transcription factors in TCGA

Screening the downloaded bulk RNA-seq data identified 1639 TFs from the initial 56,530 mRNA expression profiles. Of those, 337 TFs were dysregulated from the CC samples compared with those of normal samples, including 139 upregulated and 198 downregulated TFs. Volcano plots and heatmap visualizations of these DETFs are displayed in Fig. 2. GO and KEGG enrichment analyses showed that, in terms of biological processes, upregulated TFs were involved in epidermal development and downregulated TFs in pattern specification process and cell fate commitment. In terms of cellular composition, DETFs were enriched in the RNA polymerase II transcription regulator complex. Regarding the molecular function, they were enriched in DNA-binding transcriptional repressor activity. KEGG enrichment analysis showed that the DETFs are involved in pathways of and transcriptional dysregulation in cancer, the cell cycle, and regulation of pluripotency of stem cells (Fig. 3).

3.3 Protein-protein interaction network and subnetworks of DETFs

To further analyze the interaction of DETFs, we created a PPI network and identified 40 top hub genes (Fig. 4A). The first, second and third subnetwork contained 32, 11 and 14 TFs, respectively (Fig. 4B–E). The 14 most enriched biological processes in terms of these subnetworks are shown in Table 1.

3.4 Identification of suitable transcription factors by univariate cox regression analysis

To find out whether these genes were related to survival, Univariate Cox regression analysis was performed to determine the DETFs that have a significant effect on CC prognosis. Results identified 14 TFs closely associated with the OS of patients with CC (Table 2). An exposure was considered a risk factor when the hazard ratio was >1 and as a protective factor when <1.

3.5 A 3-TFs prognostic model construction

Subsequently, the 14 TFs were analyzed using the RSF algorithm to screen for the most suitable prognostic TFs of CC in the training cohort. When survival trees increased to a certain number, the error rate curve tended to be stable (Fig. 5A). This led to the identification of three TFs that best prognosed CC in the training cohort (Fig. 5B).

Kaplan-Meier survival curves showed a significant difference in OS between the high- and low-risk groups, and the TF-related prognostic model was significantly related to CC prognosis ($p < 0.0001$) (Fig. 6A). Predictive performance of the model with the training dataset in the ROC analysis was good and AUC values of the 1-, 3- and 5-year ROC curves all acceptable (Fig. 6B). With an increased risk score, survival time decreased significantly and the number of deaths in the prognostic model increased in the high-risk group. By analyzing the survival heatmap, early B-cell factor 2 (*EBF2*) and zinc finger protein 696 (*ZNF696*) were identified as risk-associated TFs, with their high expression associated with high risk. Pre-B-cell leukemia transcription homeobox 4 (*PBX4*) was a protective TF and its high expression associated with low risk (Fig. 6C).

3.6 Prognostic model evaluation

Results of these confirmative tests using the test dataset were consistent with those of the training set. Significant differences in OS were found between the high- and low-risk groups in the test dataset ($p = 0.0038$), and AUC values of the 1-, 3- and 5-year prediction results of the model were all acceptable (Fig. 7). The higher the risk score, the lower the survival time, whereby high-risk patients had a higher probability of dying than low-risk patients. *EBF2* and *ZNF696* were risk-associated TFs, and *PBX4* was a protective TF.

3.7 3-regulon activation

We then explored the gene expression and transcriptional activity of these three transcription factors from single cell transcriptome sequencing data. The results showed that *PBX4* was mainly expressed in tumor cells, smooth muscle cells, and T cells, and its transcriptional activity was also high in these three types of cells. After binarization, its transcriptional activity was high correspondingly (Fig. 8A–C). *EBF2* was mainly expressed in smooth muscle cells, and its transcriptional activity and binary transcriptional activity were also higher in smooth muscle cells (Fig. 8D–F). *ZNF696* was mainly expressed in tumor cells, smooth muscle cells, and endothelial cells, and its transcriptional activity and binary transcriptional activity are

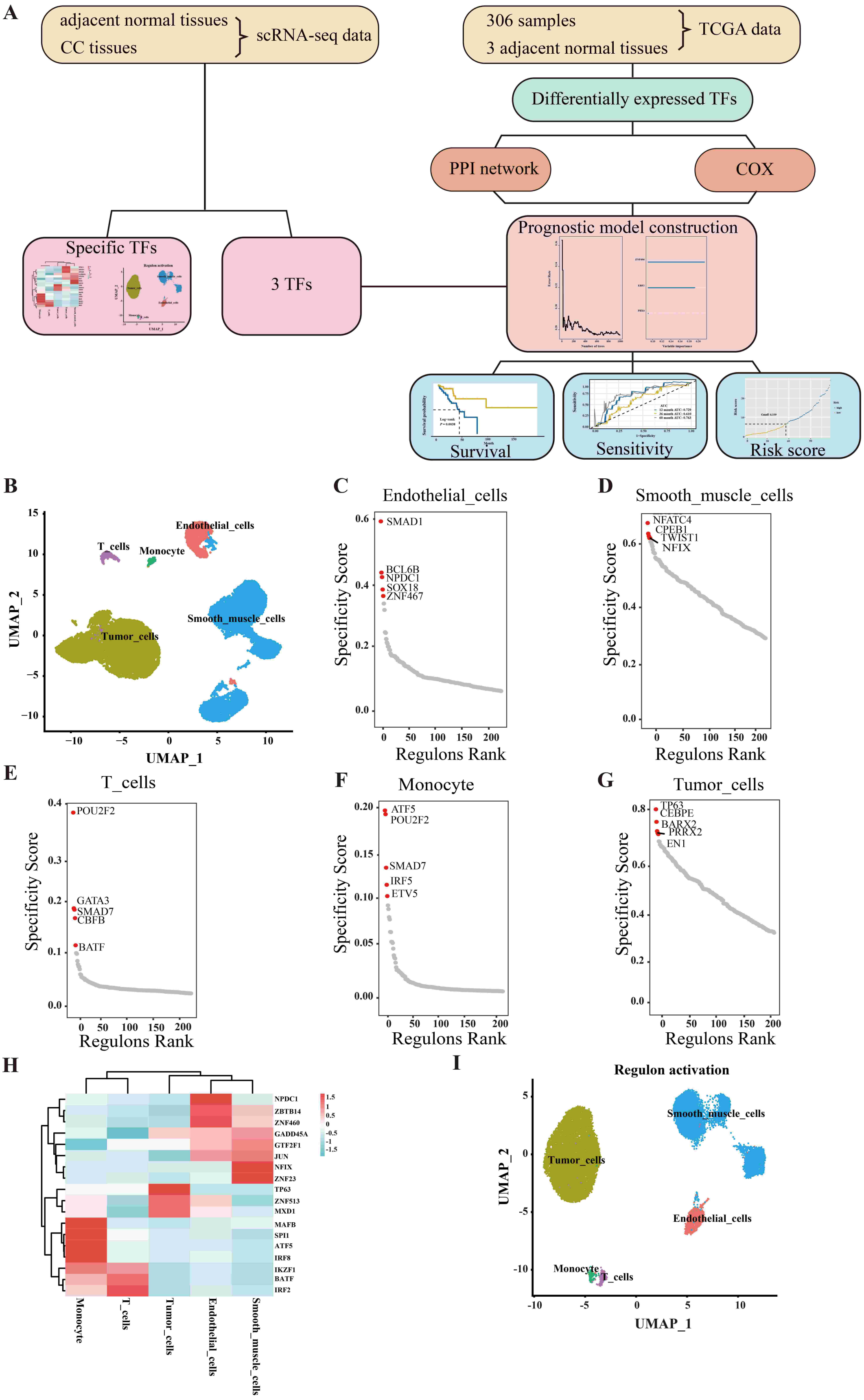


FIGURE 1. CC single-cell sequencing data analysis. (A) The Schematic diagram of this article. (B) Landscape of CC single-cell sequencing data. (C–G) Rank of regulons based on their specific score in endothelial cells, smooth muscle cells, T-cells, monocytes and tumor cells. (H) Heatmap showing selected transcription factors in each cell cluster. Scale shows expression values adjusted to a range between 0 ± 1.5 (red: highest, green: lowest expression). (I) UMAP plot of cells from both cervical cancer and adjacent normal tissue by cell type according to regulon activation.

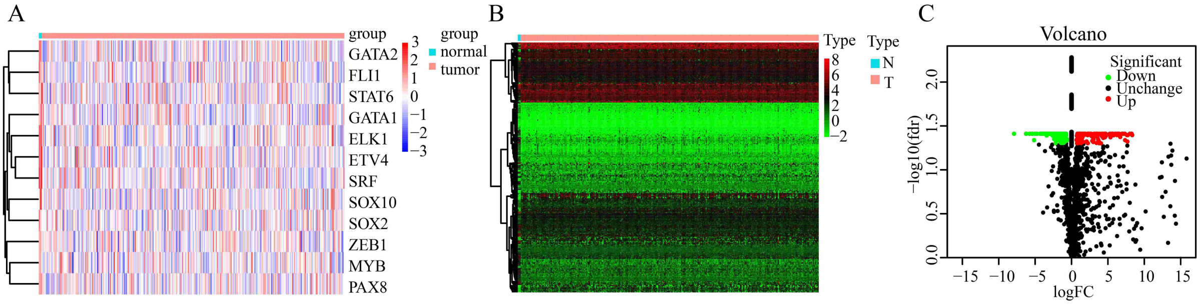


FIGURE 2. Differentially expressed transcription factors analysis from TCGA. (A) Heatmap visualization of regulons between cervical cancer (CC) and adjacent normal tissues. Red and blue colors indicate higher and lower expression, respectively. Light blue represents normal samples and pink the CC samples. (B) Heatmap visualization of 337 differentially expressed transcription factors. Red and green indicate higher and lower expression, respectively. Blue represents normal samples and pink the CC samples. (C) Volcano plot showing the expression of transcription factors (TFs). Red and green dots correspond to significantly up- and down-regulated TFs, respectively, whereas black dots show those that are not significantly expressed.

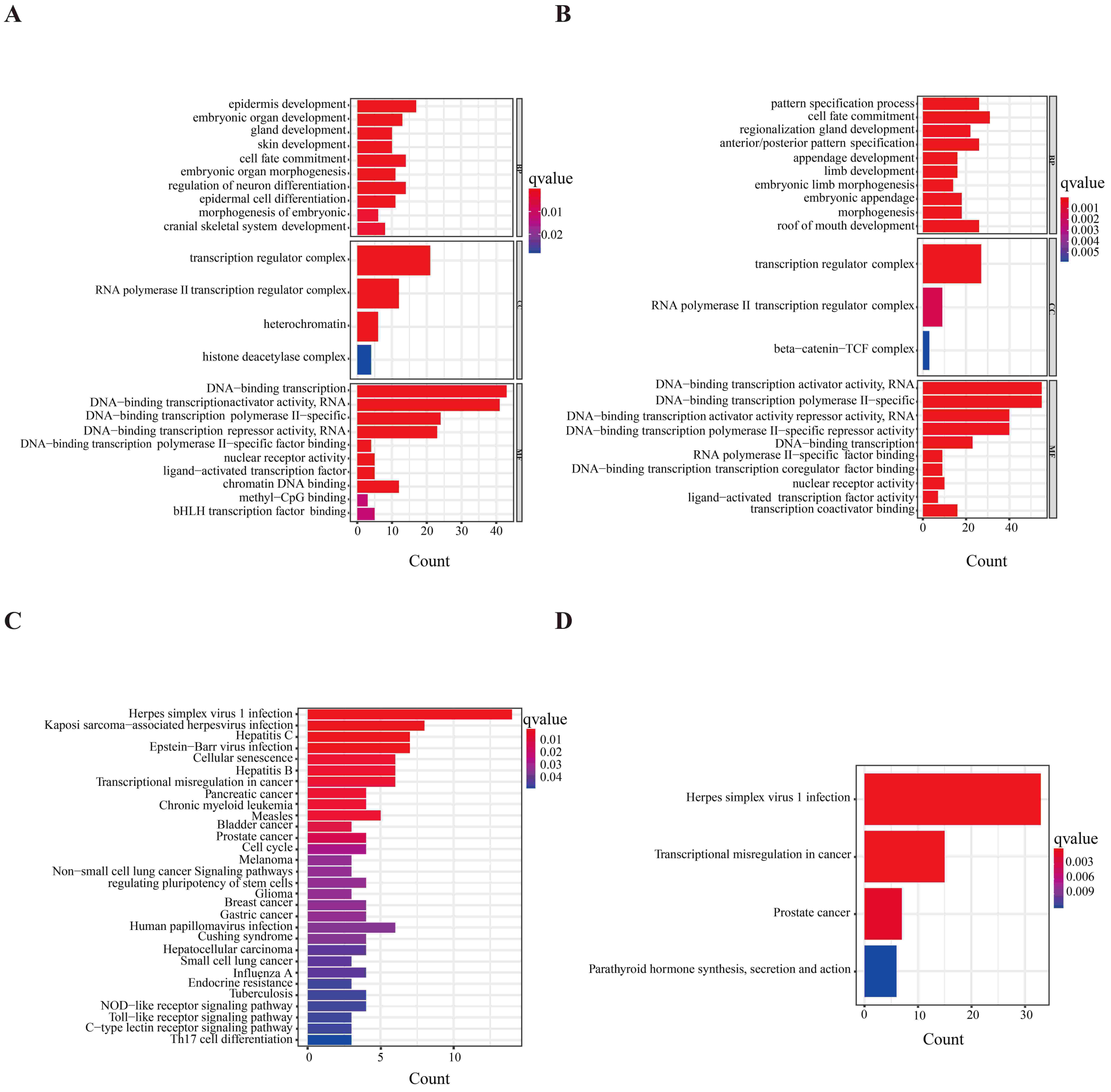
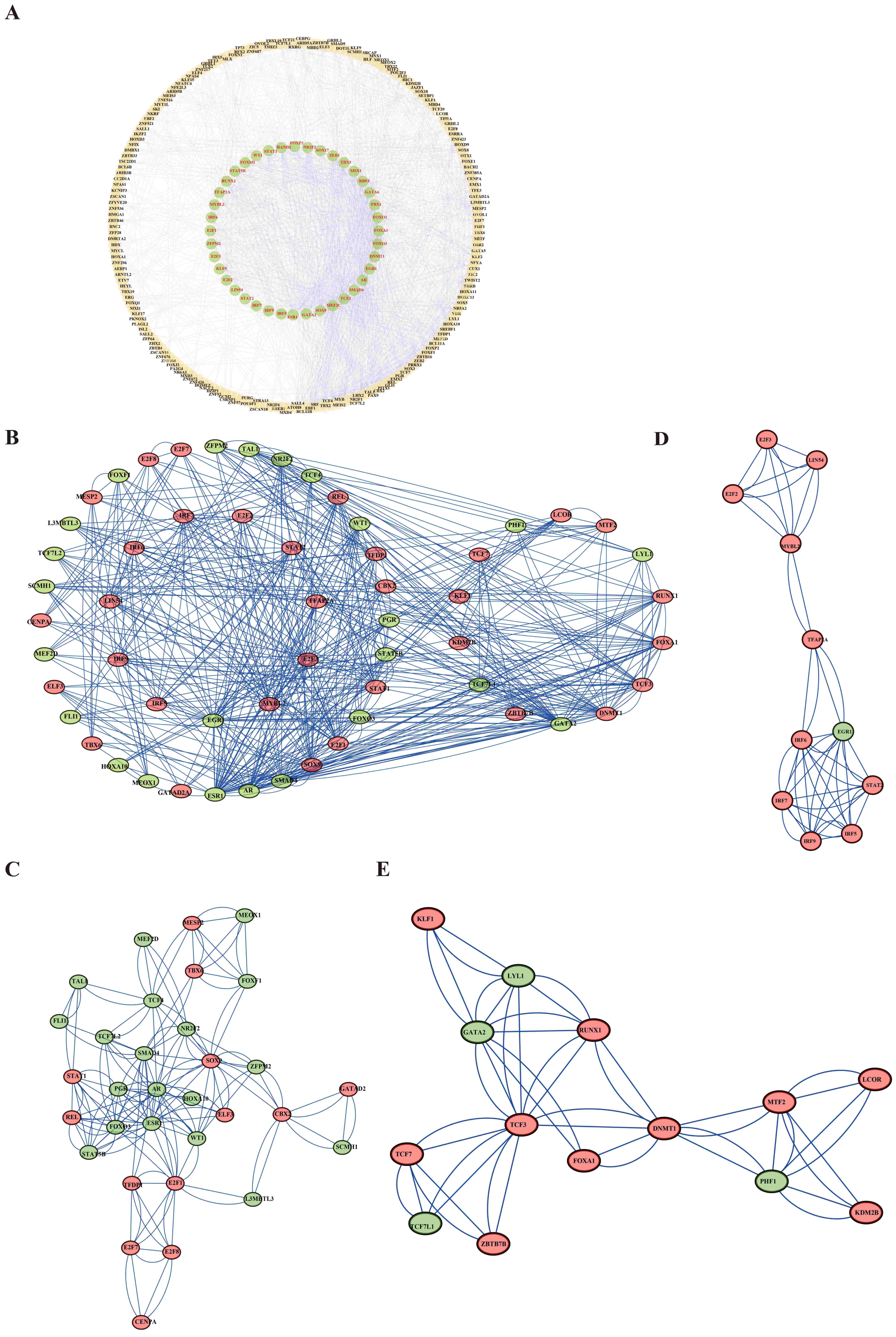


FIGURE 3. GO enrichment analysis and KEGG enrichment analysis of differentially expressed transcription factors. (A) GO enrichment analysis of up-regulated TFs. (B) GO enrichment analysis of down-regulated TFs. (C) KEGG enrichment analysis of up-regulated TFs. (D) KEGG enrichment analysis of down-regulated TFs.



F I G U R E 4. Hub genes analysis of differentially expressed transcription factors. (A) Top 40 hub genes selected using the cytoHubba plugin of the MCC (Maximal Clique Centrality) method. (B–E) Four subnetworks constructed using the MCODE plugin. Orange and green dots correspond to significantly up- and downregulated transcription factors, respectively.

TABLE 1. Most highly enriched terms for biological processes of the three created subnetworks.

Subnetwork	GO term ID	Description	Gene ratio	*p*-value
1	0048608	Reproductive structure development	13:32	8.07×10^{-14}
1	0045137	Development of primary sexual characteristics	12:32	1.37×10^{-15}
1	0007548	Sex differentiation	12:32	1.44×10^{-14}
1	0007389	Pattern specification process	10:32	1.73×10^{-9}
1	0045165	Cell fate commitment	8:32	9.22×10^{-9}
2	0051607	Defense response to virus	4:11	1.20×10^{-5}
2	0140546	Defense response to symbiont	4:11	1.20×10^{-5}
2	0019221	Cytokine-mediated signaling pathway	4:11	1.10×10^{-5}
2	1901216	Positive regulation of neuron death	3:11	2.16×10^{-5}
3	0031056	Regulation of histone modification	5:14	6.23×10^{-8}
3	0030098	Lymphocyte differentiation	5:14	5.34×10^{-6}
3	1903131	Mononuclear cell differentiation	5:14	1.01×10^{-5}
3	0006325	Chromatin organization	5:14	8.26×10^{-6}
3	0031058	Positive regulation of histone modification	4:14	5.26×10^{-7}

GO: Gene Ontology.

TABLE 2. Screening of prognostic differentially expressed transcription factors, using univariate Cox regression analysis.

Gene ID	Gene name	Hazard ratio	95% confidence interval	*p*-value
ENSG00000221818	*EBF2*	1.98	1.40–2.80	0.00010
ENSG00000157554	*ERG*	1.60	1.25–2.05	0.00020
ENSG00000183340	*JRKL*	1.85	1.30–2.64	0.00063
ENSG00000198517	*MAFK*	1.53	1.21–1.92	0.00033
ENSG00000105717	*PBX4*	0.48	0.32–0.72	0.00044
ENSG00000146587	*RBAK*	2.53	1.60–4.01	0.00008
ENSG00000232040	*ZBED9*	1.51	1.19–1.90	0.00053
ENSG00000232040	*ZMAT4*	1.96	1.40–2.73	0.00007
ENSG00000181896	*ZNF101*	0.35	0.19–0.63	0.00049
ENSG00000164631	*ZNF12*	2.79	1.60–4.87	0.00029
ENSG00000162702	*ZNF281*	1.91	1.31–2.81	0.00087
ENSG00000205903	*ZNF316*	2.40	1.51–3.81	0.00021
ENSG00000185730	*ZNF696*	2.62	1.67–4.09	0.00002
ENSG00000124203	*ZNF831*	0.50	0.34–0.74	0.00055

EBF2: EBF Transcription Factor 2; ERG: ETS transcription factor; JRKL: JRK like; MAFK: MAF BZIP Transcription Factor K; PBX4: PBX Homeobox 4; RBAK: RB Associated KRAB Zinc Finger; ZBED9: zinc finger BED-type containing 9; ZMAT4: Zinc Finger Matrin-Type 4; ZNF101: Zinc Finger Protein 101; ZNF12: Zinc Finger Protein12; ZNF281: Zinc Finger Protein 281; ZNF316: Zinc Finger Protein 316; ZNF696: Zinc Finger Protein 696; ZNF831: Zinc Finger Protein 831.

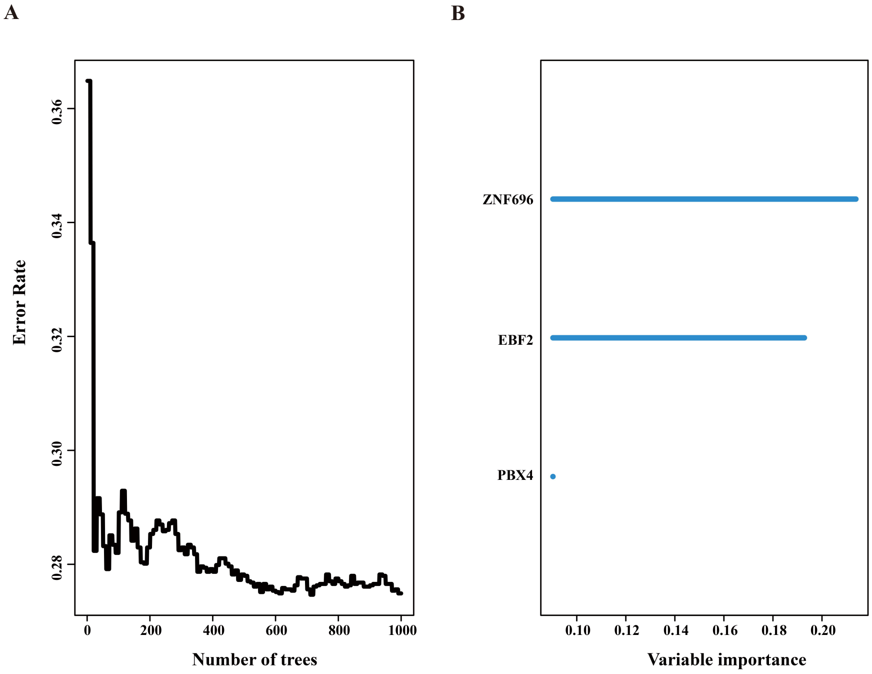


FIGURE 5. Random forest model was used to screen survival related transcription factors. (A) Relevance between the number of classification trees and error rate. (B) Importance ranking of the three most important transcription factors.

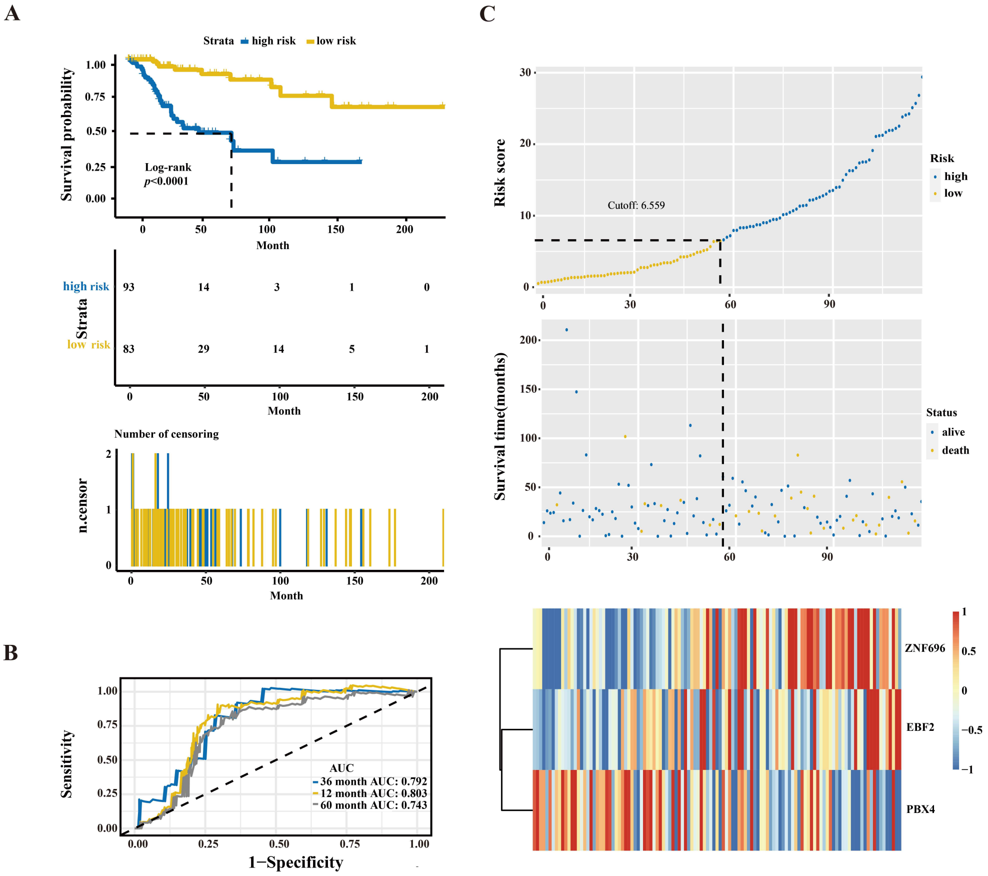
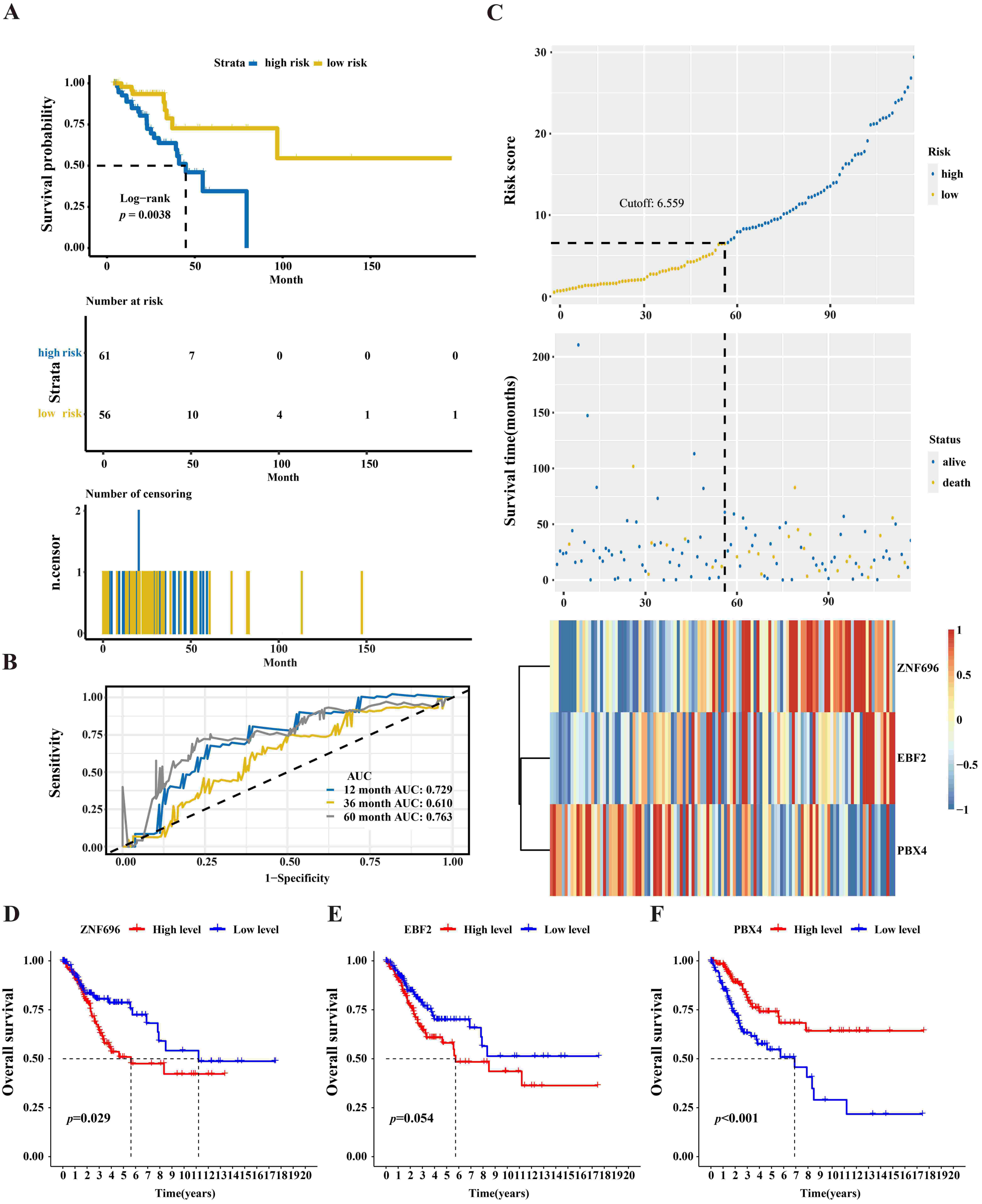


FIGURE 6. Model evaluation based on out-of-bag data using high- and low-risk groups from the training dataset. (A) Kaplan-Meier survival curve. (B) Time-dependent receiver operating characteristic curves with area under the curve (AUC) values. (C) Correlation diagram of risk factors: risk score curve (top), survival status (middle), and high (red), medium (yellow) and low (blue) expression of the transcription factors that best prognosed cervical cancer (bottom).



F I G U R E 7. Model confirmation using high- and low-risk groups from the test dataset. (A) Kaplan-Meier survival curve. (B) Time-dependent receiver operating characteristic curves with area under the curve (AUC) values. (C) Correlation diagram of risk factors: risk score curve (top), survival status (middle), and high (red), medium (yellow) and low (blue) expression of the transcription factors that best prognosed cervical cancer (bottom). (D) Overall survival of *ZNF696* (*Zinc Finger Protein 696*). (E) Overall survival of *EBF2* (*EBF Transcription Factor 2*). (F) Overall survival of *PBX4* (*PBX Homeobox 4*).

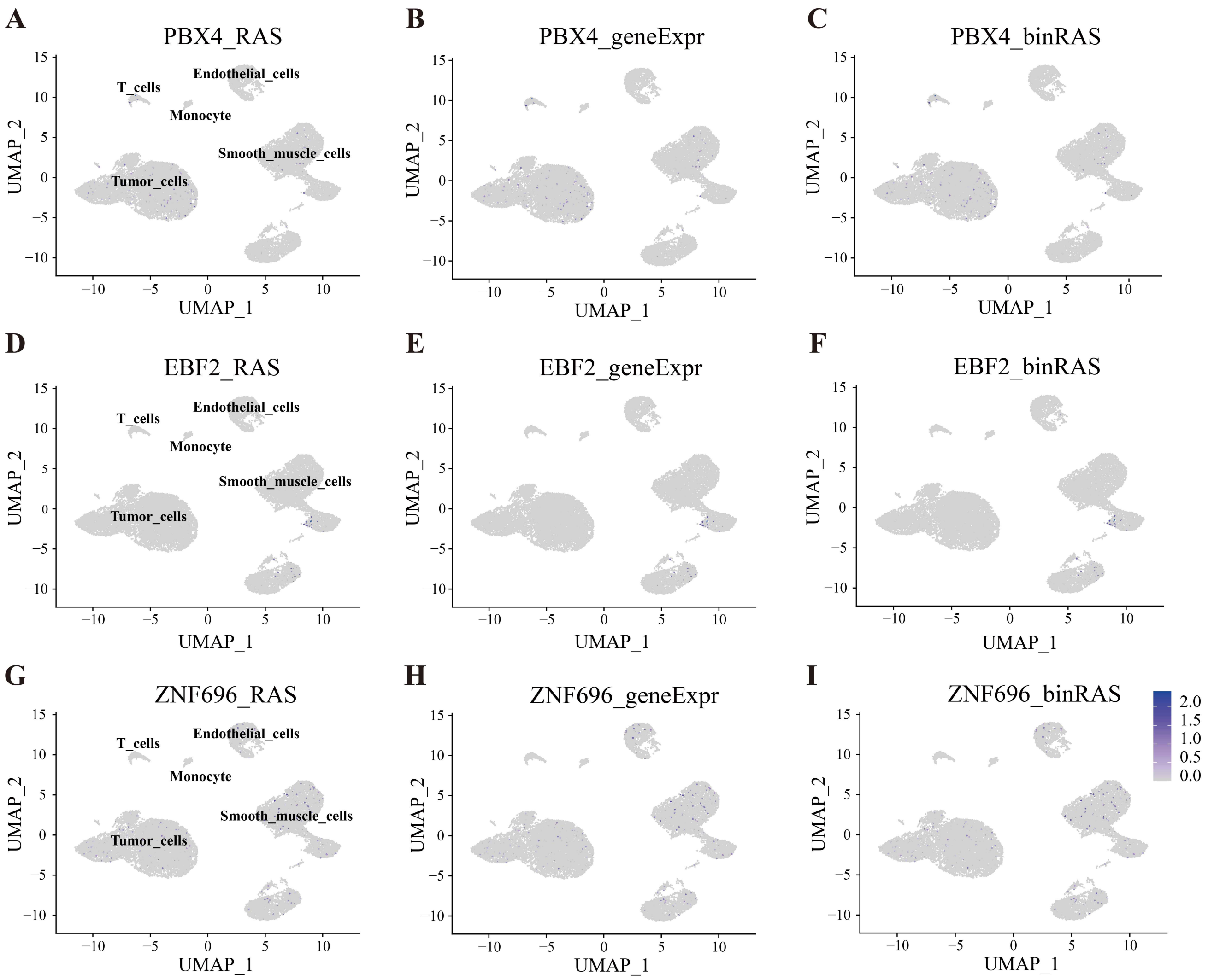


FIGURE 8. Gene expression, regulon activity score and binary regulon activity score of 3-TFs. (A–C) for PBX4 (PBX Homeobox 4), (D–F) for EBF2 (EBF Transcription Factor 2), and (G–I) for EBF2696 (Zinc Finger Protein 696).

also higher in these three types of cells (Fig. 8G–I).

4. Discussion

In developing countries, the 5-year OS of CC remains low and prognosis of patients with metastasis and recurrence is poor [15]. Owing to the biological heterogeneity of CC, no accurate method has been established for estimating its prognosis in a clinical setting. Therefore, it is particularly important to identify effective biomarkers to be used as prognostic factors for the precise diagnosis of CC.

The utilization of bioinformatic approaches has significantly advanced our understanding of CC prognosis. In recent years, various computational methodologies have been employed to uncover molecular signatures that can accurately predict disease outcomes. Several bioinformatic methods have emerged as valuable tools for determining CC prognosis. For instance, an m6A RNA methylation regulator-based prognostic signature was developed and validated in cervical squamous cell carcinoma using LASSO Cox regression analysis, which successfully determined disease prognosis [16]. In addition, a 4-miRNA model was constructed as a prognostic biomarker for CC by analyzing the miRNA expression profiles of patients with CC using Kaplan-Meier and Landmark analyses [17]. Unlike these methods, we performed an integrated analysis of TFs in CC and identified their activation in single cells using scRNA-seq data, whereby DETFs were analyzed upon mining published, high-throughput data. Combining the random forest method with traditional survival analysis tools resulted in a predictive model. The model's prediction accuracy was at least equal to or better than traditional survival analysis methods. The obtained 3-TF prognostic signature may provide a solid foundation for the clinical treatment and prognosis of CC.

The three most important TFs that best prognosed CC in the training cohort were *PBX4*, *EBF2* and *ZNF696*. *PBX4* has not been previously identified as a candidate gene in CC disease progression. PBX is a TF family member that usually interacts with the homeobox Hox gene family. Transcriptional disorders of the PBX gene family are closely related to the development and progression of cancers and may be the core regulator of signaling pathways involved in cancer progression [18–20]. Comprehensive analyses of the role of *PBX4* showed that its expression is correlated with survival prognosis and immune infiltration in various human cancers [21]. In colorectal cancer, *PBX4* is also significantly upregulated; however, its expression is thought to potentially promote tumor progression. Its overexpression significantly increased the proliferation of colorectal cancer cells through the upregulation of epithelial-mesenchymal transition and vascular markers *in vitro* [22]. Contrastingly, we found that *PBX4* was a protective TF, where its high expression was associated with low risk in patients with CC.

The second TF, *EBF2*, belongs to the Collier/Olf1/EBF family. Many diseases are related to its regulation and it plays an important role in various aspects of neural development and the immune system [23–25]. *EBF2* is differentially expressed in prostate cancer tissues, making it a candidate biomarker for prostate cancer [26]. Its expression is downregulated in bladder cancer tissues and closely related to poor prognosis [27]. Overexpression of *EBF2* inhibits apoptosis and promotes the migration and invasion of osteosarcoma cells [28]. These findings corroborate our results that indicate *EBF2* is a risk-associated TF, whereby its overexpression potentially decreases survival time.

Lastly, ZNF proteins constitute the largest transcriptional regulation family, and play multifaceted roles in various biological processes. These versatile proteins are involved in diverse functions such as DNA repair, transcriptional activation, RNA packaging, protein folding and assembly, and regulation of apoptosis [29]. Among the ZNF family members, *ZNF696* has garnered particular attention in cancer research. In the context of non-small cell lung cancer (NSCLC), *ZNF696* expression has been observed to be upregulated and associated with poor overall survival [30]. Similarly, in our investigation, *ZNF696* emerged as a risk-associated transcription factor (TF) in CC. The elevated expression of *ZNF696* was found to be correlated with poor prognosis in CC patients, further supporting its potential role as an adverse prognostic indicator in this malignancy. Overall, TFs identified in the present study may represent potential biomarkers for targeted therapy in CC, but further research is needed.

A limitation of the present study is that it is based on data from public databases. We only used a single TCGA data, without using external data for validation. Further experiments are therefore required to investigate specific roles of the three TFs in CC and to test the model *in vitro*.

5. Conclusions

Reanalysis of single cell RNA sequencing data revealed that CC has cellular heterogeneity of transcriptional activation. We identified 14 TFs of close association with the OS of patients with CC from TCGA data. After the construction of a new TF-associated prognostic model, our findings revealed that 3 TFs, including *PBX4*, *EBF2* and *ZNF696*, could effectively predict the prognostic outcomes of patients with CC. We consider that this prognostic model will be helpful in guiding effective clinical CC treatment and prognosis, and remedy the absence of prognostic biomarkers of patients with CC.

ABBREVIATIONS

CC: cervical cancer; DETF: differentially expressed transcription factor; OS: overall survival; PPI: protein-protein interaction; ROC: receiver operator characteristics; RSF: random survival forest; scRNA-seq: single-cell RNA-sequencing; TF: transcription factor; PBX4: PBX Homeobox 4; EBF2: EBF Transcription Factor 2; ZNF696: Zinc Finger Protein 696; HPV: human papillomavirus; TCGA: The Cancer Genome Atlas; UMAP: Uniform Manifold Approximation and Projection; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; TP63: Tumor Protein P63; CEBPE: CCAAT Enhancer Binding Protein Epsilon; BARX2: BarH-like homeobox 2; PRRX2: Paired Related Homeobox 2; EN1: Engrailed Homeobox 1; EBF2: EBF Transcription Factor 2; ERG: ETS transcription factor; JRKL: JRK like; MAFK: MAF BZIP Transcription Factor K; PBX4: PBX Homeobox 4; RBAK: RB Associated KRAB Zinc Finger; ZBED9: zinc finger BED-type containing 9; ZMAT4: Zinc Finger Matrin-Type 4; ZNF101: Zinc Finger Protein 101; ZNF12: Zinc Finger Protein12; ZNF281: Zinc Finger Protein 281; ZNF316: Zinc Finger Protein 316; ZNF831: Zinc Finger Protein 831.

AVAILABILITY OF DATA AND MATERIALS

The data used to support the findings of this study are available from public database. The single-cell RNA-seq dataset of CC was downloaded from GEO database (GSE No: GSE168652), the bulk RNA-seq data and clinical information were extracted from the TCGA (https://portal.gdc.cancer.gov/). The data presented in this study are available on reasonable request from the corresponding author.

AUTHOR CONTRIBUTIONS

MMY and CSL—designed the study. SRC and XRL—analyzed the data and wrote the paper. TTH—answered the response. XC, QL and XYT—assisted in the completion of the paper.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

ACKNOWLEDGMENT

Not applicable.

FUNDING

This research was funded by the National Natural Science Foundation of China-Youth Science Foundation Project (82002640), the Natural Science Research Project of Jiangsu Province University (20KJB320028), and the Natural Science Foundation of Nanjing University of Traditional Chinese Medicine (XZR2020070).

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

[1] Aalijahan H, Ghorbian S. Long non-coding RNAs and cervical cancer. Experimental and Molecular Pathology. 2019; 106: 7–16.

[2] Gupta SM, Mania-Pramanik J. Retracted article: molecular mechanisms in progression of HPV-associated cervical carcinogenesis. Journal of Biomedical Science. 2019; 26: 28.

[3] Small W, Bacon MA, Bajaj A, Chuang LT, Fisher BJ, Harkenrider MM, *et al*. Cervical cancer: a global health crisis. Cancer. 2017; 123: 2404–2412.

[4] Naga CH P, Gurram L, Chopra S, Mahantshetty U. The management of locally advanced cervical cancer. Current Opinion in Oncology. 2018; 30: 323–329.

[5] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, *et al*. The human transcription factors. Cell. 2018; 172: 650–665.

[6] Fan C, Du J, Liu N. Identification of a transcription factor signature that can predict breast cancer survival. Computational and Mathematical Methods in Medicine. 2021; 2021: 2649123.

[7] Li M, Wang H, Li W, Peng Y, Xu F, Shang J, *et al*. Identification and validation of an immune prognostic signature in colorectal cancer. International Immunopharmacology. 2020; 88: 106868.

[8] Zhang B, Wang H, Guo Z, Zhang X. A panel of Transcription factors identified by data mining can predict the prognosis of head and neck squamous cell carcinoma. Cancer Cell International. 2019; 19: 297.

[9] Taylor JMG. Random survival forests. Journal of Thoracic Oncology. 2011; 6: 1974–1975.

[10] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, *et al*. Integrated analysis of multimodal single-cell data. Cell. 2021; 184: 3573–3587.e29.

[11] Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, *et al*. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology. 2019; 20: 163–172.

[12] 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 Research. 2015; 43: e47.

[13] Yu G, Wang L, Han Y, He Q. ClusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology. 2012; 16: 284–287.

[14] Díaz-Coto S, Martínez-Camblor P, Pérez-Fernández S. SmoothROCtime: an R package for time-dependent ROC curve estimation. Computational Statistics. 2020; 35: 1231–1251.

[15] 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: 394–424.

[16] Pan J, Xu L, Pan H. Development and validation of an m6A RNA methylation regulator-based signature for prognostic prediction in cervical squamous cell carcinoma. Frontiers in Oncology. 2020; 10: 1444.

[17] Gao C, Zhou C, Zhuang J, Liu L, Liu C, Li H, *et al*. MicroRNA expression in cervical cancer: novel diagnostic and prognostic biomarkers. Journal of Cellular Biochemistry. 2018; 119: 7080–7090.

[18] Ao X, Ding W, Ge H, Zhang Y, Ding D, Liu Y. PBX1 is a valuable prognostic biomarker for patients with breast cancer. Experimental and Therapeutic Medicine. 2020; 20: 385–394.

[19] Liu Y, Ao X, Zhou X, Du C, Kuang S. The regulation of PBXs and their emerging role in cancer. Journal of Cellular and Molecular Medicine. 2022; 26: 1363–1379.

[20] Morgan R, Pandha HS. PBX3 in cancer. Cancers. 2020; 12: 431.

[21] Song Y, Ma R. Identifying the potential roles of PBX4 in human cancers based on integrative analysis. Biomolecules. 2022; 12: 822.

[22] Martinou EG, Moller-Levet CS, Angelidi AM. PBX4 functions as a potential novel oncopromoter in colorectal cancer: a comprehensive analysis of the PBX gene family. American Journal of Cancer Research. 2022; 12: 585.

[23] Badaloni A, Casoni F, Croci L, Chiara F, Bizzoca A, Gennarini G, *et al*. Dynamic Expression and New Functions of Early B cell factor 2 in cerebellar development. The Cerebellum. 2019; 18: 999–1010.

[24] Chen G, Yu W, Li Z, Wang Q, Yang Q, Du Z, *et al*. Potential regulatory effects of miR-182-3p in osteosarcoma *via* targeting EBF2. BioMed Research International. 2019; 2019: 4897905.

[25] Moruzzo D, Nobbio L, Sterlini B, Consalez GG, Benfenati F, Schenone A, *et al*. The transcription factors EBF1 and EBF2 are positive regulators of myelination in Schwann cells. Molecular Neurobiology. 2017; 54: 8117–8127.

[26] Nikitina AS, Sharova EI, Danilenko SA, Butusova TB, Vasiliev AO, Govorov AV, *et al*. Novel RNA biomarkers of prostate cancer revealed by RNA-seq analysis of formalin-fixed samples obtained from Russian patients. Oncotarget. 2017; 8: 32990–33001.

[27] Liao Y, Zou X, Wang K, Wang Y, Wang M, Guo T, *et al*. Comprehensive analysis of transcription factors identified novel prognostic biomarker in human bladder cancer. Journal of Cancer. 2021; 12: 5605–5621.

[28] Li M, Shen Y, Wang Q, Zhou X. MiR-204-5p promotes apoptosis and inhibits migration of osteosarcoma *via* targeting EBF2. Biochimie. 2019; 158: 224–232.

[29] Yan D, Shen M, Du Z, Cao J, Tian Y, Zeng P, *et al*. Developing ZNF gene signatures predicting radiosensitivity of patients with breast cancer. Journal of Oncology. 2021; 2021: 9255494.

[30] Kaya IH, Al-Harazi O, Kaya MT, Colak D. Integrated analysis of transcriptomic and genomic data reveals blood biomarkers with diagnostic and prognostic potential in non-small cell lung cancer. Frontiers in Molecular Biosciences. 2022; 9: 774738.

How to cite this article: Siru Chen, Xuerou Li, Tingting He, Xin Chen, Xiaoyu Tang, Qin Lu, *et al*. Development of a novel transcription factor signature for accurate cervical cancer prognosis. European Journal of Gynaecological Oncology. 2023; 44(5): 55-66. doi: 10.22514/ejgo.2023.079.