Integrated Bioinformatics Analysis of Differentially-Expressed Genes and Immune Cell Infiltration Characteristics in Esophageal Squamous Cell Carcinoma

Zitong Feng
Shandong University
Jingge Qu
Chinese Academy of Medical Sciences & Peking Union Medical College
Xiao Liu
Shandong University
Jinghui Liang
Shandong University
Yongmeng Li
Shandong University
Jin Jiang
Shandong University
Huiying Zhang
Shandong University
Hui Tian (tianhuiql@126.com)
Shandong University

Research Article

Keywords: ESCC, GEO, DEGs, RRA

Posted Date: February 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-153887/v1

Version of Record: A version of this preprint was published at Scientific Reports on August 17th, 2021.
See the published version at https://doi.org/10.1038/s41598-021-96274-y.

Abstract

Esophageal squamous cell carcinoma (ESCC) is a life-threatening thoracic tumor with a poor prognosis. Identifying the best-targeted therapy, appropriate biomarkers and individual treatment for patients with ESCC remains a significant challenge. The present study aimed to elucidate key candidate genes and immune cell infiltration characteristics in ESCC by integrated bioinformatics analysis. We downloaded nine gene expression datasets from the Gene Expression Omnibus (GEO) database. Differentially expressed genes (DEGs) between ESCC tissues and normal tissues in each dataset were identified by the "limma" R package, and a total of 152 robust DEGs were identified by robust rank aggregation (RRA) algorithm. Functional enrichment analyses of the robust DEGs showed that these genes were significantly associated with extracellular matrix related process. Immune cell infiltration analysis was also conducted by CIBERSORT algorithm. We found that M0 and M1 macrophages were increased dramatically in ESCC while M2 macrophages decreased. Nine hub genes were picked out from a protein-protein interaction (PPI) network used by the CytoHubba plugin in Cytoscape. According to the receiver operating characteristic (ROC) curves and Kaplan-Meier survival analysis, the genes PLAU, SPP1 and VCAN had high diagnostic and prognostic values for ESCC patients. Based on univariate and multivariate regression analyses, seven genes (IL18, PLAU, ANO1, SLCO1B3, CST1, NELL2 and MAGEA11) from the robust DEGs were used to construct a good prognostic model. A nomogram that incorporates seven genes signature was established to develop a quantitative method for ESCC prognosis. Our results might provide aid for exploring potential therapeutic targets and prognosis evaluation in ESCC.

Introduction

Esophageal cancer is the seventh most common cancer worldwide, with an estimated 572,034 new cases and 508,585 deaths in 2018[1]. Esophageal squamous cell carcinoma (ESCC) accounts for about 90% of new incident esophageal cancers each year[2]. Due to the inconspicuous symptom and inadequate endoscopic screening, patients with esophageal cancer were often diagnosed at an advanced stage, with 5-year overall survival (OS) rates ranging from 12–20%[3]. In recent years, minimally invasive esophagectomy (MIE), neoadjuvant chemoradiotherapy, targeted therapy and immunotherapy are emerging. These multimodality treatment advances have shown promising results, but a substantial fraction of patients failed to benefit, and the massive burden in terms of new ESCC cases may continue to rise given population growth and aging. Therefore, a much more comprehensive analysis of the molecular mechanisms and underlying immune microenvironment are needed to make more progress in ESCC, and it is also necessary to seek novel biomarkers used for the early detection and prognosis evaluation of ESCC.

Several studies have discovered a group of ESCC related candidate genes by public databases and high-throughput platforms such as Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). Zhang *et al.*[4] identified 345 ESCC differentially expressed genes (DEGs) based on three GEO datasets, and five hub genes can be used as potential prognostic biomarkers. Wang *et al.*[5] constructed a circRNA-miRNA-mRNA axis by integrated bioinformatics analysis of differential expression of miRNAs, predicted

mRNAs, and differential expression of circ-RNAs. They finally elucidated that the circ_0052867/miR-139-5p/RAP1B regulatory network may be a potential biomarker for ESCC patient survival. Using 5 GEO datasets, Karagoz *et al.*[6] analyzed transcriptional regulatory networks, reporter metabolic features and molecular pathways mediating ESCC development. The size of datasets and samples for these ESCC integrated omics are relatively small. The robust rank aggregation (RRA) algorithm can reduce outliers and inconsistent results caused by different platforms and analysis methods. There have been no reports of the use of the RRA algorithm in ESCC to the best of our knowledge.

This study analyzed nine mRNA gene chip datasets from GEO and identified robust (DEGs) between ESCC and normal tissues. Functions of these robust DEGs were then explored by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses. The infiltration characteristics of 22 types of immune cells were conducted by CIBERSORT algorithm. Nine hub genes were elected from robust DEGs by constructing the protein-protein interaction (PPI) network. The genes PLAU, SPP1 and VCAN had high diagnostic and prognostic values for patients with ESCC. A seven-gene signature for the prognosis of patients with ESCC was constructed by Univariate and Multivariable Cox regression analyses, which may serve as an independent prognostic factor for ESCC. Our results would help understand the molecular mechanisms of tumorigenesis and develop new therapeutic targets in ESCC.

Results

Identification of DEGs and robust DEGs.

In the present study, we conducted a systematic analysis of the biological characteristics of DEGs from nine GEO datasets (Table 1). The overall study design was illustrated in Fig. 1. A total of 665 tissue samples including 343 ESCC and 322 normal tissues were analyzed. According to the cutoff criteria of $|\log2\ FC| > 2$ and adjusted $P < 0.05$, 226 DEGs in GSE17351, 219 DEGs in GSE20347, 389 DEGs in GSE29001, 108 DEGs in GSE38129, 692 DEGs in GSE45670, 686 DEGs in GSE53625, 387 DEGs in GSE70409, 223 DEGs in GSE75241 and 147 DEGs in GSE161533 were identified. Among the DEGs, 110, 56, 168, 38, 249, 204, 115, 124 and 57 genes were upregulated while 116, 163, 221, 70, 443, 482, 272, 99 and 90 genes were downregulated in GSE17351, GSE20347, GSE29001, GSE38129, GSE45670, GSE53625, GSE70409, GSE75241 and GSE161533, respectively. For visualizing distributions of DEGs, the volcano plots (Fig. 2A-I) and heat maps (Supplementary Fig. S1) were drawn. Based on the RRA algorithm results, a total of 152 robust DEGs were determined, including 54 upregulated and 98 downregulated genes (Supplementary Table S1). According to the adjusted P of robust DEGs, the top 20 upregulated and downregulated robust DEGs were shown in a heat map (Fig. 2J).

GO and KEGG enrichment analysis of robust DEGs

To explore the biological classification of the 152 robust DEGs in ESCC, we used the “clusterprofiler” R package for GO and KEGG pathway enrichment analysis. GO enrichment analysis in the category biological process suggested that the robust DEGs were mainly accumulated in “extracellular matrix organization”, “extracellular structure organization” and “skin development” (Fig. 3A). In the cellular

component category, the robust DEGs were mainly enriched in “collagen-containing extracellular matrix”, “apical part of cell” and “endoplasmic reticulum lumen” (Fig. 3B). In the molecular function category, the robust DEGs were mainly involved in “endopeptidase activity”, “receptor ligand activity” and “signaling receptor activator activity” (Fig. 3C). KEGG pathway analysis indicated that the robust DEGs were mainly related to “IL-17 signaling pathway”, “Cytokine-cytokine receptor interaction” and “Viral protein interaction with cytokine and cytokine receptor” (Fig. 3D). The above results suggested that robust DEGs were positively associated with cancer cell development.

Immune Cell Infiltration Characteristics

Infiltrating cells of the immune system in the tumor microenvironment (TME) are accepted to be generic constituents of tumors [7]. The CIBERSORT algorithm was used to analyze the immune cell infiltration of 665 samples, and the immune infiltration results were filtered with $P < 0.05$ as the standard, then the proportions of 22 immune cells in 149 ESCC samples and 54 normal tissue samples were obtained (Fig. 4A). The heat map (Fig. 4B) and violin diagram (Fig. 4C) further provided visualization of the differences in immune cell distribution between the ESCC and normal samples. The results showed that 7 immune cells (naïve $CD4^+$T cells, activated memory $CD4^+$T cells, follicular helper T cells, resting NK cells, M0 macrophages, M1 macrophages and activated dendritic cells) were in a higher proportion in the ESCC tissues than those in the normal tissues, whereas 6 immune cells (naïve B cells, resting memory $CD4^+$T cells, gamma delta T cells, M2 macrophages, resting dendritic cells and resting mast cells) were in a higher proportion in the normal tissues. As demonstrated in the principal component analysis (PCA) results (Fig. 4D), ESCC and normal samples can be roughly distinguished using the 22 immune cells.

PPI network construction and hub genes identification

To further study the interaction of the 152 robust DEGs, we construct a PPI network using STRING database and Cytoscape software with a combined score >0.4 as the cutoff criterion. As shown in Fig. 5A, the PPI network covered 91 nodes and 304 edges, including 45 upregulated genes and 46 downregulated genes. Subsequently, the cytoHubba plugin was used to calculate the scores of topological algorithms in each node. The top 50 scoring genes calculated by each of the 12 algorithms (MCC, DMNC, MNC, Degree, EPC, BottleNeck, EcCentricity, Closeness, Radiality, Betweenness, Stress and ClusteringCoefficient) were intersected to obtain hub genes (Fig. 5B). The nine hub genes were cytidine deaminase (CDA), chemokine ligand 1(CXCL1), insulin-like growth factor binding protein 3(IGFBP3), matrix metallopeptidase 3(MMP3), matrix metallopeptidase 11(MMP11), plasminogen activator, urokinase (PLAU), Serpin Family E Member 1(SERPINE1), Secreted Phosphoprotein 1(SPP1) and Versican (VCAN). The mRNA expression of 9 hub genes were validated using the Gene Expression Profiling Interactive Analysis (GEPIA) database. As demonstrated in Fig. 6A-I, the mRNA expression levels of hub genes were markedly upregulated in Esophageal carcinoma (ESCA) tissues compared to those in normal tissues ($P < 0.01$). Moreover, ROC curves were generated to verify the diagnostic performance of nine genes based on the GSE53625 database. The area under the curve (AUC) of CDA, CXCL1, IGFBP3, MMP3, MMP11, PLAU, SERPINE1, SPP1 and VCAN were 0.8816, 0.8303, 0.9627, 0.9462, 0.9975, 0.9822, 0.9344,

0.9890 and 0.9454, respectively (Fig. 7A-I). Kaplan-Meier survival curves showed that high expression of PLAU (P < 0.001), SPP1 (P = 0.024) and VCAN (P = 0.031) were significantly correlated with poor prognosis (Fig. 7J-L).

Survival model construction and analysis

To investigate the prognostic significance of 152 robust DEGs, a total of 17 survival-related genes (P < 0.05) were identified by the Univariate Cox analysis (Table 2). After selecting the most suitable combination of candidate genes by multiple stepwise Cox regression, seven genes including Interleukin 18 (IL18), Plasminogen Activator, Urokinase (PLAU), Anoctamin 1 (ANO1), Solute Carrier Organic Anion Transporter Family Member 1B3 (SLCO1B3), Cystatin SN (CST1), Neural EGFL Like 2 (NELL2) and MAGE Family Member A11 (MAGEA11) were used to build a prognostic model (Table 3). The risk score of each patient was calculated according to the following formula: (−0.2232 × ExpIL18) + (0.4659 × ExpPLAU) + (0.1876 × ExpANO1) + (−0.0921 × ExpSLCO1B3) + (0.1844 × ExpCST1) + (−0.1203 × ExpNELL2) + (−0.1501 × ExpMAGEA11). 179 ESCC patients in GSE53625 were divided into a low-risk group and a high-risk group according to the median risk score. Kaplan-Meier curve demonstrated that the prognosis of low-risk group was significantly better than that of the high-risk group (P = 1.979e-08) (Fig. 8A). The time-dependent ROC curve was plotted, and the fact that the AUC of the risk score was 0.777 for a 5-year survival prediction indicates the exactitude of this model (Fig. 8B). As shown in Fig. 8C, the expression heat map of 7 prognostic genes was profiled. Cox regression analysis was carried out to demonstrate whether the model can be an available independent prognostic indicator. Univariate Cox regression analysis showed that N staging (P < 0.001), tumor stage (P < 0.001) and risk score (P < 0.001) were significantly correlated with prognosis (Fig. 8D). Multivariate Cox regression analysis showed that only the risk score (P< 0.001) was significantly correlated with prognosis (Fig. 8E). To better predict the prognosis of patients with ESCC at 1, 3, and 5 years after esophagectomy, we integrated the seven genes signature to establish a nomogram (Fig. 8F). A higher total point indicates a lower overall survival.

Discussion

Esophageal cancer is one of the most common cancers with high mortality worldwide due to late diagnosis and lack of efficient treatment. Esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC) are distinct diseases in terms of the cell of origin, epidemiology and molecular architecture of tumor cells[8]. Almost 50% of all esophageal cancer cases occur in China, and ESCC is the most dominant subtype[9]. An in-depth study using bioinformatics analysis on the pathogenesis and molecular mechanisms of ESCC development and progression is a practical necessity. Better biomarkers for specific prognosis and progression of ESCC are also demanded. To minimize inter-study variability and complicated statistical analyses, we integrated 9 GEO datasets using the RRA method. In our study, we identified a total of 152 robust DEGs consisting of 54 upregulated and 98 downregulated genes between ESCC and normal tissues. The GO and KEGG functional enrichment analyses demonstrated that the robust DEGs were significantly (adjusted P < 0.05) associated with extracellular matrix related process such as collagen − containing extracellular matrix, extracellular matrix organization, ECM − receptor

interaction and extracellular matrix structural constituent. The extracellular matrix (ECM) mechanics can regulate and maintain tissue cell development[10]. Dysregulation of ECM promotes tumor progression and tumor microenvironment formation[11]. The cancer hallmarks are affected by biophysical and biochemical signals from tumor-associated ECM. It has been demonstrated that the mechanical properties and configuration of ECM play important roles in sustaining proliferation, evading growth suppression, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion, avoiding immune destruction, deregulating cellular energetics, genomic instability and tumor-promoting chronic inflammation[12]. A large quantity of interactions between esophageal cancer cells and extracellular matrix seems to be intricate[13], and many studies have demonstrated that different extracellular matrix molecules play a regulatory role in the development and metastasis of ESCC[14–19]. Enrichment of robust DEGs in some KEGG pathways including IL − 17 signaling pathway[20], Cytokine – cytokine receptor interaction[21] and TNF signaling pathway[22] also demonstrated their relationship with tumorigenesis.

The tumor microenvironment (TME) is a complex environment in which tumor cells survive, mainly composed of various stromal cells and extracellular matrix[23]. Among the stromal cells, tumor-associated macrophages (TAMs) and their precursors account for the most considerable fraction of the myeloid infiltrate in the majority of solid human malignancies[24]. TAMs differentiate into M1 and M2 macrophages under the action of TME. M1 macrophages mainly play an anti-tumor effect; M2 macrophages secrete many immunoregulatory factors such as cytokines, chemokines and metalloproteinases, which affecting most aspects of ESCC progress by promoting tumor angiogenesis and lymphangiogenesis[25, 26]. In our study, the CIBERSORT algorithm showed that the infiltration level of M0 and M1 macrophages in ESCC was most significantly higher than that in normal tissues. However, M2 macrophages were less common in ESCC than in normal tissue. This “paradoxical” distribution of M2 macrophages may occur due to the high dynamics and heterogeneity of the TAMs compartment. Our study only presented the general infiltration of immune cells in ESCC, and further study was needed to investigate the diverse roles of immune cells in TME.

We identified nine hub genes among the robust DEGs by constructing a PPI network. The nine hub genes including CDA, CXCL1, IGFBP3, MMP3, MMP11, PLAU, SERPINE1, SPP1 and VCAN were all upregulated in ESCC compared with normal tissues, which was verified in the GEPIA database. In virtue of the abnormally high expression of the top 9 hub genes in ESCC, we are interested in whether they can serve as promising diagnostic biomarkers for ESCC patients. The ROC curves showed that the top 9 hub genes had relatively high diagnostic values for ESCC patients. Likewise, exploring the relationship between gene expression and prognosis of ESCC patients has important clinical significance for finding therapeutic targets of ESCC. Kaplan-Meier survival analysis of hub genes showed that the high expression of PLAU, SPP1 and VCAN genes were significantly associated with poor prognosis in ESCC patients. The three key genes were screened to explore their roles. PLAU (also named uPA), a serine protease involved in the degradation of the extracellular matrix, binds its receptor (PLAUR) to initiate a proteolytic cascade that converts plasminogen to plasmin leading to tumor cell invasion and metastasis[27]. In recent years, further progress was made in the study of PLAU in tumors. Wang *et al.* found that amiloride, a type of

synthetic PLAU inhibitor, could inhibit cervical cancer invasion by suppressing MMP2 expression[28]. Triptolide could inhibit proliferation and migration of human pancreatic cancer cells through targeting PLAU, which activated endothelial-mesenchymal transition (EMT) progression, and overexpression of PLAU protein was related to lymph node metastasis in pancreatic patients[29]. High expression of PLAU related to p38MAPK could indicate poor prognosis in esophageal cancer, and there is a close relationship between the two proteins[30]. Circulating PLAU mRNA in peripheral blood can serve as a potential unfavorable prognosis biomarker in ESCC[31]. PLAU could be a good therapeutic target for ESCC. In addition to bioinformatics analysis, *in vivo and in vitro* experiments can enhance our comprehension of its functional role in ESCC. SPP1, also called Osteopontin (OPN), is a multifunctional extracellular matrix phosphoprotein secreted by several cell types and is involved in various biological functions including wound healing and bone calcification, immune response and tumor progression[32, 33]. It has been reported that SPP1 has a multifaceted involvement in various tumors. For example, upregulation of SPP1 could affect tumor progression, prognosis, and resistance to cetuximab via the KRAS/MEK pathway in head and neck cancer[34]. SPP1 promotes aggressiveness of lung cancer cells through its phosphorylation activation of the Recepteur d'Origine Nantais(RON) signaling pathway[35]. Targeting SPP1 by microRNA-340 inhibits gastric cancer cell proliferation, migration and EMT process by inhibiting the PI3K/AKT signaling pathway[36]. A meta-analysis of 8 studies showed that SPP1 overexpression might serve as an excellent independent prognostic risk factor in 811 Chinese and Japanese ESCC patients[37]. SPP1 may serve a role in the development of ESCC, and further studies should be performed to explore the value of SPP1 as a therapeutic target in the treatment of ESCC. VCAN, a chondroitin sulfate proteoglycan, is also an essential extracellular matrix component. Previous studies showed that VCAN had been closely associated with the proliferation and metastasis of various tumor cells such as gastric cancer, leukemia and breast cancer[38–40]. However, its role in the progression of ESCC was still unclear. The three key genes are all extracellular matrix-related genes, which are consistent with our functional enrichment analysis results.

It is of crucial clinical significance to stratify patients with ESCC and construct a prognosis prediction model. Li *et al.*[41] constructed an eight-lincRNA prognostic signature and nomogram based on the GEO and TCGA database to improve the prognostic value of ESCC. Mao *et al.*[42] found a novel six-miRNA signature could be an independent biomarker for survival prediction of ESCC patients. We used Univariate Cox regression analysis to identify the robust DEGs associated with prognosis, and Multivariate Cox regression analysis was further performed to select the survival-associated robust DEGs. Finally, seven genes (IL18, PLAU, ANO1, SLCO1B3, CST1, NELL2 and MAGEA11) were used to construct a Cox regression risk model that can predict the outcome of high and low-risk groups. We found that the risk score could be used as an available independent prognostic indicator. A nomogram analysis suggested that the 1, 3, and 5-year survival rates for ESCC can be intuitively predicted based on the relative expression level of 7 genes. There are still limitations in these predictive model analyses; larger sample sizes and other databases are needed to verify these results.

In conclusion, we performed differential gene expression analysis on nine GEO datasets of ESCC using the RRA method. The immune cell infiltration was characterized, and nine hub genes were screened. Our

results revealed that genes PLAU, SPP1 and VCAN might be considered useful biomarkers of ESCC. Moreover, we identified a 7-gene prognostic signature as a potential prognostic predictor for ESCC patients. Further experiments are required to validate the current findings in the process of ESCC.

Methods

Microarray data collection

Microarray datasets GSE17351, GSE20347, GSE29001, GSE38129, GSE45670, GSE53625, GSE70409, GSE75241 and GSE161533 were obtained from GEO database (https://www.ncbi.nlm.nih.gov/geo/)[43]. The basic information for the nine GEO datasets in the current study is provided in Table 1. For these datasets, only ESCC tissue and normal tissue samples were selected for further analysis. Each included dataset contains at least ten samples. Besides, clinical data were downloaded from the GSE53625 dataset and utilized in the study.

Differential expression analysis in ESCC

We used the “limma” package[44] in R software (version 3.6.3, https://www.r-project.org/) to identify differentially expressed genes (DEGs) between ESCC tissues and normal tissues with the cutoff criteria of |log2 fold change (FC)| > 2 and adjusted P < 0.05. After the upregulated and downregulated genes were ranked by their FC in each dataset, we utilized the robust rank aggregation (RRA) algorithm to integrate the nine microarray datasets[45]. The RRA algorithm could avoid substantial heterogeneity and the error of each experiment caused by different technological platforms and difficult statistical methods. Then, the “RobustRankAggreg” R package was performed to obtain robust DEGs. The genes with |log2 FC| > 2 and adjusted P < 0.05 were considered as significant robust DEGs.

Functional and pathway enrichment analysis

To determine the biological annotation of the robust DEGs indicated above, Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were conducted using the “clusterprofiler” R package[46]. The GO analysis terms included biological process (BP), cellular component (CC) and molecular function (MF). Adjusted P < 0.05 was considered to indicate a statistically significant difference.

Immune Cells Infiltration analysis by CIBERSORT algorithm

The proportions of 22 immune cell types with gene expression profiles in each tissue sample were estimated by CIBERSORT algorithm (http://cibersort.stanford.edu/)[47]. The 22 kinds of immune cells include nine types of adaptive immunity cells [memory B cells, naïve B cells, activated memory $CD4^+$T cells, resting memory $CD4^+$T cells, naïve $CD4^+$T cells, $CD8^+$T cells, follicular helper T cells, regulatory T cells(Tregs) and gamma delta T cells] and 13 types of innate immunity cells [Activated dendritic cells, resting dendritic cells, eosinophils, macrophages (M0–M2), activated mast cells, resting mast cells,

monocytes, resting natural killer (NK) cells, activated NK cells, neutrophils and plasma cells]. All gene expression matrixes were normalized and converted to 22 kinds of immune cell matrix by the CIBERSORT algorithm. R packages evaluated the differences in the 22 immune cells subpopulations in ESCC and normal samples according to the filtered criteria of $P < 0.05$. The principal component analysis (PCA) was also performed to indicate the difference between ESCC and normal samples using the 22 immune cells[48].

Identification and validation of hub genes

The online tool Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org/) was used to obtain the predicted interactions for the robust DEGs with medium confidence >0.4[49]. The protein-protein interaction (PPI) network of the robust DEGs was visualized with the Cytoscape software (Version 3.72, http://www.cytoscape.org/)[50]. The CytoHubba plugin in Cytoscape provides 12 different algorithms to analyze the topology of the PPI network, and it consists of Maximal Clique Centrality (MCC), Density of Maximum Neighborhood Component (DMNC), Maximum Neighborhood Component (MNC), Degree, Component (EPC), BottleNeck, EcCentricity, Closeness, Radiality, Betweenness, Stress and ClusteringCoefficient[51]. These algorithms perform together to identify hub genes. Furthermore, the differential expression of hub genes in ESCC was validated using GEPIA database (http://gepia.cancer-pku.cn/)[52]. GSE53625 dataset was used to evaluate the diagnostic performance and Kaplan-Meier survival analysis of expression levels of hub genes. Receiver-operating characteristic (ROC) curves were utilized to assess performance of hub genes as biomarkers for classifying patients with ESCC. The ROC curve and the area under the curve (AUC) were calculated and drawn using the GraphPad Prism 8.0 (GraphPad Software, Inc., La Jolla, California). Based on each best-separation cutoff value of hub gene, ESCC patients within the GEO dataset were divided into two groups to get the Kaplan-Meier survival curves.

Prognostic model construction

179 ESCC samples with reliable clinical prognostic information from GSE53625 were used to perform survival analysis. Univariate Cox proportional hazards regression analysis was applied on 152 robust DEGs to identify prognosis-related genes using the “survival” R package. Next, based on the above preliminary significant genes, we constructed a Multivariate Cox proportional hazards regression model and calculated a risk score for predicting the prognosis in ESCC patients[53]. The risk score formula of prognostic signature was as follows: Risk score=$\sum(\beta_i \times Exp_i)$ (β_i represented the coefficient value, and EXP_i represented the gene expression level). According to the median risk score, the ESCC patients were divided into low-risk and high-risk groups. The time-dependent ROC curve was performed to assess the predictive power of prognostic model by using the “SurvivalROC” R package[54]. To further explore the correlation between robust DEGs in the prediction model and clinicopathological characteristics, univariate regression analysis and multivariate regression analysis were used to identify independent prognostic factors (including age, gender, grade, T staging, lymph node metastasis, stage and risk score) in patients with ESCC. Additionally, the nomogram with calibration plots was conducted using the “rms” R

package to forecast survival probability for one year, three years and five years. P < 0.05 was considered statistically significant.

Declarations

Data Availability

The datasets generated during and/or analysed during the current study are available in the GEO repository, https://www.ncbi.nlm.nih.gov/geo/.

Acknowledgements

This work was supported by grants from the Major Scientific and Technological Innovation Project of Shandong Province (Grant No.2020CXGC011303), the Taishan Scholar Program of Shandong Province (Grant No.ts201712087) and the National Natural Science Foundation of China (Grant No. 81672292).

Author contributions

H.T., Z.F., and J.Q. conceived and designed the study. J.L., J.J., and Y.L. performed data collection and literature search. Z.F., H.Z., and X.L. analyzed the data and drafted the manuscript. J.Q. and H.T. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394-424. https://doi.org/10.3322/caac.21492 (2018).
2. Abnet, C. C., Arnold, M. & Wei, W.-Q. Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology* **154**, 360-373. https://doi.org/10.1053/j.gastro.2017.08.023 (2018).
3. Napier, K. J., Scheerer, M. & Misra, S. Esophageal cancer: A Review of epidemiology, pathogenesis, staging workup and treatment modalities. *J. Gastrointest. Oncol.* **6**, 112-120. https://doi.org/10.4251/wjgo.v6.i5.112 (2014).
4. Zhang, H. *et al.* Integrated Bioinformatics Analysis Identifies Hub Genes Associated with the Pathogenesis and Prognosis of Esophageal Squamous Cell Carcinoma. *Biomed Res. Int.* **2019**, 2615921. https://doi.org/10.1155/2019/2615921 (2019).
5. Wang, Z., Li, H., Li, F., Su, X. & Zhang, J. Bioinformatics-Based Identification of a circRNA-miRNA-mRNA Axis in Esophageal Squamous Cell Carcinomas. *Oncol.* **2020**, 8813800. https://doi.org/10.1155/2020/8813800 (2020).

6. Karagoz, K., L. Lehman, H., B. Stairs, D., Sinha, R. & Y. Arga, K. Proteomic and Metabolic Signatures of Esophageal Squamous Cell Carcinoma. *Cancer Drug Targets* **16**, 721-736. https://doi.org/10.2174/1568009616666160203113721 (2016).
7. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674. https://doi.org/10.1016/j.cell.2011.02.013 (2011).
8. Talukdar, F. R. *et al.* Molecular landscape of esophageal cancer: implications for early detection and personalized therapy. *N. Y. Acad. Sci.* **1434**, 342-359. https://doi.org/10.1111/nyas.13876 (2018).
9. Arnold, M., Soerjomataram, I., Ferlay, J. & Forman, D. Global incidence of oesophageal cancer by histological subtype in 2012. *Gut* **64**, 381-387. https://doi.org/10.1136/gutjnl-2014-308124 (2015).
10. Mammoto, T. & Ingber, D. E. Mechanical control of tissue and organ development. *Development* **137**, 1407-1420. https://doi.org/10.1242/dev.024166 (2010).
11. Lu, P., Weaver, V. M. & Werb, Z. The extracellular matrix: a dynamic niche in cancer progression. *Cell Biol.* **196**, 395-406. https://doi.org/10.1083/jcb.201102147 (2012).
12. Pickup, M. W., Mouw, J. K. & Weaver, V. M. The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep.* **15**, 1243-1253. https://doi.org/10.15252/embr.201439246 (2014).
13. Palumbo, A. *et al.* Esophageal Cancer Development: Crucial Clues Arising from the Extracellular Matrix. *Cells* **9**. https://doi.org/10.3390/cells9020455 (2020).
14. Yoshinaga, K. *et al.* Activin A enhances MMP-7 activity via the transcription factor AP-1 in an esophageal squamous cell carcinoma cell line. *J. Oncol.* **33**, 453-459 (2008).
15. Xiao, J. *et al.* Expression of fibronectin in esophageal squamous cell carcinoma and its role in migration. *BMC Cancer* **18**, 976. https://doi.org/1186/s12885-018-4850-3 (2018).
16. Kuo, I. Y. *et al.* Low SOX17 expression is a prognostic factor and drives transcriptional dysregulation and esophageal cancer progression. *J. Cancer* **135**, 563-573. https://doi.org/10.1002/ijc.28695 (2014).
17. Li, F. *et al.* Expression of Integrin β6 and HAX-1 Correlates with Aggressive Features and Poor Prognosis in Esophageal Squamous Cell Carcinoma. *Cancer Manag. Res.* **12**, 9599-9608. https://doi.org/10.2147/CMAR.S274892 (2020).
18. Ohtsuka, M. *et al.* Concurrent expression of C4.4A and Tenascin-C in tumor cells relates to poor prognosis of esophageal squamous cell carcinoma. *J. Oncol.* **43**, 439-446. https://doi.org/10.3892/ijo.2013.1956 (2013).
19. Qiao, L. *et al.* Gene silencing of galectin-3 changes the biological behavior of Eca109 human esophageal cancer cells. *Med. Report.* **13**, 160-166. https://doi.org/10.3892/mmr.2015.4543 (2016).
20. Wu, L. *et al.* A novel IL-17 signaling pathway controlling keratinocyte proliferation and tumorigenesis via the TRAF4-ERK5 axis. *Exp. Med.* **212**, 1571-1587. https://doi.org/10.1084/jem.20150204 (2015).
21. Smyth, M. J., Cretney, E., Kershaw, M. H. & Hayakawa, Y. Cytokines in cancer immunity and immunotherapy. *Rev.* **202**, 275-293 (2004).

22. Sethi, G., Sung, B. & Aggarwal, B. B. TNF: a master switch for inflammation to cancer. *Biosci.* **13**, 5094-5107 (2008)
23. Hui, L. & Chen, Y. Tumor microenvironment: Sanctuary of the devil. *Cancer Lett.* **368**. https://doi.org/10.1016/j.canlet.2015.07.039 (2015).
24. Vitale, I., Manic, G., Coussens, L. M., Kroemer, G. & Galluzzi, L. Macrophages and Metabolism in the Tumor Microenvironment. *Cell Metab.* **30**, 36-50. https://doi.org/10.1016/j.cmet.2019.06.001 (2019).
25. Shigeoka, M. *et al.* Tumor associated macrophage expressing CD204 is associated with tumor aggressiveness of esophageal squamous cell carcinoma. *Cancer Sci.* **104**, 1112-1119. https://doi.org/10.1111/cas.12188 (2013).
26. Sun, M.-M. *et al.* The synergistic effect of esophageal squamous cell carcinoma KYSE150 cells and M2 macrophages on lymphatic endothelial cells. *Am J Transl Res* **9**, 5105-5115 (2017).
27. Dass, K., Ahmad, A., Azmi, A. S., Sarkar, S. H. & Sarkar, F. H. Evolving role of uPA/uPAR system in human cancers. *Cancer Treat. Rev.* **34**, 122-136 (2008).
28. Wang, X. *et al.* Effect of a synthetic inhibitor of urokinase plasminogen activator on the migration and invasion of human cervical cancer cells in vitro. *Med. Report.* **17**, 4273-4280. https://doi.org/10.3892/mmr.2018.8414 (2018).
29. Zhao, X. *et al.* Triptolide inhibits pancreatic cancer cell proliferation and migration via down-regulating PLAU based on network pharmacology of Tripterygium wilfordii Hook F. *J. Pharmacol.* **880**, 173225. https://doi.org/10.1016/j.ejphar.2020.173225 (2020).
30. Liu, Q., Li, W., Yang, S. & Liu, Z. High expression of uPA related to p38MAPK in esophageal cancer indicates poor prognosis. *Onco Targets Ther.* **11**, 8427-8434. https://doi.org/10.2147/OTT.S181701 (2018).
31. He, X., Xu, X., Zhu, G. & Ye, H. Circulating uPA as a potential prognostic biomarker for resectable esophageal squamous cell carcinoma. *Medicine (Baltimore).* **98**, e14717. https://doi.org/10.1097/MD.0000000000014717 (2019).
32. McKee, M. D., Pedraza, C. E. & Kaartinen, M. T. Osteopontin and wound healing in bone. *Cells, tissues, organs* **194**. 313-319. https://doi.org/10.1159/000324244 (2011).
33. Lamort, A.-S., Giopanou, I., Psallidas, I. & Stathopoulos, G. T. Osteopontin as a Link between Inflammation and Cancer: The Thorax in the Spotlight. *Cells* **8**. https://doi.org/10.3390/cells8080815 (2019).
34. Liu, K. *et al.* Upregulation of secreted phosphoprotein 1 affects malignant progression, prognosis, and resistance to cetuximab via the KRAS/MEK pathway in head and neck cancer. *Carcinog.* **59**, 1147-1158. https://doi.org/10.1002/mc.23245 (2020).
35. Hao, C. *et al.* OPN promotes the aggressiveness of non-small-cell lung cancer cells through the activation of the RON tyrosine kinase. *Rep.* **9**, 18101. https://doi.org/10.1038/s41598-019-54843-2 (2019).
36. Song, S.-Z. *et al.* Targeting of SPP1 by microRNA-340 inhibits gastric cancer cell epithelial-mesenchymal transition through inhibition of the PI3K/AKT signaling pathway. *Cell. Physiol.* **234**,

18587-18601. https://doi.org/10.1002/jcp.28497 (2019).

37. Wang, Y. *et al.* Prognostic value of osteopontin expression in esophageal squamous cell carcinoma: A meta-analysis. *Res. Pract.* **215**, 152571. https://doi.org/10.1016/j.prp.2019.152571 (2019).
38. Cheng, Y. *et al. V*Up-Regulation of VCAN Promotes the Proliferation, Invasion and Migration and Serves as a Biomarker in Gastric Cancer. *Onco Targets Ther.* **13**, 8665-8675. https://doi.org/10.2147/OTT.S262613 (2020).
39. Yang, L. *et al.* Up-regulation of EMT-related gene VCAN by NPM1 mutant-driven TGF-β/cPML signalling promotes leukemia cell invasion. *Cancer* **10**, 6570-6583. https://doi.org/10.7150/jca.30223 (2019).
40. Zhang, Y. *et al.* Enhanced PAPSS2/VCAN sulfation axis is essential for Snail-mediated breast cancer cell migration and metastasis. *Cell Death Differ.* **26**, 565-579. https://doi.org/10.1038/s41418-018-0147-y (2019).
41. Li, W., Liu, J. & Zhao, H. Identification of a nomogram based on long non-coding RNA to improve prognosis prediction of esophageal squamous cell carcinoma. *Aging* **12**, 1512-1526. https://doi.org/10.18632/aging.102697 (2020).
42. Mao, Y. *et al.* A six-microRNA risk score model predicts prognosis in esophageal squamous cell carcinoma. *Cell. Physiol.* **234**, 6810-6819. https://doi.org/10.1002/jcp.27429 (2019).
43. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **41**, D991-D995. https://doi.org/10.1093/nar/gks1193 (2013).
44. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47. https://doi.org/10.1093/nar/gkv007 (2015).
45. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573-580. https://doi.org/10.1093/bioinformatics/btr709 (2012).
46. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. **16**, 284-287. https://doi.org/10.1089/omi.2011.0118 (2012).
47. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Methods* **12**, 453-457. https://doi.org/10.1038/nmeth.3337 (2015).
48. Ringnér, M. What is principal component analysis? *Biotechnol.* **26**, 303-304. https://doi.org/10.1038/nbt0308-303 (2008).
49. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362-D368. https://doi.org/10.1093/nar/gkw937 (2017).
50. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504 (2003).
51. Chin, C.-H. *et al.* cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8 Suppl 4**, S11. https://doi.org/10.1186/1752-0509-8-S4-S11 (2014).

52. Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**. https://doi.org/10.1093/nar/gkx247 (2017).
53. Bøvelstad, H. M. *et al.* Predicting survival from microarray data--a comparative study. *Bioinformatics* **23**, 2080-2087 (2007).
54. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337-344 (2000).

Tables

Table 1
Basic information of 9 GEO microarray datasets.

Datasets	Year	County	Sample size(T/N)	Platform	Number of rows
GSE17351	2009	USA	5/5	GPL570	54675
GSE20347	2010	USA	17/17	GPL571	22277
GSE29001	2011	USA	24/21	GPL571	22277
GSE38129	2012	USA	30/30	GPL571	22277
GSE45670	2013	China	28/10	GPL570	54675
GSE53625	2013	China	179/179	GPL18109	71584
GSE70409	2013	China	17/17	GPL13287	29187
GSE75241	2015	Brazil	15/15	GPL5175	316919
GSE161533	2020	China	28/28	GPL570	54675

Table 2
Univariate Cox regression analysis of the 17 genes.

Gene	HR	Lower 95%CI	Upper 95%CI	P
MYH11	1.214316	1.016687	1.450361	0.032148
CRCT1	0.81209	0.692012	0.953004	0.010785
IL18	0.792458	0.663518	0.946454	0.010248
CNN1	1.206791	1.01106	1.440414	0.037362
SERPINH1	1.527643	1.064343	2.192614	0.021552
SERPINB2	0.901575	0.817279	0.994566	0.03857
PLAU	1.414795	1.056845	1.893982	0.019729
SULT2B1	0.832297	0.735726	0.941542	0.003532
TMPRSS11E	0.840419	0.736972	0.958387	0.009481
KLK11	0.862082	0.76162	0.975795	0.018897
ANO1	1.265394	1.066067	1.50199	0.007113
SLCO1B3	0.863263	0.776586	0.959614	0.006458
CST1	1.205793	1.055109	1.377997	0.006004
NELL2	0.86076	0.779304	0.95073	0.003116
MAGEA6	0.91218	0.837226	0.993845	0.035634
MAGEA4	0.920913	0.859059	0.98722	0.020204
COL11A1	0.754961	0.571866	0.996678	0.047318
HR, hazard ratio; CI, confidence interval				

Table 3
Multivariate Cox regression analysis of the 7-gene signature.

Gene	coef	HR	Lower 95%CI	Upper 95%CI	P
IL18	-0.22318	0.799971	0.660934	0.968256	0.021956
PLAU	0.465937	1.593507	1.177179	2.157075	0.002563
ANO1	0.187636	1.206394	1.022512	1.423344	0.026161
SLCO1B3	-0.09211	0.912006	0.812737	1.023399	0.117213
CST1	0.184351	1.202437	1.045682	1.382691	0.009689
NELL2	-0.12029	0.886663	0.801369	0.981035	0.019753
MAGEA11	-0.15012	0.860602	0.777913	0.95208	0.003583
HR, hazard ratio; CI, confidence interval					

Figures

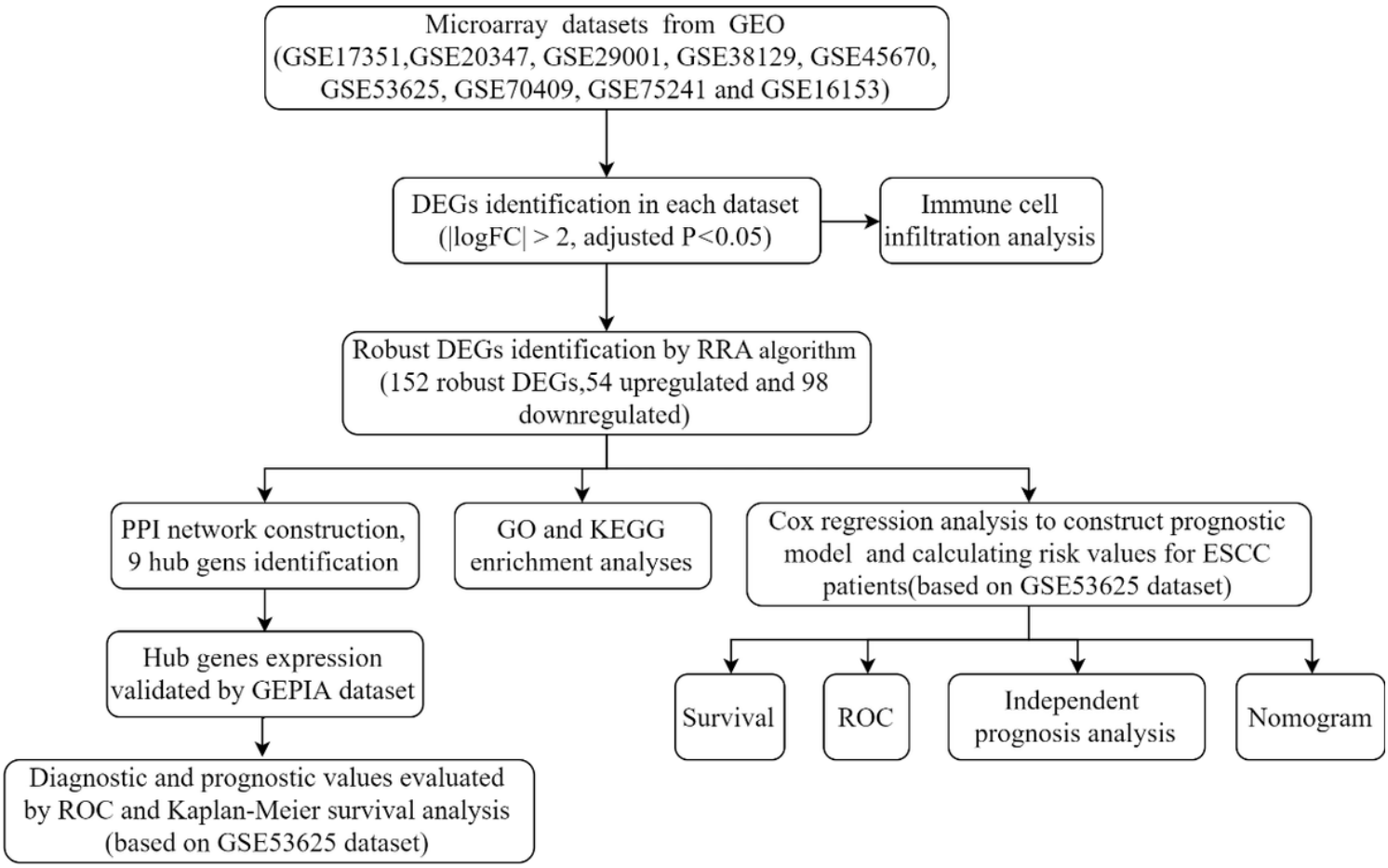

Figure 1

Whole study workflow for analyzing differentially-expressed genes in ESCC.

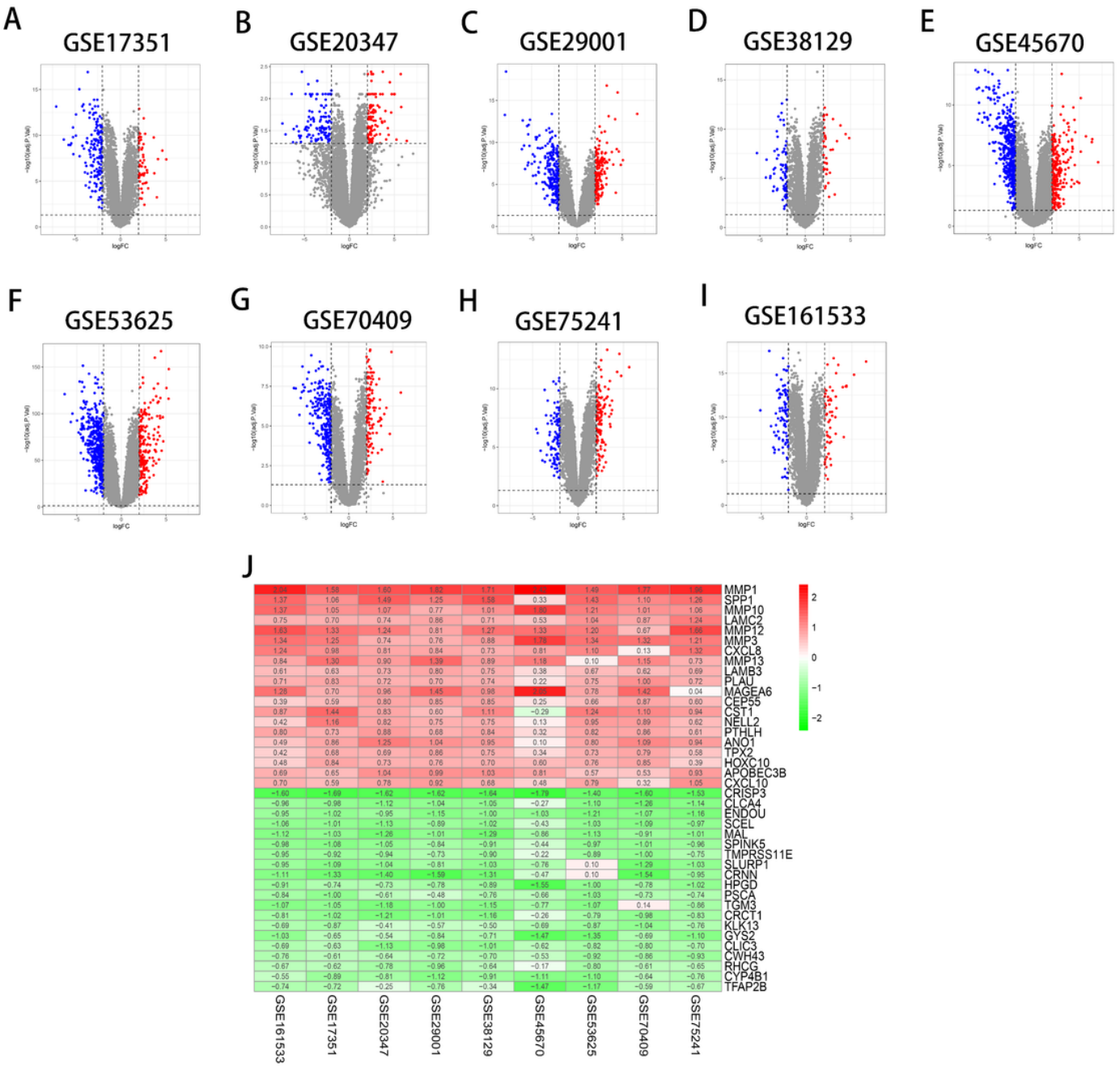

Figure 2

Identification of DEGs and robust DEGs in ESCC vs. normal tissues. The volcano plots of DEGs in GSE17351(A), GSE20347 (B), GSE29001 (C), GSE38129(D), GSE45670(E), GSE53625(F), GSE70409(G), GSE75241(H) and GSE161533(I). Red and blue dots represent the upregulated and downregulated genes, respectively. (J) The heatmap of top 20 upregulated and downregulated robust DEGs identified by the RRA algorithm. Red represents high expression robust DEGs, while green represents low expression robust DEGs. DEGs, differentially expressed genes; RRA, robust rank aggregation.

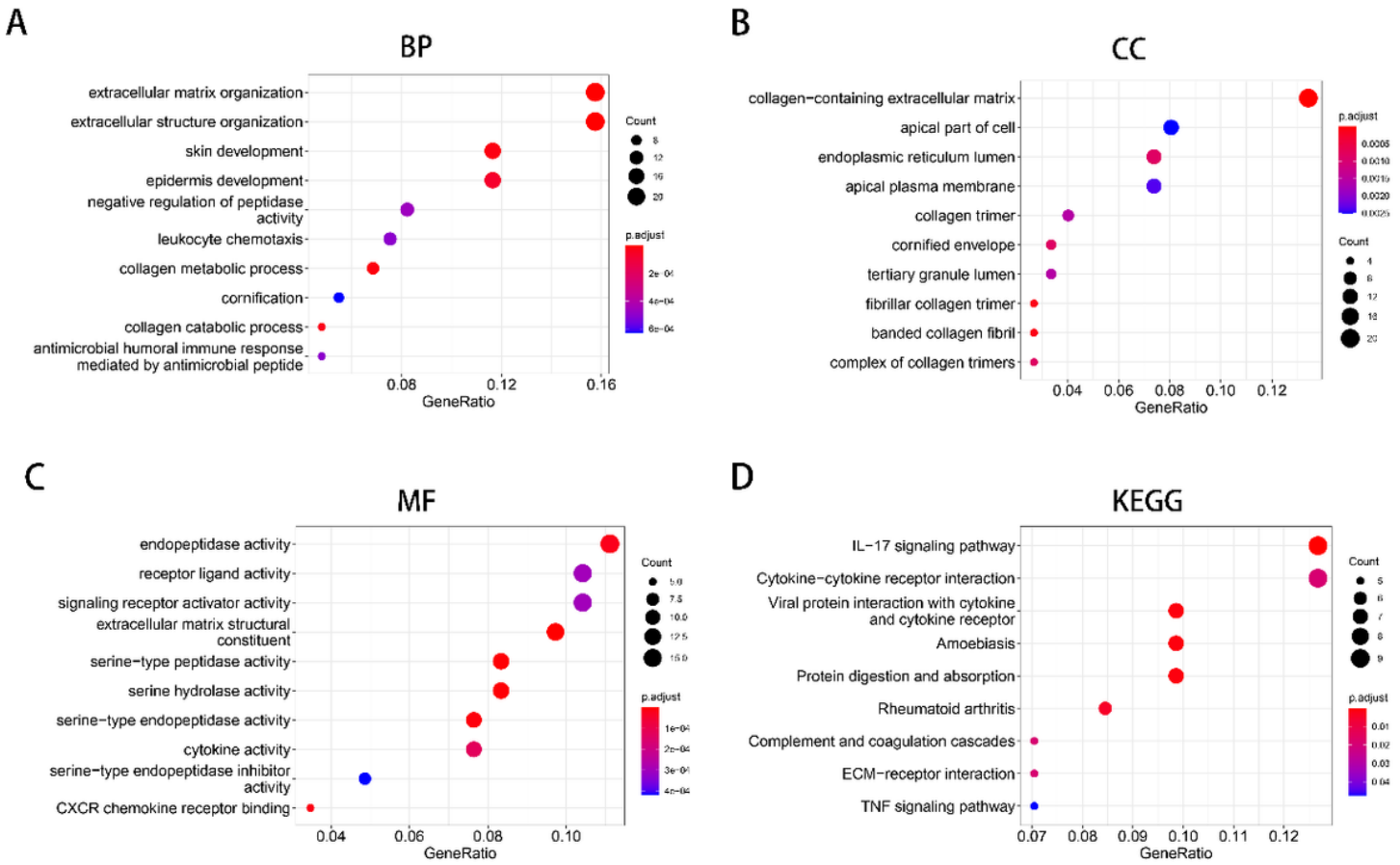

Figure 3

GO and KEGG pathway of robust DEGs in ESCC. (A) Biological process terms for robust DEGs. (B) Cellular component terms for robust DEGs. (C) Molecular function terms for robust DEGs. (D) KEGG analysis for robust DEGs. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

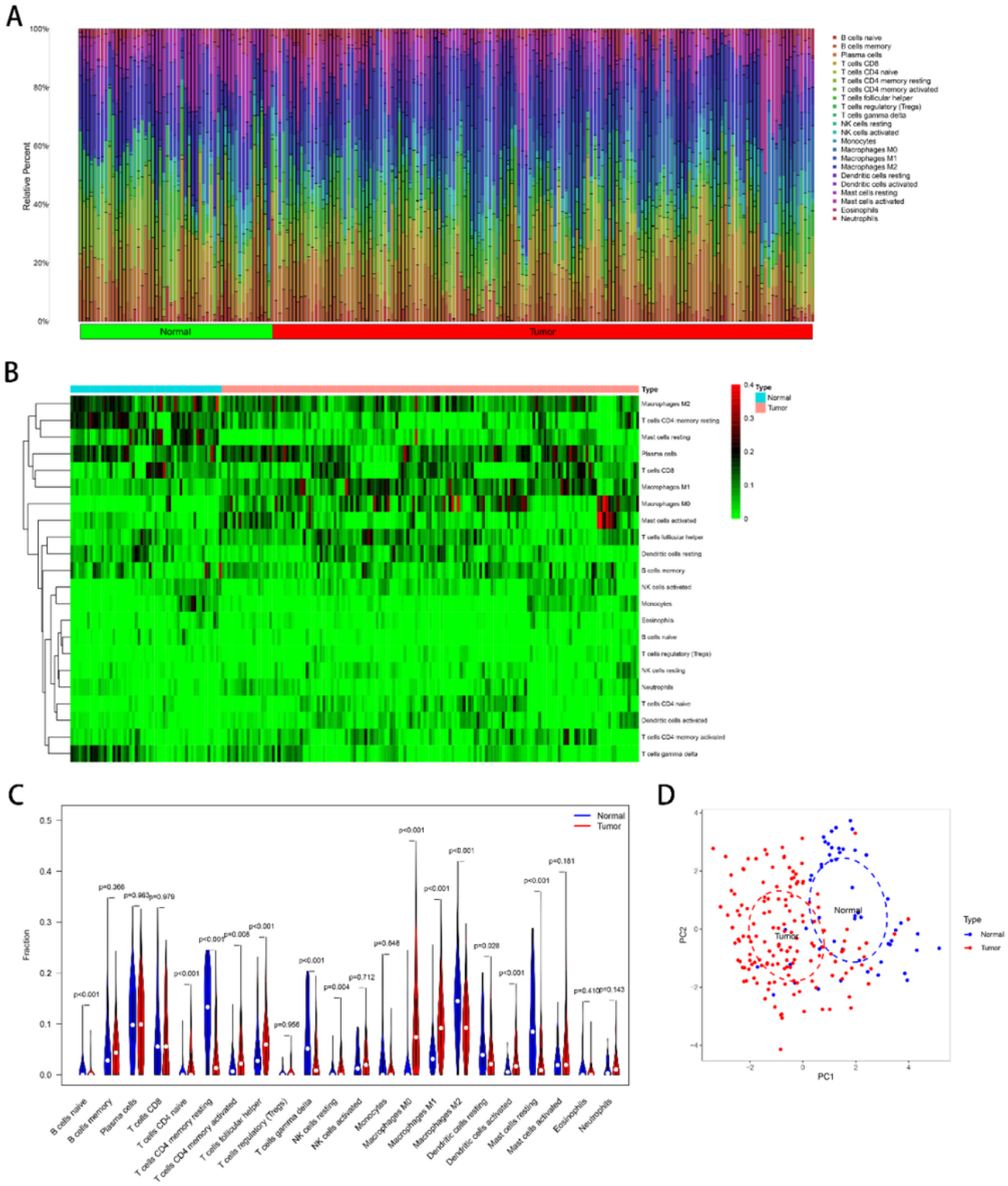

Figure 4

Infiltrating immune cells Characteristics. (A) The fraction of 22 immune cell subpopulations between ESCC and normal tissues. (B) Heat map visualizing the difference of immune cells between ESCC and normal tissues. (C) Violin plot showing the differentially infiltrated immune cells ($P < 0.05$). (D) The principal component analysis showed that 22 immune cells could roughly distinguish ESCC and normal tissues.

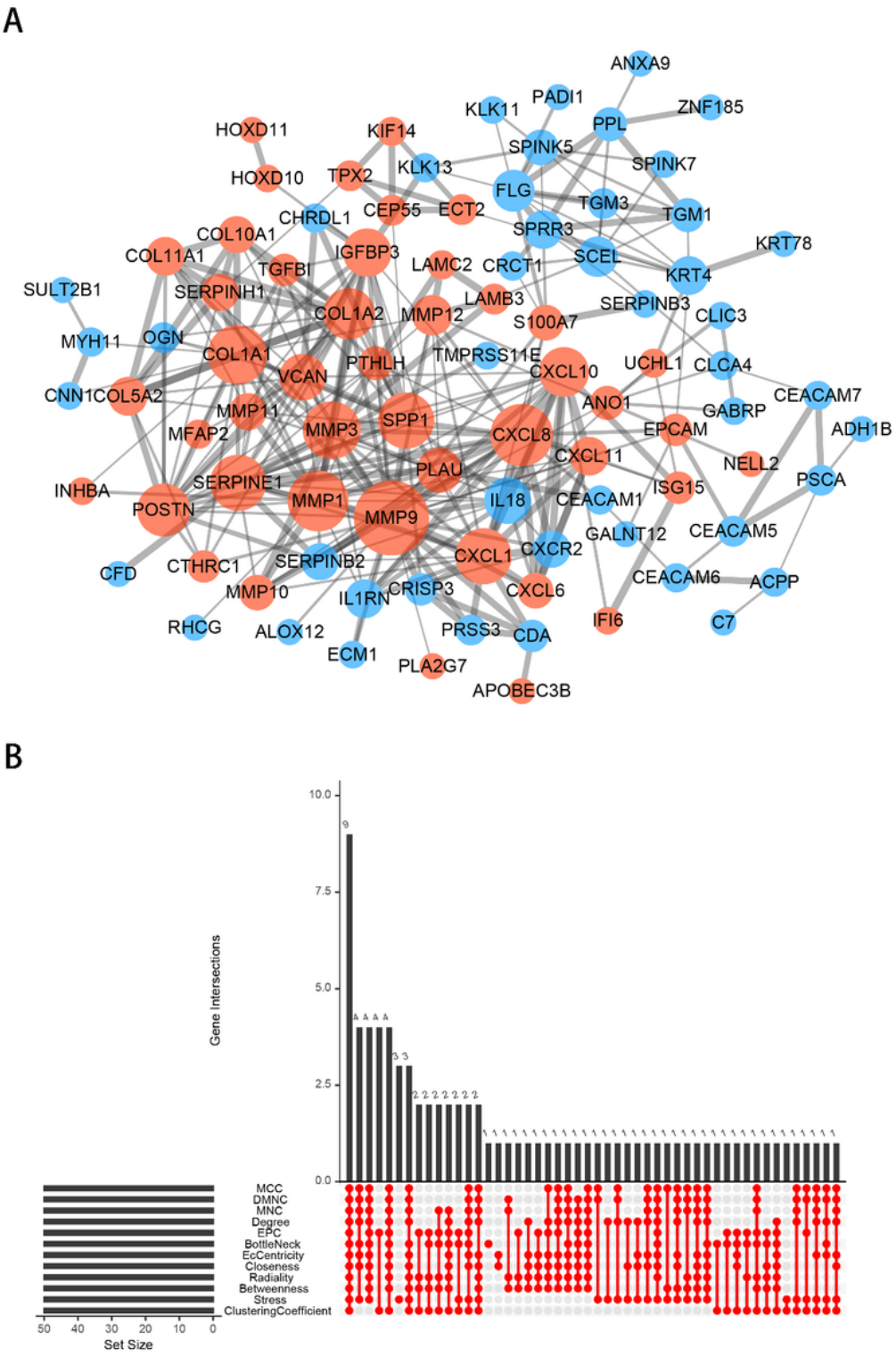

Figure 5

PPI network construction and hub genes identification. (A) The PPI network consisted of 91 nodes and 304 edges. Red nodes represent the upregulated genes, and blue nodes represent the downregulated DEGs. The spot size represents the connectivity degree. The edge thickness represents the combined score. (B) Nine hub genes were identified by the intersection of the top 50 genes from 12 algorithms in cytoHubba plugin.

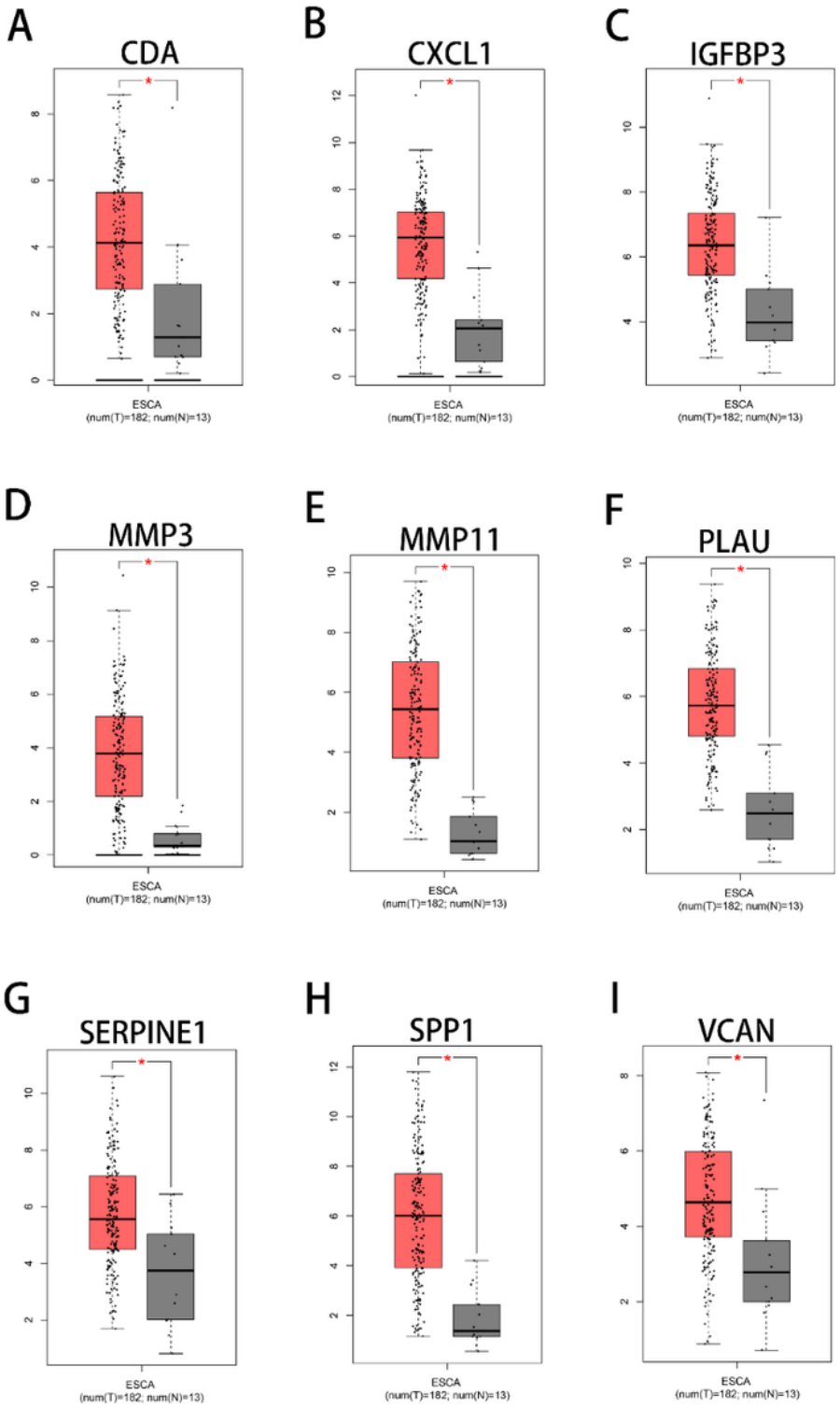

Figure 6

Validation of mRNA expression levels of hub genes between ESCA tissues and normal tissues in GEPIA database. (A-I) CDA, CXCL1, IGFBP3, MMP3, MMP11, PLAU, SERPINE1, SPP1 and VCAN were significantly upregulated in ESCC tissues compared with normal tissues. * P<0.01 was considered statistically significant. ESCA, Esophageal carcinoma.

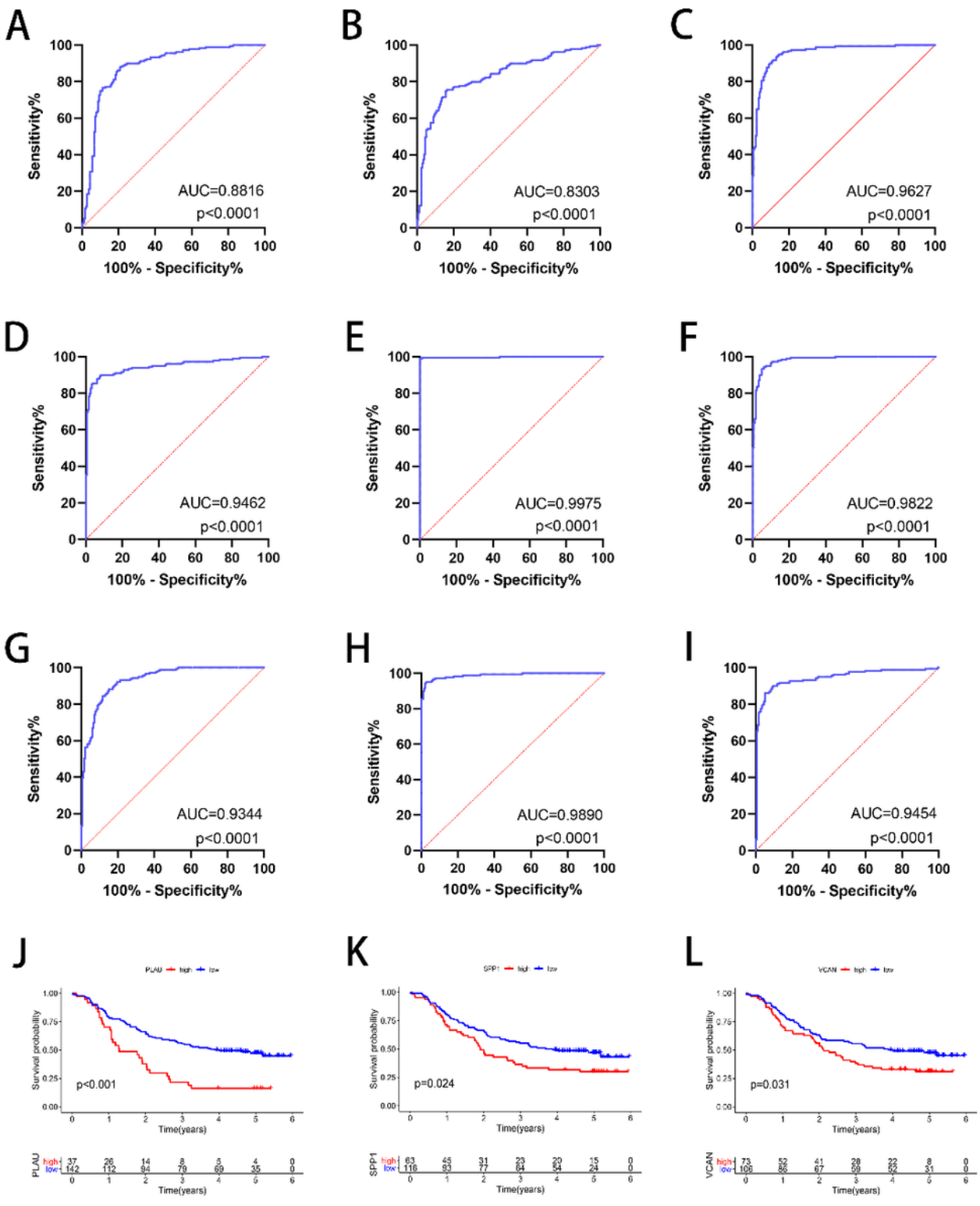

Figure 7

ROC curve analysis of hub genes diagnosis and Kaplan-Meier survival curve of hub genes. The AUC of CDA(A), CXCL1(B), IGFBP3(C), MMP3(D), MMP11(E), PLAU(F), SERPINE1(G), SPP1(H) and VCAN(I) were 0.8816, 0.8303, 0.9627, 0.9462, 0.9975, 0.9822, 0.9344, 0.9890 and 0.9454, respectively. High expression levels of PLAU (J), SPP1 (K) and VCAN (L) were associated with poor overall survival. ROC, receiver operating characteristic curve; AUC, area under the curve.

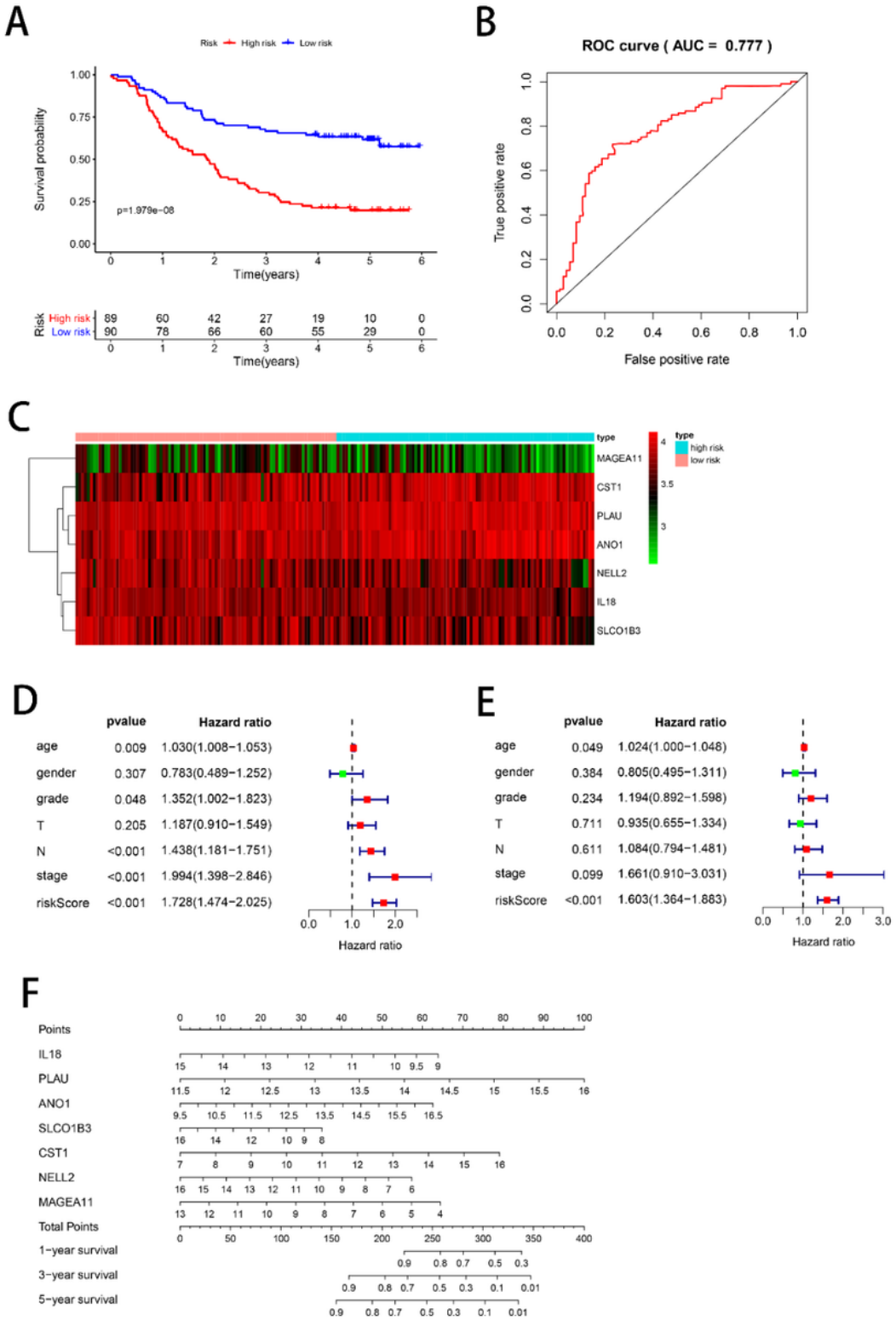

Figure 8

Prognostic risk score model analysis of 7 prognostic genes in ESCC patients. (A) Kaplan-Meier curves for the overall survival of patients in the high-risk group and low-risk group. (B) ROC curves for forecasting 5-year survival based on the risk score. (C) A risk heat map constructed from 7 genes from 179 ESCC patients. (D, E) Univariate and multivariate independent prognostic analysis forest map of the prognostic

genes model and ESCC clinicopathological characteristics. The red and green dots represented high and low-risk factors, respectively. (F) Nomogram for predicting 1-, 3- and 5-year survival of ESCC patients.

Supplementary Files

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

- SupplementaryZitongFeng.pdf