The clinical significance of collagen family gene expression for esophageal squamous cell carcinoma

Jieling Li Equal first author, [1] , **Xiao Wang** Equal first author, [2] , **Kai Zheng** [1] , **Ying Liu** [2] , **Junjun Li** [1] , **Shaoqi Wang** [3] , **Kaisheng Liu** [2] , **Nan Li** [2] , **Shouxia Xie** Corresp., [2] , **Shaoxiang Wang** Corresp. [1]

[1] School of Pharmaceutical Sciences, Shenzhen University Health Science Center, Shenzhen, China

[2] Department of Pharmacy, The Second Clinical Medical College (Shenzhen People's Hospital), Jinan University, Shenzhen, China

[3] Department of Oncology, Hubei Provincial Corps Hospital, Chinese People Armed Police Forces, Wuhan, China

Corresponding Authors: Shouxia Xie, Shaoxiang Wang
Email address: szshouxia@163.com, wsx@szu.edu.cn

Background: Esophageal squamous cell carcinoma (ESCC) is a subtype of esophageal cancer with high incidence and mortality . Due to the poor five-year survival rates of patients with ESCC, exploring novel diagnostic markers for early ESCC is emergent. Collagen, the abundant constituent of extracellular matrix, plays a critical role in tumor growth and epithelial-mesenchymal transition. However, the clinical significance of collagen genes in ESCC has been rarely studied. In this work, we systematically analyzed the gene expression of whole collagen family in ESCC, aiming to search for ideal biomarkers.

Methods: Clinical data and gene expression profiles of ESCC patients were collected from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases. Bioinformatics methods, including differential expression analysis, survival analysis, gene sets enrichment analysis (GSEA) and co-expression network analysis, were performed to investigate the correlation between the expression patterns of 44 collagen family genes and the development of ESCC.

Results: 22 genes of collagen family were identified as differentially expressed genes (DEGs) in both the two datasets. Among them, COL1A1, COL10A1 and COL11A1 were particularly up-regulated in ESCC tissues compared to normal controls, while COL4A4, COL6A5 and COL14A1 were notably down-regulated. Besides, patients with low COL6A5 expression or high COL18A1 expression showed poor survival. In addition, a 7-gene prediction model was established based on collagen gene expression to predict patient survival, which had better predictive accuracy than the tumor-node-metastasis (TNM) staging based model. Finally, GSEA results suggested that collagen genes might be tightly associated with PI3K/Akt/mTOR pathway, p53 pathway, apoptosis, cell cycle, etc.

Conclusion: Several collagen genes could be potential diagnostic and prognostic biomarkers for ESCC. Moreover, a novel 7-gene prediction model is probably useful for predicting survival outcomes of ESCC patients. These findings may facilitate early detection of ESCC and help improves prognosis of the patients.

The clinical significance of collagen family gene expression for esophageal squamous cell carcinoma

Jieling Li[1*], Xiao Wang[2*], Kai Zheng[1], Ying Liu[2], Junjun Li[1], Shaoqi Wang[3], Kaisheng Liu[2], Nan Li[2], Shouxia Xie[2], Shaoxiang Wang[1]

[1]School of Pharmaceutical Sciences, Shenzhen University Health Science Center, Shenzhen, China
[2]Department of Pharmacy, The Second Clinical Medical College (Shenzhen People's Hospital), Jinan University, Shenzhen, China
[3]Department of Oncology, Hubei Provincial Corps Hospital, Chinese People Armed Police Forces, Wuhan, China
[*]These authors contributed equally to this work.

Corresponding Author:
Shaoxiang Wang[1]
Email address: wsx@szu.edu.cn
Shouxia Xie[2]
Email address: szshouxia@163.com

Abstract

Background: Esophageal squamous cell carcinoma (ESCC) is a subtype of esophageal cancer with high incidence and mortality. Due to the poor five-year survival rates of patients with ESCC, exploring novel diagnostic markers for early ESCC is emergent. Collagen, the abundant constituent of extracellular matrix, plays a critical role in tumor growth and epithelial-mesenchymal transition. However, the clinical significance of collagen genes in ESCC has been rarely studied. In this work, we systematically analyzed the gene expression of whole collagen family in ESCC, aiming to search for ideal biomarkers.

Methods: Clinical data and gene expression profiles of ESCC patients were collected from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO) databases. Bioinformatics methods, including differential expression analysis, survival analysis, gene sets enrichment analysis (GSEA) and co-expression network analysis, were performed to investigate the correlation between the expression patterns of 44 collagen family genes and the development of ESCC.

Results: 22 genes of collagen family were identified as differentially expressed genes (DEGs) in both the two datasets. Among them, COL1A1, COL10A1 and COL11A1 were particularly up-regulated in ESCC tissues compared to normal controls, while COL4A4, COL6A5 and COL14A1 were notably down-regulated. Besides, patients with low COL6A5 expression or high COL18A1 expression showed poor survival. In addition, a 7-gene prediction model was established based on collagen gene expression to predict patient survival, which had better predictive accuracy than the tumor-node-metastasis (TNM) staging based model. Finally, GSEA results suggested that collagen genes might be tightly associated with PI3K/Akt/mTOR pathway, p53 pathway, apoptosis, cell cycle, etc.

Conclusion: Several collagen genes could be potential diagnostic and prognostic biomarkers for ESCC. Moreover, a novel 7-gene prediction model is probably useful for predicting survival outcomes of ESCC patients. These findings may facilitate early detection of ESCC and help improves prognosis of the patients.

Introduction

Esophageal cancer is the seventh most commonly diagnosed cancer and the sixth leading cause of cancer death (Bray et al. 2018). It is classified into two histological subtypes, esophageal adenocarcinoma (EAC) and esophageal squamous cell carcinoma (ESCC), the latter of which is the predominant type worldwide (Pennathur et al. 2013). Despite the effective treatments (e.g. surgery, chemotherapy and radiotherapy) for ESCC, the 5-year survival rates of patients with advanced ESCC are still less than 20% (Codipilly et al. 2018). However, the survival rates could be improved to over 80% if patients were diagnosed with an early stage (Lao-Sirieix & Fitzgerald 2012; Wang et al. 2004). Although a few tumor markers, carcinoembryonic antigen (CEA), carbohydrate antigen (CA) 19-9, and squamous cell carcinoma (SCC) antigen, have been used in the diagnosis of ESCC, they are not suitable for early detection due to the lack of sensitivity

(Kosugi et al. 2004). Thus, it is urgent to search for novel biomarkers to help early detection of ESCC and improve survival rates of the patients.

Collagen is the most abundant extracellular matrix protein that promotes cell growth and provides mechanical resilience of connective tissues (Sorushanova et al. 2018). The collagen family comprises 28 types with different α Chains encoded by more than 40 genes (Ricard-Blum 2011). It has been reported that the expression of collagen-encoding genes was significantly related to the prognosis of certain types of cancers (Giussani et al. 2018; Liu et al. 2018; Rong et al. 2018; Shen et al. 2016; Zhang et al. 2018c). In addition, a couple of collagen genes, such as COL11A1 and COL6A1, were expressed aberrantly in ESCC tissues and possibly affected the progression of ESCC (Fan et al. 2012; He et al. 2017; Zhang et al. 2018a). However, most of these works focused on specific collagen gene, and the potential roles of other members remain to be clarified.

Here we provided a systematic analysis on gene expression of the whole collagen family and its corresponding clinical significance in ESCC. Clinical data and gene expression profiles of ESCC patients were extracted from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO), two public databases with substantial information of cancers. Different bioinformatics methods, including differential expression analysis, survival analysis, pathway analysis and co-expression network analysis were used to analyze the data to sift important hits possibly involved in the initiation and development of ESCC. According to collagen family genes, we also established a prediction model with high performance to predict the prognosis of ESCC patients. Collectively, our works mainly explored the relation of collagen gene expression to ESCC and illuminated the potential mechanism.

Materials & Methods

Patient data

Basic data of ESCC patients were downloaded from the TCGA database (https://portal.gdc.cancer.gov/) and the GSE53625 dataset of the GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53625), 95 cases from TCGA and 179 cases from GSE53625. Univariate and multivariate Cox regression analyses were carried out to investigate the correlation between overall survival and clinicopathological characteristics of the patients by SPSS (v23.0). The relation between collagen family gene expression and clinicopathological characteristics of the patients was also examined via SPSS.

Differential expression analysis

Gene expression profiles of tumor and adjacent normal tissues in ESCC patients were also obtained from the two datasets. 81 of 95 patient cases in TCGA and all patient cases in GEO had RNA-sequence data. In total, 81 tumor samples with 11 normal controls from TCGA and 179 tumor samples with 179 normal controls from GEO (Li et al. 2014) were included in analysis (each sample was taken from a different patient). Differential expression analysis was conducted using the edgeR (Robinson et al. 2010) and the limma (Ritchie et al. 2015) packages respectively for TCGA and GEO data by R software (https://www.r-project.org/, v3.5.3). Gene expression levels were normalized by the calcNormFactors function in edgeR (Law et al. 2016) and by the normalizeBetweenArrays function in limma (Smyth & Speed 2003), to make the expression distributions of each sample are similar across the entire matrix. Then based on the exact test in edgeR which is analogous to Fisher's exact test (Robinson et al. 2010) and the Empirical Bayes statistical test in limma (Phipson et al. 2016), fold change (FC), *P* value and false discovery rate (FDR) (or adjusted *P* value) were figured out to show the expression difference between tumor and normal samples. Genes with $P < 0.05$ and FDR < 0.05 were considered as differentially expressed genes (DEGs). Accordingly, DEGs of collagen family were identified. Then heatmaps, boxplots and Venn diagram were drawn by R software.

Survival analysis

First, hazard ratio (HR) and *P* value of each DEG of collagen family were figured out based on gene expression and overall survival of patients by the univariate Cox regression model with the survival package through R software. The HR is an estimate of the ratio of the hazard rate in the treated versus the control group (Spruance et al. 2004), while in this study it is defined as the hazard in the high expression group divided by the hazard in the low expression group. HR > 1 and HR < 1 mean higher expression of the gene is associated with worse and better overall survival respectively. Survival curves were plotted according to the Kaplan-Meier method using the survival and the qvalue packages in R, to compare the overall survival of patients between high and low expression groups. $P < 0.05$ was considered statistically significant.

Prediction models

Prediction models were established to predict patient survival based on gene expression of 22 DEGs of collagen family and overall survival of patients by the multivariate Cox regress analysis with the survival package via R software. Several candidate genes were eventually selected out by the analysis to form the model, with a formula calculating the risk score of each patient. The general formula is given below:

$$\text{Risk score} = \sum_{i=1}^{n} Coef_i \times Exp_i \qquad (1)$$

where n, Coef, and Exp indicate the number of included genes, the coefficient of each gene, and gene expression level, respectively. Then receiver operating characteristic (ROC) curves were plotted based on the risk scores and overall survival of patients by the survivalROC package in R, with area under curve (AUC) values which represented the accuracy of predicting 3-year survival. Also, survival curves were obtained by dividing the patients into high- and low-risk groups according to the median risk score using the survival package.

Pathway analysis

Potential mechanism of collagen family genes was explored by the gene sets enrichment analysis (GSEA), a method to determine whether members of a previously defined gene set are correlated with the phenotypic class distinction (Subramanian et al. 2005). GSEA was conducted using the gene expression profiles of patients' tumor samples via javaGSEA software (http://software.broadinstitute.org/gsea/downloads.jsp), and the patient samples were divided into high- and low-risk groups by the median risk score obtained from the prediction models. Oncogenic Signatures Gene Sets (v6.2), Hallmark Gene Sets (v6.2) and KEGG Gene Sets (v6.2) (http://software.broadinstitute.org/gsea/msigdb/collections.jsp) were respectively used as references. Based on these gene sets databases, the expression profiles were analyzed to find out if a set of genes were mostly up-regulated (or down-regulated) in the high-risk group (or low-risk group). Normalized enrichment score (NES) reflected the degree to which a gene set was overrepresented in the groups, and gene sets in the results with P < 0.05 and FDR < 0.25 were considered as significant ones (Subramanian et al. 2005).

Co-expression network analysis

Patients' tumor samples from TCGA were separated into high- and low-risk groups by the risk scores calculated by the 7-gene prediction model. Risk-score-based DEGs that were differentially expressed between the two groups were determined using the gene expression profiles of tumor samples by the same method as differential expression analysis. Then the relationships between collagen family genes and the risk-score-based DEGs as well as the representative enriched gene sets from GSEA were assessed by the Weighted Gene Co-Expression Network Analysis (WCGNA) with the WGCNA package through R software, which is a method to describe the correlation patterns among genes across different samples (Langfelder & Horvath 2008). Genes of each gene set were extracted from http://software.broadinstitute.org/gsea/msigdb/genesets.jsp. Finally the genes co-expressed with collagen family genes were obtained, and the networks of them were drawn via Cytoscape (http://www.cytoscape.org/, v3.7.1).

Results

Clinicopathological information of the ESCC patients

A total of 95 patient cases in TCGA and 179 cases in GEO were collected and analyzed by univariate and multivariate Cox regression analyses. As a result, poor overall survival was significantly correlated with sex, TNM stage and N stage in TCGA (*P* = 0.020, *P* = 0.015, and *P* = 0.012, respectively) (Table 1), and was notably associated with age, TNM stage and N stage in GEO (*P* = 0.021, *P* < 0.001, and *P* = 0.030, respectively) (Table 2). Besides, investigation into the correlation between collagen family gene expression and the clinicopathological characteristics revealed that the expression of several collagen genes was significantly related to advanced TNM stages or tumor grades. (Table 3 and Table 4).

Identification of DEGs of collagen family in ESCC tissues

Differential expression analysis showed that more than 2/3 of the 44 collagen family genes were up-regulated in tumor tissues in both TCGA and GEO (Tables S1 and S2). 22 members in TCGA and 35 members in GEO were identified as DEGs, and their expression patterns were shown by heatmaps (Figs. 1A and 1B). Then the Venn diagram demonstrated that there were 22 mutual DEGs between the two datasets (Fig. 1C), which meant the DEGs observed in TCGA were also

DEGs in GEO. Obviously from the heatmaps, COL1A1, COL10A1 and COL11A1 ranked in the top five among the up-regulated DEGs in both datasets (Figs. 1D-1I), further presented by boxplots. Likewise, COL4A4, COL6A5 and COL14A1 were the most down-regulated candidates (Figs. 1J-1O).

Survival analysis of collagen family genes in ESCC patients

HRs and *P* values of the 22 DEGs were calculated and shown by heatmaps (Figs. 2A and 2B). Among them, HRs of COL6A5 and COL18A1 were the lowest and highest respectively. Survival curves of the DEGs were plotted according to the Kaplan-Meier method. Consistently, COL6A5 and COL18A1 were the two genes most relevant to the overall survival of ESCC patients. Patients with lower COL6A5 expression exhibited poorer overall survival ($P = 0.008$ in TCGA, Fig. 2C; $P = 0.060$ in GEO, Fig. 2D). By contrast, patients with higher COL18A1 expression had worse overall survival ($P = 0.393$ in TCGA, Fig. 2E; $P = 0.009$ in GEO, Fig. 2F). These results suggested that COL6A5 and COL18A1 are tightly associated with the prognosis of ESCC.

DEGs-based prediction models to predict the prognosis of ESCC patients

ROC curves have been extensively used to evaluate the predictive effect of one or more genes. The AUC value represents predictive accuracy and usually makes sense when it exceeds 0.60 (Ludemann et al. 2006; Metz 1978; Obuchowski 2003). ROC curves of COL6A5 and COL18A1 indicated that good predictive performance could only be attained by COL6A5 in TCGA (AUC=0.679, Fig. S1A), while COL18A1 had no predictive ability (Figs. S1C and S1D), suggesting that a single gene is not suitable for survival prediction of ESCC patients. Therefore, we established multi-gene prediction models based on expression levels of the DEGs to assess the joint effect of selected collagen genes on patient survival. There were 7 genes in TCGA and 9 genes in GEO finally included to form the models respectively, and risk scores of the patients were calculated according to the below formulas:

$$\text{Risk score (TCGA)} = (1.528*\text{COL1A1}_{\text{Exp}}) + (0.265*\text{COL4A4}_{\text{Exp}}) + (-0.539*\text{COL6A5}_{\text{Exp}}) + (-0.638*\text{COL11A1}_{\text{Exp}}) + (-1.193*\text{COL12A1}_{\text{Exp}}) + (-0.244*\text{COL19A1}_{\text{Exp}}) + (0.417*\text{COL24A1}_{\text{Exp}}). \qquad (2)$$

$$\text{Risk score (GEO)} = (7.700*\text{COL1A1}_{\text{Exp}}) + (-8.800*\text{COL1A2}_{\text{Exp}}) + (-5.800*\text{COL3A1}_{\text{Exp}}) + (6.320*\text{COL5A1}_{\text{Exp}}) + (-0.708*\text{COL6A5}_{\text{Exp}}) + (-0.790*\text{COL11A1}_{\text{Exp}}) + (1.990*\text{COL14A1}_{\text{Exp}}) + (1.300*\text{COL22A1}_{\text{Exp}}) + (2.400*\text{COL24A1}_{\text{Exp}}). \qquad (3)$$

Notably, AUCs on the ROC curves of the DEGs-based models in TCGA and GEO reached 0.86 and 0.68 respectively (Figs. 3A and 3C), which were higher than those of the prediction models based on TNM staging in the two datasets with AUCs of 0.625 and 0.646 respectively (Figs. 3E and 3G). The TNM staging system is a generally recognized standard for classifying the spreading extent of cancer (D'Journo 2018) and is commonly used to predict prognosis of cancer in clinical application. The prediction models respectively based on T-stage and N-stage were also examined but the AUCs were all less than 0.6 (Fig. S2). Furthermore, survival curves showed that patients with high risk were significantly correlated with poor survival (Figs. 3B, 3D, 3F and 3H). The 7-gene model in TCGA with true positive rate of 86% was more accurate than that of the TNM staging-based model, whereas predictive accuracy of the 9-gene model in GEO exhibited no difference. Therefore, the model in TCGA was used for our further studies. Finally, a heatmap was plotted to show the expression patterns of the 7 genes in TCGA between high-risk and low-risk groups (Fig. 3I). The risk score distribution was exhibited in ascending order, and patients were divided into high- and low-risk groups by the median point (Fig. 3J). Overall, it can be seen that patients with high risk score had higher mortality rates and shorter survival time than those with low risk score (Fig. 3K). Taken together, above results indicated that the 7-gene model could be more accurate to predict patient survival.

Pathway analysis of collagen family genes

GSEA results showed that most of the gene sets were up-regulated in the high-risk group, and the top twenty enriched gene sets were given in Tables S3-S8. The gene sets that were closely associated with tumorigenesis were shown in Fig. 4. For instance, gene sets of PDGF, RB/P107, AKT/MTOR and p53 were significantly up-regulated according to Oncogenic Signatures Gene Sets (Figs. 4A-4F). Based on Hallmark Gene Sets, the enriched gene sets included p53 pathway, oxidative phosphorylation, apoptosis, mitotic spindle, G2/M checkpoint and notch signaling (Figs. 4G-4L). Using KEGG Gene Sets as reference, the high-risk group was tightly correlated with oxidative phosphorylation, renal cell carcinoma, bladder cancer, small cell lung cancer, adherens junction and cell cycle (Figs. 4M-4R).

Co-expression network analysis

WCGNA was performed to find out the genes that were co-expressed with collagen family genes in ESCC tissues. Risk-score-based DEGs that were differentially expressed between high- and low-risk groups were determined and presented by the volcano plot (Fig. S3). The co-expression network of collagen genes and the risk-score-based DEGs were given in several modules (Fig. 5). Collagen family genes were displayed as red nodes, and the genes included in the 7-gene prediction model in TCGA were marked as bigger red nodes. The blue nodes represented the co-expressed genes. Another network was drawn to show the association between collagen family genes and seven representative enriched gene sets (PDGF, RB/p107, PI3K/Akt/mTOR pathway, p53 pathway, oxidative phosphorylation, apoptosis and cell cycle) from the GSEA results (Fig. S4). The red nodes were the collagen family genes with close connections to those gene sets. A big blue circle represented a gene set and the blue nodes were genes included in each set. Genes closer to the center were more tightly associated with the collagen genes.

Discussion

Although extensive research efforts have been focused on this field in past decades, efficient detection methods for early ESCC and accurate prediction against complicated ESCC patients still remain an open issue. Recently, studies have found that the expression of certain genes, such as MCT4, ZNF750, Gli1, etc. was highly related to the occurrence and development of ESCC, and they might be applied as ideal biomarkers for ESCC (Cheng et al. 2018; Li et al. 2018; Nambara et al. 2017; Yang et al. 2017; Zhang et al. 2018b). In addition, the aberrant expression of a few collagen family genes has also been reported to be significantly associated with the prognosis of ESCC patients. However, most works only focused on single or limited genes, and the predictive ability was barely satisfactory. Herein, we provided a more systematic analysis of the whole collagen family gene expression to evaluate the potential roles and clinical significance of collagen genes in ESCC.

We found that most of the collagen genes were up-regulated in ESCC tissues when compared to normal controls, half of which were identified as DEGs (Figs. 1A and 1B). Among them, the expression of COL1A1, COL10A1 and COL11A1 was particularly higher, and that of COL4A4, COL6A5 and COL14A1 was especially lower in tumor tissues, indicating their possible roles as diagnostic markers for ESCC. Consistently, several studies have shown that COL1A1, COL10A1 and COL11A1 were notably overexpressed in ESCC compared to normal tissues (Fang et al. 2019; He et al. 2017; Karagoz et al. 2016; Senthebane et al. 2018; Zhang et al. 2018a). Also, COL4A4 was also found to be down-regulated in esophageal tumor tissues (Chattopadhyay et al. 2009). Additionally, among the DEGs, COL7A1 was observed to be up-regulated in ESCC tissues (Kita et al. 2009). In our works, COL6A5, COL14A1 and some other collagen genes were reported to be significantly up- or down-regulated in ESCC tissues for the first time.

In the survival analysis, COL6A5 and COL18A1 were validated to be significantly related to overall survival of ESCC patients. Previous studies demonstrated that the COL6A5 expression was significantly associated with depressed behavior and atopic dermatitis (Soderhall et al. 2007; Zhan et al. 2017), but no articles manifested its correlation with cancer. In addition, COL18A1 has been proved to be a promising biomarker for ovarian cancer and was possibly involved in the progression of bladder cancer (Fang et al. 2013; Peters et al. 2005). In this study, ESCC patients with low COL6A5 expression or high COL18A1 expression showed poor overall survival (Figs. 2C-2F), implying the expression of COL6A5 or COL18A1 as a potential indicator for the prognosis of ESCC patients. Moreover, the variations that affect the expression of COL6A5 and COL18A1 possibly have effects on the progression of ESCC. Activating COL6A5 or inhibiting COL18A1 might improve the therapeutic efficiency and the life-span of ESCC patients.

Because the expression of one gene is usually influenced by various factors, ideal effect may not be attained by using a single gene as a predictor. Indeed, COL6A5 achieved an AUC value over 0.60 only in TCGA (Fig. S1), making the requirement of another more powerful prediction method. Based on the selected collagen DEGs (7 genes in TCGA and 9 genes in GEO, both including COL6A5), we established two new prediction models. Importantly, such DEGs-based models exhibited better predictive ability than conventional prognostic models according to TNM staging. The 7-gene model in TCGA had especially higher predictive accuracy of 86%. One possible reason was that the RNA sequencing technology applied to TCGA was more accurate than the gene chip technology used in GEO. In summary, this 7-gene prediction model is greatly promising to predict the prognosis of ESCC patients and help determine next therapeutic regimens.

Furthermore, GSEA was used to identify significantly enriched gene sets and potentially relevant pathways (Fig. 4). The results showed that based on Oncogenic Signatures Gene Sets, gene sets of PDGF, RB/P107 and AKT/MTOR were significantly enriched in the high-risk group.

It has been reported that PDGF receptor-beta increased the expression of COL1A2 through Akt/mTORC1 signaling pathway (Das et al. 2017). Besides, the pRB-related p107 protein was found to be required for quantitatively normal COL2A1 expression (Rossi et al. 2002). According to Oncogenic Signatures Gene Sets and Hallmark Gene Sets, the high-risk group was significantly related to p53 and p53 pathway, which suggested that collagen genes might be highly associated with the p53 or its related pathway in ESCC. Earlier studies proved that enhanced expression of ectopic p53 in dermal fibroblasts inhibited basal and TGF-beta-stimulated collagen gene expression, and the absence of cellular p53 was correlated with increased transcriptional activity of the Type I collagen gene (COL1A2) and collagen synthesis (Ghosh et al. 2004). Moreover, the type IV collagen expression was inversely related to p53 in malignant tumors (Bar et al. 2004). Oxidative phosphorylation related genes were found to be up-regulated in the high-risk group by both Hallmark Gene Sets and KEGG Gene Sets. Indeed, some reports demonstrated that oxidative phosphorylation signature occurred when collagen density was decreased, and the change of collagen density microenvironment regulated the metabolism of cancer cells (Mah et al. 2018; Morris et al. 2016). As for apoptosis, recent studies have shown that Type IV collagen could stimulate cancer cell proliferation, migration, and inhibit apoptosis (Ohlund et al. 2013), and the negative regulation of COL10A1 induced apoptosis (Guo et al. 2018). Additionally, the gene sets of mitotic spindle, G2/M checkpoint and cell cycle were enriched in the high-risk group as well, implying that collagen might regulate the cell cycle of ESCC cells. Furthermore, it was indicated that the high-risk group was markedly associated with renal cell carcinoma, bladder cancer and small cell lung cancer. These results were consistent with previous studies that collagen gene expression was correlated with the poor prognosis of those cancers (Koskimaki et al. 2010; Wan et al. 2015; Xu et al. 2017; Zeng et al. 2018).

As shown by the co-expression network (Fig. 5), a few collagen family genes such as COL1A1, COL11A1, COL6A6, and COL19A1, were co-expressed with NETO1, NEUROD2, and NRG3, which are the genes involved in neural functions. These findings could be verified by earlier articles to some extent (McCarthy & Hay 1991; Perris et al. 1993a; Perris et al. 1993b). COL11A1 was also observed to be co-expressed with tumor suppressor candidate 7 (TUSC7), further validating the possible role of COL11A1 in the occurrence of ESCC. Beyond that, some potassium channel related genes (KCNA2, KCNE1B, KCNH1, KCNJ4, and KCNK4) were co-expressed with collagen genes in a way, revealing that collagen genes might be correlated with the regulation of potassium channels in ESCC. As for the two potential prognostic biomarkers, COL18A1 only showed close relations with collagen family members, while COL6A5 was associated with two other genes in this network, ROBO2 and MIR548A3. ROBO2 has been identified as a candidate tumor suppressor (Trifonov et al. 2013), and the alteration of its expression might play a role in malignant tumors of digestive tract including gastric and colorectal cancers (Je et al. 2013).

Apart from what is aforementioned, there are still some limitations of this research. For instance, the prediction model was comprised of several genes, making it difficult to conduct cellular experiments by targeting a single gene to confirm its predictive effect. Aside from it, the characteristics of patient samples, as well as the methodology utilized in TCGA, were somewhat different from that in GEO, which may explain the different results coming from the two datasets.

For example, TCGA uses the RNA sequence technology while GEO applies the gene chip technology to detect gene expression of patient tissues. Besides, TCGA mainly collected data from white people, whereas the majority of patients in GEO (GSE53625) were Asian. Therefore, there was no a single gene that exhibited significant *P* values in both datasets in the survival analysis, and the selected genes driving the prediction model in one dataset were not completely identical to those in another dataset. Further validation of these outcomes requires more clinical information and biological experiments in the future.

Conclusions

In summary, this study identified 22 collagen family genes that were significantly expressed higher or lower in ESCC compared to normal tissues. Among them, COL1A1, COL10A1, COL11A1, COL4A4, COL6A5 and COL14A1 were the most distinct ones and possessed the potential in ESCC diagnosis. Besides, COL6A5 and COL18A1 showed strong correlations with overall survival of ESCC patients and might be robust prognostic biomarkers for ESCC. Furthermore, we established a 7-gene prediction model with high performance to predict the prognosis of ESCC patients. In terms of the underlying mechanism, collagen genes might be associated with PI3K/Akt/mTOR pathway, p53 pathway, oxidative phosphorylation, apoptosis and cell cycle during the progression of ESCC. Our works may further benefit the diagnosis, prognosis and treatments for ESCC patients.

References

Bar JK, Grelewski P, Popiela A, Noga L, and Rabczynski J. 2004. Type IV collagen and CD44v6 expression in benign, malignant primary and metastatic ovarian tumors: correlation with Ki-67 and p53 immunoreactivity. *Gynecol Oncol* 95:23-31. 10.1016/j.ygyno.2004.06.046

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, and Jemal A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 10.3322/caac.21492

Chattopadhyay I, Phukan R, Singh A, Vasudevan M, Purkayastha J, Hewitt S, Kataki A, Mahanta J, Kapur S, and Saxena S. 2009. Molecular profiling to identify molecular mechanism in esophageal cancer with familial clustering. *Oncol Rep* 21:1135-1146.

Cheng B, Chen X, Li Y, Huang X, and Yu J. 2018. Prognostic value of monocarboxylate transporter 4 in patients with esophageal squamous cell carcinoma. *Oncol Rep* 40:2906-2915. 10.3892/or.2018.6706

Codipilly DC, Qin Y, Dawsey SM, Kisiel J, Topazian M, Ahlquist D, and Iyer PG. 2018. Screening for esophageal squamous cell carcinoma: recent advances. *Gastrointest Endosc* 88:413-426. 10.1016/j.gie.2018.04.2352

D'Journo XB. 2018. Clinical implication of the innovations of the 8(th) edition of the TNM classification for esophageal and esophago-gastric cancer. *J Thorac Dis* 10:S2671-S2681. 10.21037/jtd.2018.03.182

Das F, Ghosh-Choudhury N, Venkatesan B, Kasinath BS, and Ghosh Choudhury G. 2017. PDGF receptor-beta uses Akt/mTORC1 signaling node to promote high glucose-induced renal proximal tubular cell collagen I (alpha2) expression. *Am J Physiol Renal Physiol* 313:F291-F307. 10.1152/ajprenal.00666.2016

Fan NJ, Gao CF, Wang CS, Zhao G, Lv JJ, Wang XL, Chu GH, Yin J, Li DH, Chen X, Yuan XT, and Meng NL. 2012. Identification of the up-regulation of TP-alpha, collagen alpha-1(VI) chain, and S100A9 in esophageal squamous cell carcinoma by a proteomic method. *J Proteomics* 75:3977-3986. 10.1016/j.jprot.2012.05.008

Fang S, Dai Y, Mei Y, Yang M, Hu L, Yang H, Guan X, and Li J. 2019. Clinical significance and biological role of cancer-derived Type I collagen in lung and esophageal cancers. *Thorac Cancer*. 10.1111/1759-7714.12947

Fang ZQ, Zang WD, Chen R, Ye BW, Wang XW, Yi SH, Chen W, He F, and Ye G. 2013. Gene expression profile and enrichment pathways in different stages of bladder cancer. *Genet Mol Res* 12:1479-1489. 10.4238/2013.May.6.1

Ghosh AK, Bhattacharyya S, and Varga J. 2004. The tumor suppressor p53 abrogates Smad-dependent collagen gene induction in mesenchymal cells. *J Biol Chem* 279:47455-47463. 10.1074/jbc.M403477200

Giussani M, Landoni E, Merlino G, Turdo F, Veneroni S, Paolini B, Cappelletti V, Miceli R, Orlandi R, Triulzi T, and Tagliabue E. 2018. Extracellular matrix proteins as diagnostic markers of breast carcinoma. *J Cell Physiol* 233:6280-6290. 10.1002/jcp.26513

Guo Q, Zheng M, Xu Y, Wang N, and Zhao W. 2018. miR-384 induces apoptosis and autophagy of non-small cell lung cancer(NSCLC) cells through the negative regulation of Collagen alpha-1(X) chain(COL10A1) gene. *Biosci Rep*. 10.1042/BSR20181523

He Y, Liu J, Zhao Z, and Zhao H. 2017. Bioinformatics analysis of gene expression profiles of esophageal squamous cell carcinoma. *Dis Esophagus* 30:1-8. 10.1093/dote/dow018

Je EM, Gwak M, Oh H, Choi MR, Choi YJ, Lee SH, and Yoo NJ. 2013. Frameshift mutations of axon guidance genes ROBO1 and ROBO2 in gastric and colorectal cancers with microsatellite instability. *Pathology* 45:645-650. 10.1097/PAT.0000000000000007

Karagoz K, Lehman HL, Stairs DB, Sinha R, and Arga KY. 2016. Proteomic and Metabolic Signatures of Esophageal Squamous Cell Carcinoma. *Curr Cancer Drug Targets*.

Kita Y, Mimori K, Tanaka F, Matsumoto T, Haraguchi N, Ishikawa K, Matsuzaki S, Fukuyoshi Y, Inoue H, Natsugoe S, Aikou T, and Mori M. 2009. Clinical significance of LAMB3 and COL7A1 mRNA in esophageal squamous cell carcinoma. *Eur J Surg Oncol* 35:52-58. 10.1016/j.ejso.2008.01.025

Koskimaki JE, Karagiannis ED, Tang BC, Hammers H, Watkins DN, Pili R, and Popel AS. 2010. Pentastatin-1, a collagen IV derived 20-mer peptide, suppresses tumor growth in a small cell lung cancer xenograft model. *BMC Cancer* 10:29. 10.1186/1471-2407-10-29

Kosugi S, Nishimaki T, Kanda T, Nakagawa S, Ohashi M, and Hatakeyama K. 2004. Clinical significance of serum carcinoembryonic antigen, carbohydrate antigen 19-9, and squamous cell carcinoma antigen levels in esophageal cancer patients. *World J Surg* 28:680-685. 10.1007/s00268-004-6865-y

Langfelder P, and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. 10.1186/1471-2105-9-559

Lao-Sirieix P, and Fitzgerald RC. 2012. Screening for oesophageal cancer. *Nat Rev Clin Oncol* 9:278-287. 10.1038/nrclinonc.2012.35

Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, and Ritchie ME. 2016. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Res* 5. 10.12688/f1000research.9005.3

Li J, Chen Z, Tian L, Zhou C, He MY, Gao Y, Wang S, Zhou F, Shi S, Feng X, Sun N, Liu Z, Skogerboe G, Dong J, Yao R, Zhao Y, Sun J, Zhang B, Yu Y, Shi X, Luo M, Shao K, Li N, Qiu B, Tan F, Chen R, and He J. 2014. LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut* 63:1700-1710. 10.1136/gutjnl-2013-305806

Li K, Liu Y, Xu S, and Wang J. 2018. PPM1D Functions as Oncogene and is Associated with Poor Prognosis in Esophageal Squamous Cell Carcinoma. *Pathol Oncol Res*. 10.1007/s12253-018-0518-1

Liu W, Li L, Ye H, Tao H, and He H. 2018. Role of COL6A3 in colorectal cancer. *Oncol Rep* 39:2527-2536. 10.3892/or.2018.6331

Ludemann L, Grieger W, Wurm R, Wust P, and Zimmer C. 2006. Glioma assessment using quantitative blood volume maps generated by T1-weighted dynamic contrast-enhanced magnetic resonance imaging: a receiver operating characteristic study. *Acta Radiol* 47:303-310.

Mah EJ, Lefebvre A, McGahey GE, Yee AF, and Digman MA. 2018. Collagen density modulates triple-negative breast cancer cell metabolism through adhesion-mediated contractility. *Sci Rep* 8:17094. 10.1038/s41598-018-35381-9

McCarthy RA, and Hay ED. 1991. Collagen I, laminin, and tenascin: ultrastructure and correlation with avian neural crest formation. *Int J Dev Biol* 35:437-452.

Metz CE. 1978. Basic principles of ROC analysis. *Semin Nucl Med* 8:283-298.

Morris BA, Burkel B, Ponik SM, Fan J, Condeelis JS, Aguirre-Ghiso JA, Castracane J, Denu JM, and Keely PJ. 2016. Collagen Matrix Density Drives the Metabolic Shift in Breast Cancer Cells. *EBioMedicine* 13:146-156. 10.1016/j.ebiom.2016.10.012

Nambara S, Masuda T, Tobo T, Kidogami S, Komatsu H, Sugimachi K, Saeki H, Oki E, Maehara Y, and Mimori K. 2017. Clinical significance of ZNF750 gene expression, a novel tumor suppressor gene, in esophageal squamous cell carcinoma. *Oncol Lett* 14:1795-1801. 10.3892/ol.2017.6341

Obuchowski NA. 2003. Receiver operating characteristic curves and their use in radiology. *Radiology* 229:3-8. 10.1148/radiol.2291010898

Ohlund D, Franklin O, Lundberg E, Lundin C, and Sund M. 2013. Type IV collagen stimulates pancreatic cancer cell proliferation, migration, and inhibits apoptosis through an autocrine loop. *BMC Cancer* 13:154. 10.1186/1471-2407-13-154

Pennathur A, Gibson MK, Jobe BA, and Luketich JD. 2013. Oesophageal carcinoma. *Lancet* 381:400-412. 10.1016/S0140-6736(12)60643-6

Perris R, Kuo HJ, Glanville RW, and Bronner-Fraser M. 1993a. Collagen type VI in neural crest development: distribution in situ and interaction with cells in vitro. *Dev Dyn* 198:135-149. 10.1002/aja.1001980207

Perris R, Kuo HJ, Glanville RW, Leibold S, and Bronner-Fraser M. 1993b. Neural crest cell interaction with type VI collagen is mediated by multiple cooperative binding sites within triple-helix and globular domains. *Exp Cell Res* 209:103-117. 10.1006/excr.1993.1290

Peters DG, Kudla DM, Deloia JA, Chu TJ, Fairfull L, Edwards RP, and Ferrell RE. 2005. Comparative gene expression analysis of ovarian carcinoma and normal ovarian epithelium by serial analysis of gene expression. *Cancer Epidemiol Biomarkers Prev* 14:1717-1723. 10.1158/1055-9965.EPI-04-0704

Phipson B, Lee S, Majewski IJ, Alexander WS, and Smyth GK. 2016. Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression. *Ann Appl Stat* 10:946-963. 10.1214/16-AOAS920

Ricard-Blum S. 2011. The collagen family. *Cold Spring Harb Perspect Biol* 3:a004978. 10.1101/cshperspect.a004978

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47. 10.1093/nar/gkv007

Robinson MD, McCarthy DJ, and Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140. 10.1093/bioinformatics/btp616

Rong L, Huang W, Tian S, Chi X, Zhao P, and Liu F. 2018. COL1A2 is a Novel Biomarker to Improve Clinical Prediction in Human Gastric Cancer: Integrating Bioinformatics and Meta-Analysis. *Pathol Oncol Res* 24:129-134. 10.1007/s12253-017-0223-5

Rossi F, MacLean HE, Yuan W, Francis RO, Semenova E, Lin CS, Kronenberg HM, and Cobrinik D. 2002. p107 and p130 Coordinately regulate proliferation, Cbfa1 expression, and hypertrophic differentiation during endochondral bone development. *Dev Biol* 247:271-285.

Senthebane DA, Jonker T, Rowe A, Thomford NE, Munro D, Dandara C, Wonkam A, Govender D, Calder B, Soares NC, Blackburn JM, Parker MI, and Dzobo K. 2018. The Role of Tumor Microenvironment in Chemoresistance: 3D Extracellular Matrices as Accomplices. *Int J Mol Sci* 19. 10.3390/ijms19102861

Shen L, Yang M, Lin Q, Zhang Z, Zhu B, and Miao C. 2016. COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. *Oncol Rep* 36:877-885. 10.3892/or.2016.4869

Smyth GK, and Speed T. 2003. Normalization of cDNA microarray data. *Methods* 31:265-273.

Soderhall C, Marenholz I, Kerscher T, Ruschendorf F, Esparza-Gordillo J, Worm M, Gruber C, Mayr G, Albrecht M, Rohde K, Schulz H, Wahn U, Hubner N, and Lee YA. 2007. Variants in a novel epidermal collagen gene (COL29A1) are associated with atopic dermatitis. *PLoS Biol* 5:e242. 10.1371/journal.pbio.0050242

Sorushanova A, Delgado LM, Wu Z, Shologu N, Kshirsagar A, Raghunath R, Mullen AM, Bayon Y, Pandit A, Raghunath M, and Zeugolis DI. 2018. The Collagen Suprafamily: From Biosynthesis to Advanced Biomaterial Development. *Adv Mater*:e1801651. 10.1002/adma.201801651

Spruance SL, Reid JE, Grace M, and Samore M. 2004. Hazard ratio in clinical trials. *Antimicrob Agents Chemother* 48:2787-2792. 10.1128/AAC.48.8.2787-2792.2004

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545-15550. 10.1073/pnas.0506580102

Trifonov V, Pasqualucci L, Dalla Favera R, and Rabadan R. 2013. MutComFocal: an integrative approach to identifying recurrent and focal genomic alterations in tumor samples. *BMC Syst Biol* 7:25. 10.1186/1752-0509-7-25

Wan F, Wang H, Shen Y, Zhang H, Shi G, Zhu Y, Dai B, and Ye D. 2015. Upregulation of COL6A1 is predictive of poor prognosis in clear cell renal cell carcinoma patients. *Oncotarget* 6:27378-27387. 10.18632/oncotarget.4860

Wang GQ, Jiao GG, Chang FB, Fang WH, Song JX, Lu N, Lin DM, Xie YQ, and Yang L. 2004. Long-term results of operation for 420 patients with early squamous cell esophageal carcinoma discovered by screening. *Ann Thorac Surg* 77:1740-1744. 10.1016/j.athoracsur.2003.10.098

Xu F, Chang K, Ma J, Qu Y, Xie H, Dai B, Gan H, Zhang H, Shi G, Zhu Y, Zhu Y, Shen Y, and Ye D. 2017. The Oncogenic Role of COL23A1 in Clear Cell Renal Cell Carcinoma. *Sci Rep* 7:9846. 10.1038/s41598-017-10134-2

Yang Z, Cui Y, Ni W, Kim S, and Xuan Y. 2017. Gli1, a potential regulator of esophageal cancer stem cell, is identified as an independent adverse prognostic factor in esophageal squamous cell carcinoma. *J Cancer Res Clin Oncol*

143:243-254. 10.1007/s00432-016-2273-6

Zeng XT, Liu XP, Liu TZ, and Wang XH. 2018. The clinical significance of COL5A2 in patients with bladder cancer: A retrospective analysis of bladder cancer gene expression data. *Medicine (Baltimore)* 97:e0091. 10.1097/MD.0000000000010091

Zhan H, Huang F, Yan F, Zhao Z, Zhang J, Cui T, Yang F, Hai G, Jia X, and Shi Y. 2017. Alterations in splenic function and gene expression in mice with depressive-like behavior induced by exposure to corticosterone. *Int J Mol Med* 39:327-336. 10.3892/ijmm.2017.2850

Zhang B, Zhang C, Yang X, Chen Y, Zhang H, Liu J, and Wu Q. 2018a. Cytoplasmic collagen XIalphaI as a prognostic biomarker in esophageal squamous cell carcinoma. *Cancer Biol Ther* 19:364-372. 10.1080/15384047.2018.1423915

Zhang Y, Xu Y, Li Z, Zhu Y, Wen S, Wang M, Lv H, Zhang F, and Tian Z. 2018b. Identification of the key transcription factors in esophageal squamous cell carcinoma. *J Thorac Dis* 10:148-161. 10.21037/jtd.2017.12.27

Zhang Z, Fang C, Wang Y, Zhang J, Yu J, Zhang Y, Wang X, and Zhong J. 2018c. COL1A1: A potential therapeutic target for colorectal cancer expressing wild-type or mutant KRAS. *Int J Oncol* 53:1869-1880. 10.3892/ijo.2018.4536

Table 1(on next page)

Univariate and multivariate analyses of clinicopathological characteristics for overall survival in ESCC patients from the TCGA dataset (N=95).

Characteristics with *P* < 0.3 in the univariate analysis were further screened in the multivariate analysis. HR, hazard ratio; CI, confidence interval; TNM stage, tumor-node-metastasis stage; T stage, stage of tumor invasion; N stage, stage of regional lymph node invasion.

Variables	n (%)	Univariate analysis		Multivariate analysis	
		HR (95% CI)	*P*	HR (95% CI)	*P*
Age					
<60	56 (58.9%)	1 (Reference)			
≥60	39 (41.1%)	1.296 (0.631-2.662)	0.461		
Sex					
Male	80 (84.2%)	1(Reference)		1 (Reference)	
Female	15 (15.8%)	0.175 (0.041-0.756)	0.020	0.206 (0.043-0.978)	0.047
TNM Stage					
I+II	63 (66.3%)	1 (Reference)		1 (Reference)	
III+IV	31 (32.6%)	2.443 (1.191-5.011)	0.015	0.921 (0.321-2.643)	0.879
Missing	1 (1.1%)				
T Stage					
T1+T2	40 (42.1%)	1 (Reference)			
T3+T4	54 (56.8%)	1.351 (0.649-2.811)	0.422		
Missing	1 (1.1%)				
Tumor Grade					
G1+G2	65 (68.4%)	1 (Reference)			
G3	21 (22.1%)	0.736 (0.277-1.950)	0.537		
Missing	9 (9.5%)				
N Stage					
N0+N1	84 (88.4%)	1 (Reference)		1 (Reference)	
N2+N3	9 (9.5%)	3.265 (1.302-8.189)	0.012	6.738 (1.493-30.399)	0.013
Missing	2 (2.1%)				
Tumor Location					
Upper+Middle	50 (52.6%)	1 (Reference)			
Lower	44 (46.3%)	0.958 (0.448-2.051)	0.913		
Missing	1 (1.1%)				
Alcohol Use					
No	25 (26.3%)	1 (Reference)		1 (Reference)	
Yes	68 (71.6%)	2.172 (0.751-6.276)	0.152	4.755 (1.054-21.457)	0.043
Missing	2 (2.1%)				
Tobacco use					
No	44 (46.3%)	1 (Reference)		1 (Reference)	
Yes	51 (53.7%)	1.965 (0.901-4.285)	0.089	1.095 (0.440-2.725)	0.845
Race					
Asian	45 (47.4%)	1 (Reference)		1 (Reference)	
White+Other	47 (49.5%)	1.570 (0.688-3.581)	0.284	2.021(0.782-5.223)	0.146
Missing	3 (3.2%)				

Table 2(on next page)

Univariate and multivariate analyses of clinicopathological characteristics for overall survival in ESCC patients from the GEO dataset (N=179).

Characteristics with *P* < 0.3 in the univariate analysis were further screened in the multivariate analysis. HR, hazard ratio; CI, confidence interval; TNM stage, tumor-node-metastasis stage; T stage, stage of tumor invasion; N stage, stage of regional lymph node invasion.

Variables	n (%)	Univariate analysis		Multivariate analysis	
		HR (95% CI)	*P*	HR (95% CI)	*P*
Age					
<60	91 (50.8%)	1 (Reference)		1 (Reference)	
≥60	88 (49.2%)	1.574 (1.072-2.311)	0.021	1.451 (0.980-2.147)	0.063
Sex					
Male	146 (81.6%)	1(Reference)			
Female	33 (18.4%)	1.277 (0.798-2.044)	0.307		
TNM Stage					
I+II	87 (48.6%)	1 (Reference)		1 (Reference)	
III+IV	92 (51.4%)	2.155 (1.448-3.207)	<0.001	2.066 (1.322-3.228)	0.001
T Stage					
T1+T2	39 (21.8%)	1 (Reference)			
T3+T4	140 (78.2%)	1.091 (0.687-1.732)	0.712		
Tumor Grade					
G1+G2	99 (55.3%)	1 (Reference)		1 (Reference)	
G3	80 (44.7%)	1.391 (0.951-2.037)	0.089	1.269 (0.860-1.873)	0.230
N Stage					
N0+N1	145 (81.0%)	1 (Reference)		1 (Reference)	
N2+N3	34 (19.0%)	1.644 (1.048-2.577)	0.030	1.062 (0.644-1.751)	0.814
Tumor Location					
Upper+Middle	117 (65.4%)	1 (Reference)			
Lower	62 (34.6%)	0.823 (0.546-1.242)	0.354		
Alcohol Use					
No	73 (40.8%)	1 (Reference)			
Yes	106 (59.2%)	0.864 (0.588-1.269)	0.456		
Tobacco Use					
No	65 (36.3%)	1 (Reference)		1 (Reference)	
Yes	114 (63.7%)	0.749 (0.508-1.105)	0.145	0.753 (0.505-1.122)	0.163
Pneumonia					
No	164 (91.6%)	1 (Reference)			
Yes	15 (8.4%)	1.425 (0.719-2.824)	0.310		

Table 3(on next page)

Correlation of collagen family gene expression and clinicopathological characteristics of ESCC patients from the TCGA dataset.

Superscripts of the correlation coefficients represent *P* values. * correlation with *P* < 0.05; ** correlation with *P* < 0.01.

Gene	Ag≥60	Sex	TNM Stage	N stage	Tumor Grade	Tumor Location
		(Female)	III/IV	(N1+N2)	(G3)	(Lower)
COL1A1		-0.222$^{*0.048}$				
COL1A2		-0.222$^{*0.048}$				
COL2A1						
COL3A1		-2.225$^{*0.045}$				
COL4A1						
COL4A2						
COL4A3						
COL4A4						
COL4A5						
COL4A6						
COL5A1						
COL5A2		-0.231$^{*0.039}$				
COL5A3		-0.229$^{*0.041}$				
COL6A1						
COL6A2						
COL6A3						
COL6A5						
COL6A6						
COL7A1					-0.226$^{*0.046}$	-0.226$^{*0.046}$
COL8A1						
COL8A2						
COL9A1						
COL9A2						
COL9A3		0.318$^{**0.004}$				
COL10A1						
COL11A1						
COL11A2						
COL12A1						-0.288$^{*0.010}$
COL13A1						
COL14A1						
COL15A1						
COL16A1			-0.280$^{*0.013}$		-0.280$^{*0.013}$	
COL17A1			-0.299$^{**0.008}$		-0.299$^{**0.008}$	
COL18A1						
COL19A1				0.367$^{**0.00}$		
COL20A1						
COL21A1		0.243$^{*0.030}$				
COL22A1						
COL23A1						

COL24A1	
COL25A1	
COL26A1	
COL27A1	$-0.245^{*0.02}$
COL28A1	

Table 4(on next page)

Correlation of collagen family gene expression and clinicopathological characteristics of ESCC patients in GEO.

Superscripts of the correlation coefficients represent *P* values. * correlation with *P* < 0.05; ** correlation with *P* < 0.01.

Gene	Age≥60	Sex	TNM Stage	N stage	Tumor Grade	Tumor Location
		(Female)	III+IV	(N1+N2)	(G3)	(Lower)
COL1A1						
COL1A2						
COL2A1						
COL3A1						
COL4A1						
COL4A2						
COL4A3					$0.149^{*0.046}$	$-0.162^{*0.030}$
COL4A4						$-0.168^{*0.024}$
COL4A5						
COL4A6						
COL5A1						
COL5A2						
COL5A3						$0.167^{*0.026}$
COL6A1						
COL6A2						
COL6A3						
COL6A5					$-0.173^{*0.020}$	
COL6A6						
COL7A1						
COL8A1		$0.188^{*0.012}$				
COL8A2						
COL9A1						
COL9A2				$-0.175^{*0.019}$		
COL9A3					$0.162^{*0.030}$	
COL10A1					$-0.151^{*0.044}$	
COL11A1						
COL11A2						
COL12A1						
COL13A1						
COL14A1						
COL15A1						
COL16A1						
COL17A1						
COL18A1						
COL19A1					$0.174^{*0.020}$	
COL20A1						
COL21A1			$-0.163^{*0.029}$			
COL22A1						
COL23A1						

COL24A1							
COL25A1							0.147$^{*0.049}$
COL26A1	0.174$^{*0.020}$						0.206$^{**0.006}$
COL27A1		-0.174$^{*0.020}$					
COL28A1							

Figure 1

Differential expression analysis of collagen family genes between ESCC and normal tissues.

(A) and (B) Heatmaps of the DEGs in TCGA and GEO in descending order of logFC. The red and blue colors represent high and low expression, respectively. * *P* < 0.05; ** *P* < 0.01; *** *P* < 0.001. (C) The Venn diagram showing the overlapped DEGs between the two datasets. (D-I) Boxplots of three representative up-regulated genes, COL1A1, COL10A1 and COL11A1 in TCGA and GEO. (J-O) Boxplots of three representative down-regulated genes, COL4A4, COL6A5 and COL14A1 in TCGA and GEO.

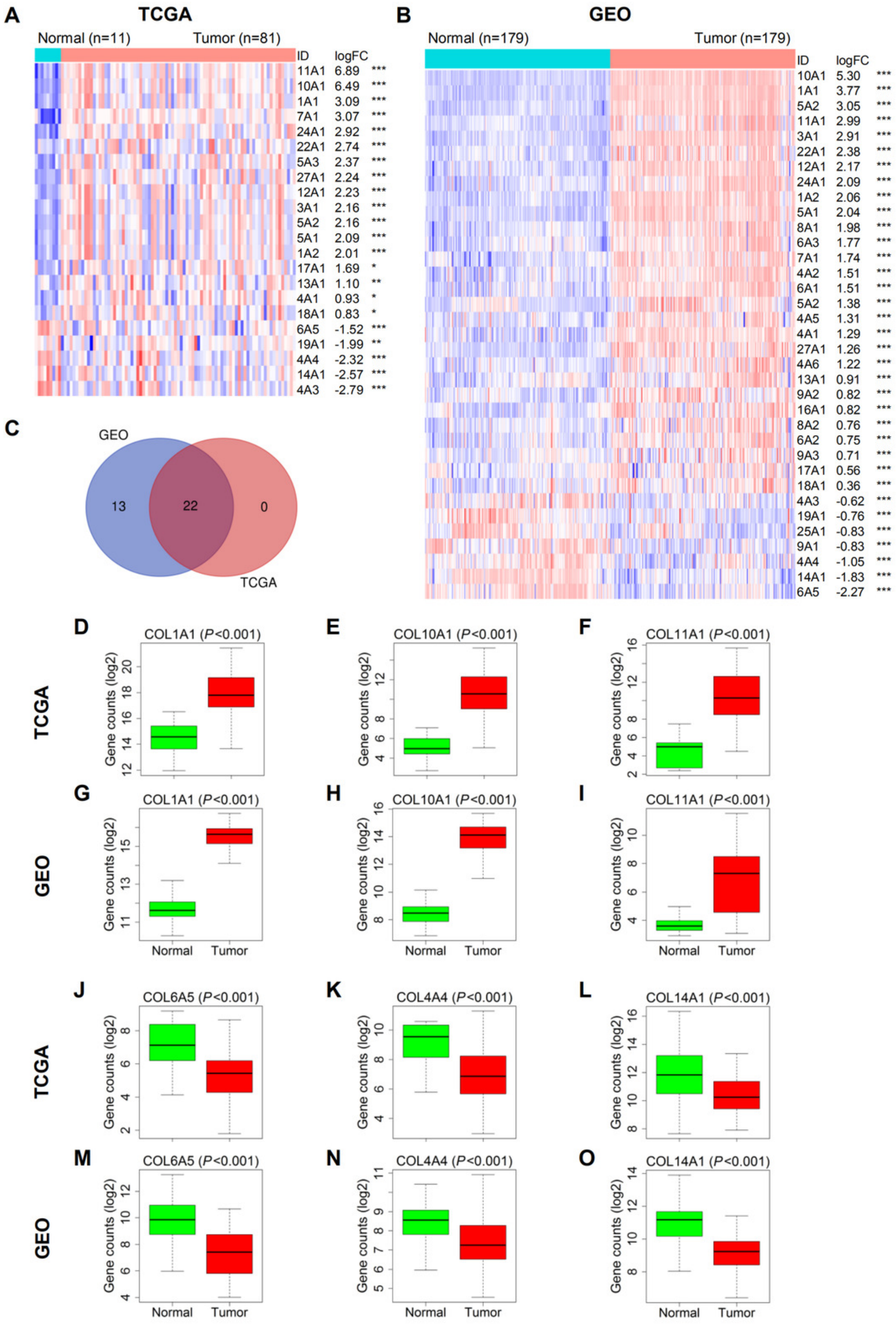

A
TCGA
Normal (n=11)
Tumor (n=81)
ID logFC
11A1 6.89 ***
10A1 6.49 ***
1A1 3.09 ***
7A1 3.07 ***
24A1 2.92 ***
22A1 2.74 ***
5A3 2.37 ***
27A1 2.24 ***
12A1 2.23 ***
3A1 2.16 ***
5A2 2.16 ***
5A1 2.09 ***
1A2 2.01 ***
17A1 1.69 *
13A1 1.10 **
4A1 0.93 *
18A1 0.83 *
6A5 -1.52 ***
19A1 -1.99 **
4A4 -2.32 ***
14A1 -2.57 ***
4A3 -2.79 ***
B
GEO
Normal (n=179)
Tumor (n=179)
ID logFC
10A1 5.30 ***
1A1 3.77 ***
5A2 3.05 ***
11A1 2.99 ***
3A1 2.91 ***
22A1 2.38 ***
12A1 2.17 ***
24A1 2.09 ***
1A2 2.06 ***
5A1 2.04 ***
8A1 1.98 ***
6A3 1.77 ***
7A1 1.74 ***
4A2 1.51 ***
6A1 1.51 ***
5A2 1.38 ***
4A5 1.31 ***
4A1 1.29 ***
27A1 1.26 ***
4A6 1.22 ***
13A1 0.91 ***
9A2 0.82 ***
16A1 0.82 ***
8A2 0.76 ***
6A2 0.75 ***
9A3 0.71 ***
17A1 0.56 ***
18A1 0.36 ***
4A3 -0.62 ***
19A1 -0.76 ***
25A1 -0.83 ***
9A1 -0.83 ***
4A4 -1.05 ***
14A1 -1.83 ***
6A5 -2.27 ***
C
GEO
13
22
0
TCGA
TCGA
GEO
TCGA
GEO
D
COL1A1 (P<0.001)
E
COL10A1 (P<0.001)
F
COL11A1 (P<0.001)
G
COL1A1 (P<0.001)
H
COL10A1 (P<0.001)
I
COL11A1 (P<0.001)
J
COL6A5 (P<0.001)
K
COL4A4 (P<0.001)
L
COL14A1 (P<0.001)
M
COL6A5 (P<0.001)
N
COL4A4 (P<0.001)
O
COL14A1 (P<0.001)
Gene counts (log2)
Normal
Tumor

Figure 2

Survival analysis of the DEGs of collagen family in ESCC patients.

(A) and (B) HRs and *P* values of the DEGs related to overall survival in ascending order of HR in TCGA and GEO. (C) and (D) Kaplan-Meier survival curves of COL6A5 in TCGA and GEO. (E) and (F) Kaplan-Meier survival curves of COL18A1 in TCGA and GEO.

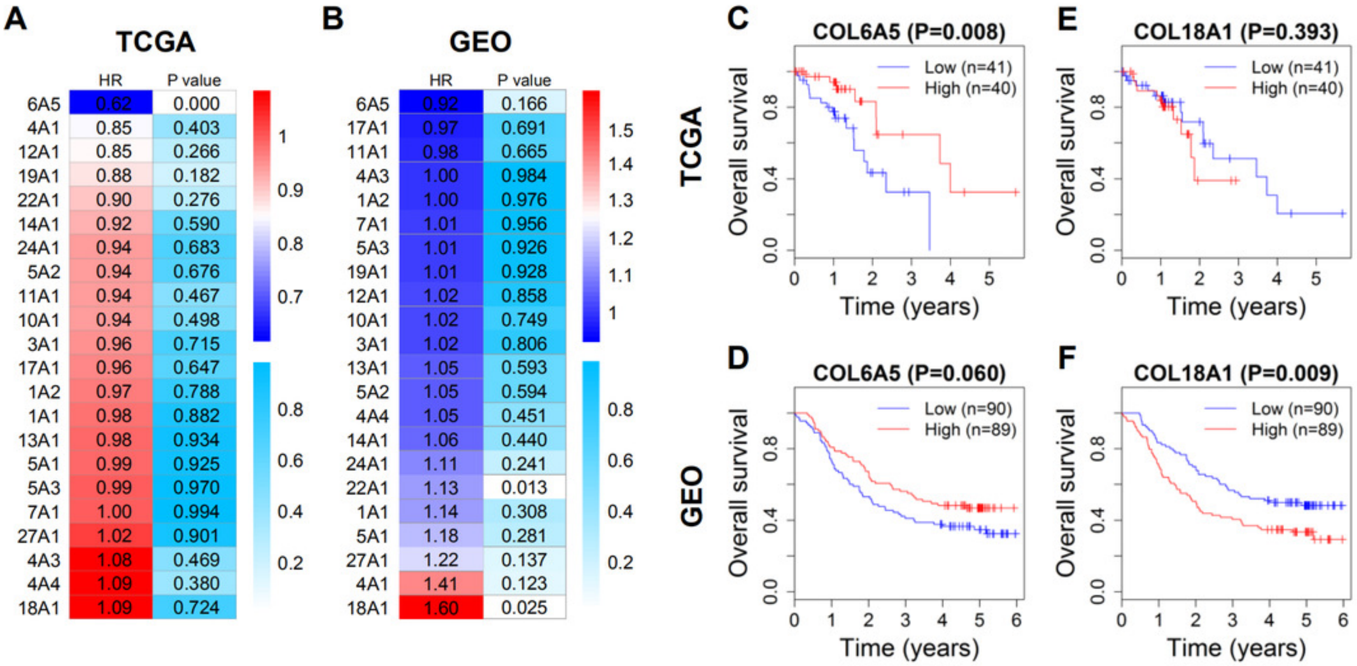


Figure 3

Prediction models to predict the survival of ESCC patients.

(A-D) ROC and survival curves of the models based on expression of 7 and 9 collagen DEGs respectively in TCGA and GEO. (E-H) ROC and survival curves of the models according to TNM staging in TCGA and GEO. (I) A heatmap showing the expression patterns of the 7 genes driving the prediction model in TCGA. (J) Risk score distribution of the patients in ascending order and divided into low-risk (green) and high-risk (red) in TCGA. (K) Survival time and status of the patients in order of increasing risk scores in TCGA. The red and green dots represent dead and alive, respectively.

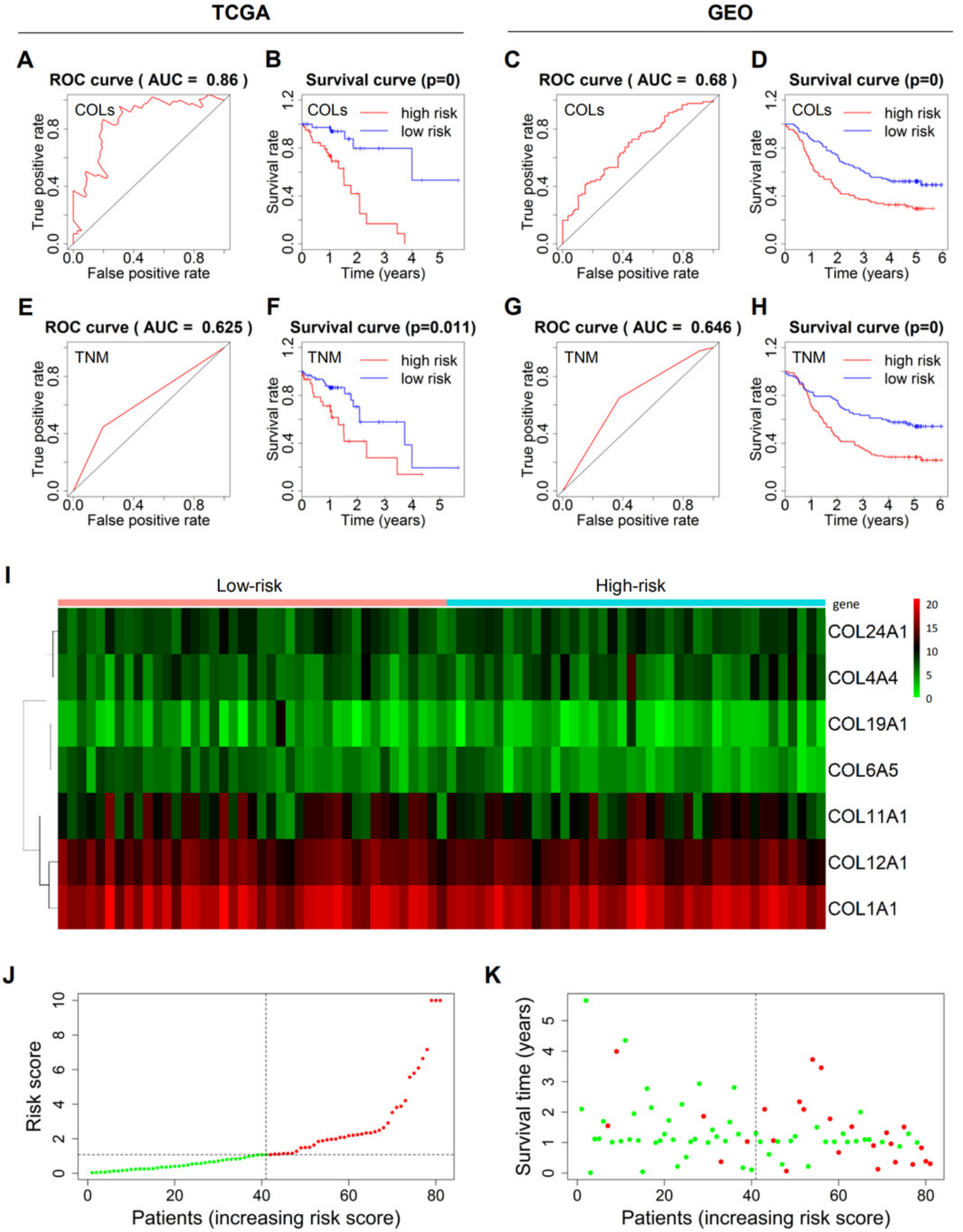

TCGA
GEO
A
ROC curve (AUC = 0.86)
COLs
True positive rate
False positive rate
B
Survival curve (p=0)
COLs
high risk
low risk
Survival rate
Time (years)
C
ROC curve (AUC = 0.68)
COLs
D
Survival curve (p=0)
COLs
high risk
low risk
E
ROC curve (AUC = 0.625)
TNM
F
Survival curve (p=0.011)
TNM
high risk
low risk
G
ROC curve (AUC = 0.646)
TNM
H
Survival curve (p=0)
TNM
high risk
low risk
I
Low-risk
High-risk
gene
COL24A1
COL4A4
COL19A1
COL6A5
COL11A1
COL12A1
COL1A1
20
15
10
5
0
J
Risk score
Patients (increasing risk score)
K
Survival time (years)
Patients (increasing risk score)

Figure 4

GSEA results based on patient risk scores calculated by the prediction models in TCGA and GEO

(A-F) Representative enriched gene sets according to Oncogenic Signatures Gene Sets. (G-L) Representative enriched gene sets according to Hallmark Gene Sets. (M-R) Representative enriched gene sets according to KEGG Gene Sets.

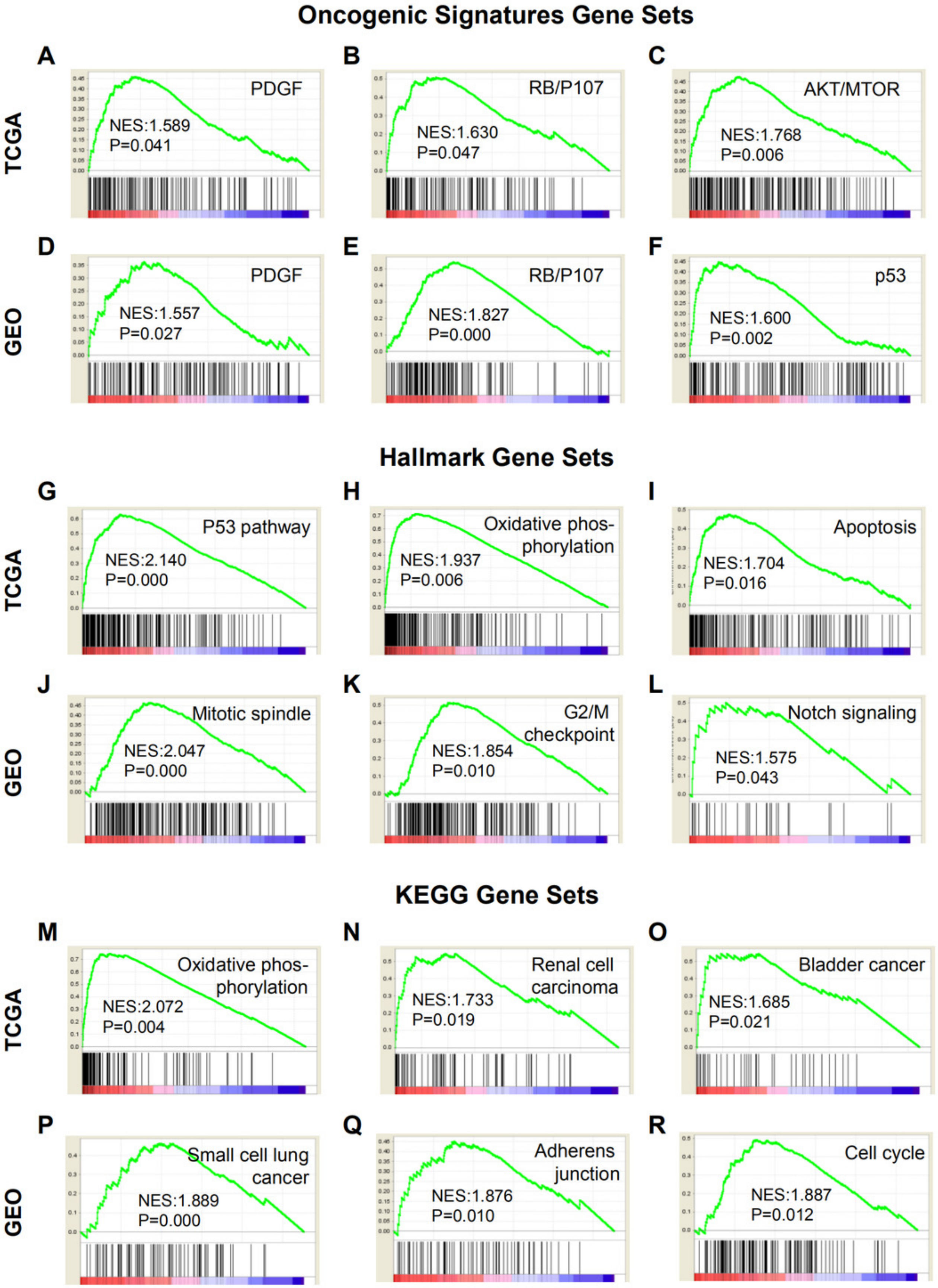
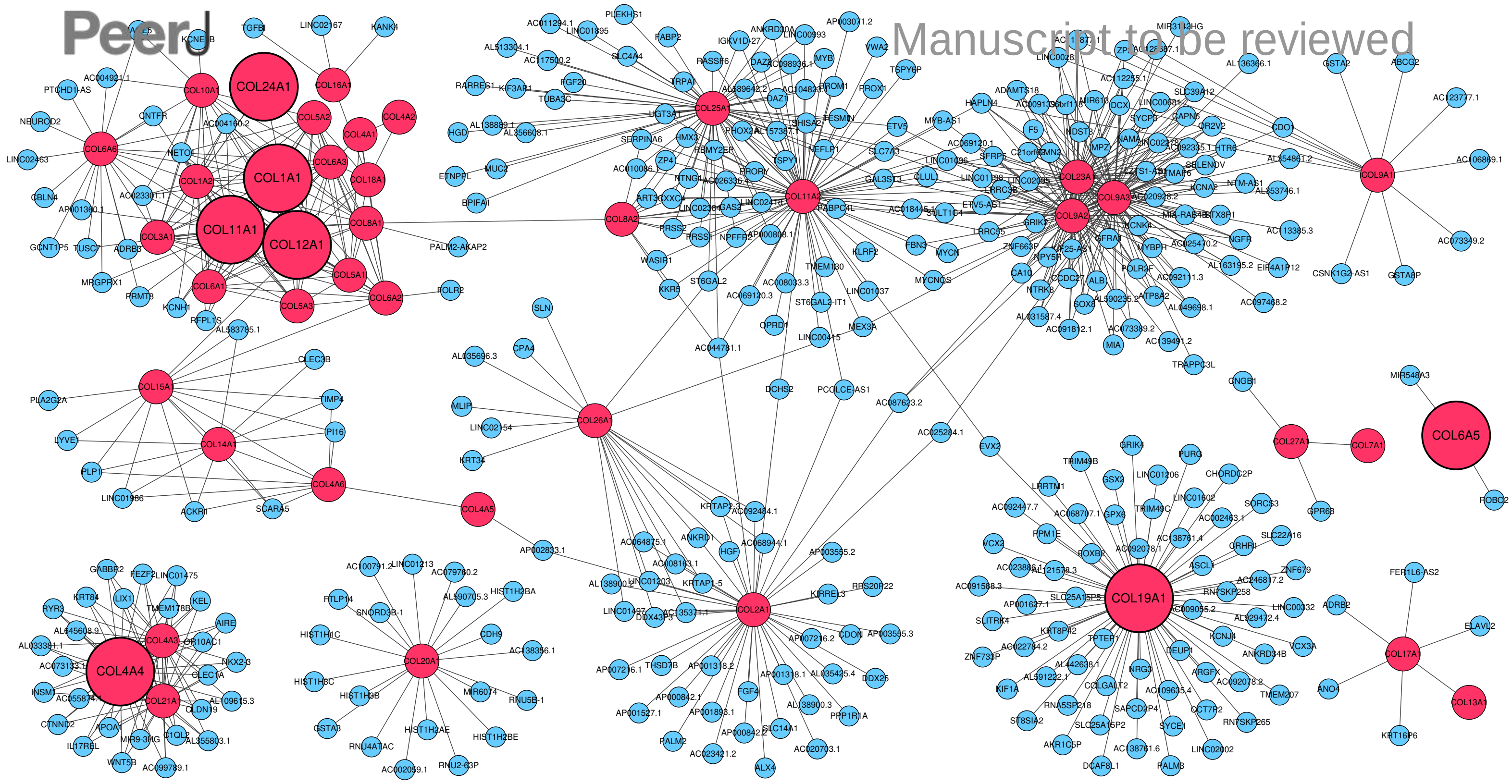

Oncogenic Signatures Gene Sets
A
B
C
TCGA
PDGF
NES:1.589
P=0.041
RB/P107
NES:1.630
P=0.047
AKT/MTOR
NES:1.768
P=0.006
D
E
F
GEO
PDGF
NES:1.557
P=0.027
RB/P107
NES:1.827
P=0.000
p53
NES:1.600
P=0.002
Hallmark Gene Sets
G
H
I
TCGA
P53 pathway
NES:2.140
P=0.000
Oxidative phos-phorylation
NES:1.937
P=0.006
Apoptosis
NES:1.704
P=0.016
J
K
L
GEO
Mitotic spindle
NES:2.047
P=0.000
G2/M checkpoint
NES:1.854
P=0.010
Notch signaling
NES:1.575
P=0.043
KEGG Gene Sets
M
N
O
TCGA
Oxidative phos-phorylation
NES:2.072
P=0.004
Renal cell carcinoma
NES:1.733
P=0.019
Bladder cancer
NES:1.685
P=0.021
P
Q
R
GEO
Small cell lung cancer
NES:1.889
P=0.000
Adherens junction
NES:1.876
P=0.010
Cell cycle
NES:1.887
P=0.012

Figure 5(on next page)

Co-expression network of collagen family genes.

Visualization of the co-expression between collagen family genes and the risk-scores-based DEGs. The red nodes are collagen family genes, and the bigger ones are the genes included in the 7-gene prediction model in TCGA. The blue nodes are the co-expressed genes.


COL24A1
COL1A1
COL11A1
COL12A1
COL10A1
COL16A1
COL5A2
COL4A1
COL4A2
COL6A6
COL1A2
COL6A3
COL18A1
COL3A1
COL8A1
COL6A1
COL5A1
COL5A3
COL6A2
COL15A1
COL14A1
COL4A6
COL4A5
COL4A4
COL4A3
COL21A1
COL20A1
COL25A1
COL11A2
COL8A2
COL26A1
COL2A1
COL23A1
COL9A3
COL9A2
COL9A1
COL27A1
COL7A1
COL6A5
COL19A1
COL17A1
COL13A1
TGFBI
LINC02167
KANK4
KCNE3B
NEUROD2
LINC02463
CBLN4
GCNT1P5
TUSC7
ADRB3
MRGPRX1
PRMT8
KCNH1
RFPL1S
CNTFR
NETO1
PALM2-AKAP2
FOLR2
CLEC3B
TIMP4
PI16
PLA2G2A
LYVE1
PLP1
LINC01986
ACKR1
SCARA5
GABBR2
FEZF2
LINC01475
RYR3
KRT84
LIX1
KEL
TMEM178B
AIRE
NKX2-3
CLEC1A
CLDN19
INSM1
CTNND2
APOA1
IL17REL
MIR9-3HG
WNT5B
C1QL2
HIST1H2BA
CDH9
HIST1H1C
HIST1H3C
HIST1H3B
GSTA3
RNU4ATAC
HIST1H2AE
HIST1H2BE
MIR6074
RNU5B-1
RNU2-63P
PLEKHS1
SLC4A4
FABP2
MYB
PROM1
PROX1
VWA2
TSPY6P
ETV5
MYB-AS1
SLC7A3
GAL3ST3
CLUL1
FBN3
MYCN
MYCNOS
TMEM130
KLRF2
MEX3A
LINC00415
DCHS2
PCOLCE-AS1
SLN
CPA4
MLIP
LINC02154
KRT34
EVX2
HGF
KRTAP1-5
ANKRD1
KIRREL3
PRSS20P22
CDON
DDX25
FGF4
PPP1R1A
SLC14A1
ALX4
PALM2
HAPLN4
ADAMTS18
F5
DCX
SYCP3
CAPN6
OR2V2
HTR6
SELENOV
KCNA2
NTM-AS1
NGFR
MIA
TRAPPC3L
GSTA2
ABCG2
CDO1
CSNK1G2-AS1
GSTA8P
EIF4A1P12
CNGB1
GPR68
MIR548A3
ROBO2
TRIM49B
GRIK4
PURG
CHORDC2P
SORCS3
SLC22A16
ZNF679
VCX3A
TMEM207
RNTSKP265
LINC02002
PALM3
DCAF8L1
AKR1C6P
ST8SIA2
KIF1A
ZNF733P
SLITRK4
VCX2
ASCL1
KCNJ4
NRG3
SYCE1
FER1L6-AS2
ADRB2
ELAVL2
ANO4
KRT16P6