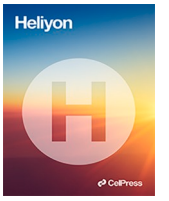
Heliyon 10 (2024) e36164

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Potential regulation and prognostic model of colorectal cancer with extracellular matrix genes

Xiaobao Yang [a,b,c], Jiale Gao [a,b,c], Tianzhen Zhang [a,b,c], Lu Yang [a,b,c], Chao Jing [a,b,c], Zhongtao Zhang [a,b,c,**], Dan Tian [a,b,c,d,*]

[a] Department of General Surgery, Beijing Friendship Hospital, Capital Medical University, Beijing, China
[b] State Key Lab of Digestive Health, Beijing Friendship Hospital, Capital Medical University, Beijing, China
[c] National Clinical Research Center for Digestive Diseases, Beijing, China
[d] Immunology Research Center for Oral and Systemic Health, Beijing Friendship Hospital, Capital Medical University, Beijing, China

ARTICLE INFO

Keywords:
Colorectal cancer
Extracellular matrix
Single-cell sequencing
Prognostic model

ABSTRACT

Background: The tumor microenvironment (TME) of colorectal cancer (CRC) mainly comprises immune cells, stromal cells, tumor cells, as well as the extracellular matrix (ECM), which holds a pivotal position. The ECM affects cancer progression, but its regulatory roles and predictive potential in CRC are not fully understood.
Methods: We analyzed transcriptomes from CRC tumors and paired normal tissues to study ECM features. Up-regulated ECM components were examined through functional enrichment analysis, and single-cell sequencing identified cell types producing collagen, regulators, and secreted factors. Transcription factor analysis and cell-cell interaction studies were conducted to identify potential regulators of ECM changes. Additionally, a prognostic model was developed using TCGA-CRC cohort data, focusing on up-regulated core ECM components.
Results: Bulk RNA-seq analysis revealed a unique ECM pattern in tumors, with ECM abundance and composition significantly related to patient survival. Up-regulated ECM components were linked to various cancer-related pathways. Fibroblasts and non-fibroblasts interactions were crucial in forming the TME. Key potential regulators identified included ZNF469, PRRX2, TWIST1, and AEBP1. A prognostic model based on five ECM genes (THBS3, LAMB3, ESM1, SPRX, COL9A3) demonstrated strong associations with immune suppression and tumor angiogenesis.
Conclusions: The ECM components were involved in various cell-cell interactions and correlated with tumor development and poor survival outcomes. The ECM prognostic model components could be potential targets for novel therapeutic interventions in colorectal cancer.

1. Introduction

The CRC ranks third place among the most common types of cancer and is the second highest leading cause globally [1]. In recent years, despite advances in surgical techniques, chemotherapy, targeted therapy, and immunotherapy approaches, the overall

* Corresponding author. Department of General Surgery, Beijing Friendship Hospital, Capital Medical University 95 Yong'an Road, Xicheng District, Beijing, 100050, China.
** Corresponding author. Department of General Surgery, Beijing Friendship Hospital, Capital Medical University, 95Yong'an Road, Xicheng District, Beijing, 100050, China.
E-mail addresses: zhangzht@ccmu.edu.cn (Z. Zhang), tiand@ccmu.edu.cn (D. Tian).

https://doi.org/10.1016/j.heliyon.2024.e36164
Received 19 January 2024; Received in revised form 10 August 2024; Accepted 11 August 2024
Available online 13 August 2024

prognosis remains poor [2]. Thus, it's crucial to find innovative biomarkers and therapeutic targets for CRC. As the key element of the evolutionary and ecological process in cancer development and cancer therapy, the tumor microenvironment (TME) has attracted more and more attention [3,4]. The TME consists of a variety of cell types, such as tumor cells, endothelial cells, fibroblast cells, as well as the transformed ECM [5,6]. In addition to providing support for tumor cells, the ECM also acts as a key regulator between cells and between cells and the matrix in the TME [7]. Moreover, the ECM undergoes dynamic remodeling during tumorigenesis, resulting in a cancer-supporting matrix that promotes tumor growth, invasion, angiogenesis, immune evasion, and drug resistance [7–10]. However, the molecular mechanisms underlying ECM remodeling and its impact on the TME in CRC are not fully understood. Moreover, the prognostic value of ECM components in CRC has yet to be established. Therefore, this research seeks to investigate the role of ECM in CRC tumorigenesis and prognosis by combining bulk sequencing and single-cell sequencing data.

We first performed an extensive analysis of the landscape of ECM-associated cells in CRC using publicly available datasets. We identified differentially expressed ECM components between normal and CRC tissues, and characterized their correlation with clinical traits and survival outcomes. We also conducted single-cell RNA sequencing (scRNA-seq) analysis to find the specific cell type that was responsible for the ECM changes in tumor tissues, reveal the cellular communication in the TME, and explore the potential molecular mechanisms of ECM action at the single-cell level. Furthermore, we constructed an ECM regulatory network based on transcription factor binding motifs and gene co-expression patterns. Finally, we constructed a prognostic model using the core ECM genes and validated its predictive performance successfully.

Collectively, this study provides a detailed overview of the ECM's role in CRC and reveals novel insights into how the ECM affects tumorigenesis, tumor progression, and prognosis. Our findings may also suggest new avenues for precision medicine approaches to improve the outcome of CRC patients.

2. Methods

2.1. Data acquisition and processing

We obtained the gene expression matrix and clinical characteristics of 618 CRC patients from The Cancer Genome Atlas (TCGA) database utilizing the TCGAbiolinks package (version 2.24.3) [11]. The data included 618 tumor samples and 51 paired normal samples. We used the raw read count and log2 (x+1) transformed transcripts per million (TPM) format data for further analysis. To ensure reliable survival analysis, individuals with an overall survival (OS) period of less than 30 days were excluded from our study. We also obtained six additional CRC datasets (GSE17536, GSE17537, GSE33113, GSE38832, and GSE161158) which used the Affymetrix GPL570 platform from the Gene Expression Omnibus (GEO). These datasets were used to validate our findings from the TCGA data.

We obtained the extracellular matrix (ECM) genes, also known as the Matrisome, from MatrisomeDB 2.0 (https://matrisomedb.org/) on March 6th, 2023. This database provides comprehensive information on the ECM components of various tissues and species [12]. We preprocessed the data and included 941 ECM genes in this study (see Supplementary Table 1 for details). The ECM components in MatrisomeDB 2.0 are categorized as core ECM and ECM-associated proteins. The core ECM consists of glycoproteins, collagens, and proteoglycans, which are highly insoluble, while the ECM-associated proteins comprise ECM-affiliated proteins, ECM regulators, and secreted factors that are known or anticipated to bind with ECM's structural elements [12].

2.2. Differentially expressed genes analysis

We utilized the DESeq2 package (version 1.36.0) [13] to detect differentially expressed genes (DEGs) in tumor tissues compared with normal samples. DESeq2 is a tool for analyzing count-based gene expression data, such as RNA sequencing. Genes exhibiting an absolute log2 fold change greater than 1.5 and an adjusted *p*-value below 0.05 were classified as DEGs. We then extracted the differentially expressed ECMs from the DEGs list.

2.3. Principal component analysis

We used principal component analysis (PCA) to discern extracellular matrix (ECM) patterns in tumor and normal samples of colorectal cancer (CRC) patients. We applied PCA to the log-transformed transcripts per million (TPM) data of 941 ECM genes, employing the FactoMineR R package (version 1.34) [14]. PCA is a technique used in RNA-Seq data to reduce the dimensionality and reveal hidden subtypes or heterogeneity within the data.

2.4. The immune landscape of CRC patients

We used four computational algorithms to evaluate the immune cell infiltration (ICI) within the tumor microenvironment of patients from the TCGA-CRC cohort. The algorithms we used were EPIC [15], CIBERSORT [16], ESTIMATE [17], and quanTIseq [18]. These algorithms estimate the proportions of various immune cell subsets by evaluating specific gene expression levels. These analyses were conducted using the IOBR R package (version 0.99.9) [19]. We also assessed the association between the gene signature score and the ICI level for each sample. We calculated the correlation coefficients employing the ggcorrplot R package (version 0.1.3) and visualized them as heatmaps with the ComplexHeatmap R package (version 2.13.2) [20].

2.5. Single sample gene set enrichment analysis

In addition to conducting immune cell infiltration analysis, we also utilized the single-sample gene set enrichment analysis (ssGSEA) method to evaluate the abundance of ECM components in individual samples based on the ECM (Extracellular Matrix) gene set. The ssGSEA is a widely recognized approach for scoring individual samples according to specific molecular characteristics or gene sets. This algorithm was implemented using the GSVA package (version 1.44.2) [21].

2.6. Survival analysis

To identify ECM genes that were associated with prognosis, we used the progression-free interval (PFI) as the main endpoint, which measures the time from diagnosis to disease progression or death (15). Then we applied the univariate Cox regression model, with the survival R package (version 3.5–3), to test the correlation between each ECM gene and PFI. We considered genes with a *p*-value below 0.05 and a hazard ratio (HR) greater than 1 as unfavorable genes, while genes with an HR less than 1 as favorable genes.

We also compared the prognosis of patients with varying stromal scores, which were calculated using the ESTIMATE algorithm. Individuals were categorized as high and low stromal score groups, with the cutoff points established using the survminer R package (version 0.4.9). Subsequently, the clinical outcomes between the two groups were compared with the log-rank test.

2.7. Gene functional enrichment analysis

We performed functional enrichment analysis on the dysregulated ECM genes with the clusterProfiler R package (version 4.4.4) [22]. This analysis revealed significantly enriched Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with these genes. GO terms provide insights into the biological processes, molecular functions, and cellular components of genes, while KEGG pathways show the interactions and relationships of genes in various biological systems. We adjusted the p-values by the Benjamini and Hochberg method to control for multiple comparisons, and a *p*-value lower than 0.05 was regarded as statistically significant.

2.8. Single-cell RNAseq data processing

We downloaded the raw count data of scRNA-seq from the GEO database, under accession numbers GSE132465 and GSE144735. We merged the count matrices from the two datasets using the Seurat R package (version 4.2.0) [23]. We removed the low-quality cells that had more than 20 % of mitochondrial genes or less than 200 detected genes. To normalize the count matrix, we applied the SCTransform to regress out the percentage of the mitochondrial genes, resulting in corrected counts for all features and Pearson residuals of 3000 highly variable features. We reduced the dimensionality of the dataset by conducting a PCA procedure using the RunPCA function in Seurat. To eliminate the batch effect, We corrected the PCA embeddings utilizing the RunHarmony (harmony, version 0.1.0). We further reduced the dimensionality of the corrected PCA embeddings by applying uniform manifold approximation and projection (UMAP). We constructed a shared nearest neighbor (SNN) graph utilizing FindNeighbors in Seurat, which measures the similarity between cells by analyzing their gene expression patterns. We found cell clusters with FindClusters in Seurat with a resolution of 0.6, which assigns labels to cells based on their SNN graph membership. We annotated the cell clusters with known markers, which are genes that are either highly expressed or unique to certain cell types.

2.9. Single-cell gene signature scoring for ECM genes

To assess the overall ECM gene expression patterns across various cell types, we calculated the gene signature score of ECM genes in our single-cell datasets with the UCell R package (version 2.0.1) [24]. The gene signature score is a measure of the expression level of a group of genes that are associated with a certain biological process or cell type. We used UCell because it is a fast and accurate tool for scoring gene signatures in single-cell data, which can reveal cellular heterogeneity and diversity in a tissue or organ.

2.10. Cell-cell communication analysis

We explored the interactions among cells and the mechanisms of the communicating molecules in our scRNA-seq datasets with the CellChat R package (version 1.5.0) [25]. CellChat is a tool designed to identify and predict the possible cell-cell interactions and signaling pathways in scRNA-seq data [25]. The communication patterns in CellChat were categorized into three main types: signaling through secreted factors, interactions between ECM and receptors, and direct cell-cell contact signaling driven by heterodimers. The ligand-receptor pair data for these signaling pathways was sourced from the KEGG database.

We applied CellChat to 10 cell groups in our data, such as CD4 T cells, CD8 T cells, regulatory T cells (Tregs), plasma cells, B cells, macrophages, monocytes, endothelial cells, epithelial cells, and fibroblasts. We used the aggregateNet function in CellChat to calculate the aggregated cell-cell communication network, which shows the overall signaling activity between different cell groups. We also visualized the signaling from each cell group using CellChat.

2.11. Transcription factor enrichment analysis

We performed an analysis of transcription factors to explore the potential upstream mediators of the ECM changes in CRC patients. These proteins attach to specific DNA sequences and modulate gene expression. We used the chEA web interface (https://maayanlab.cloud/chea3/) [26] to perform the transcription factor analysis on the differentially expressed core ECM genes. chEA is a tool that

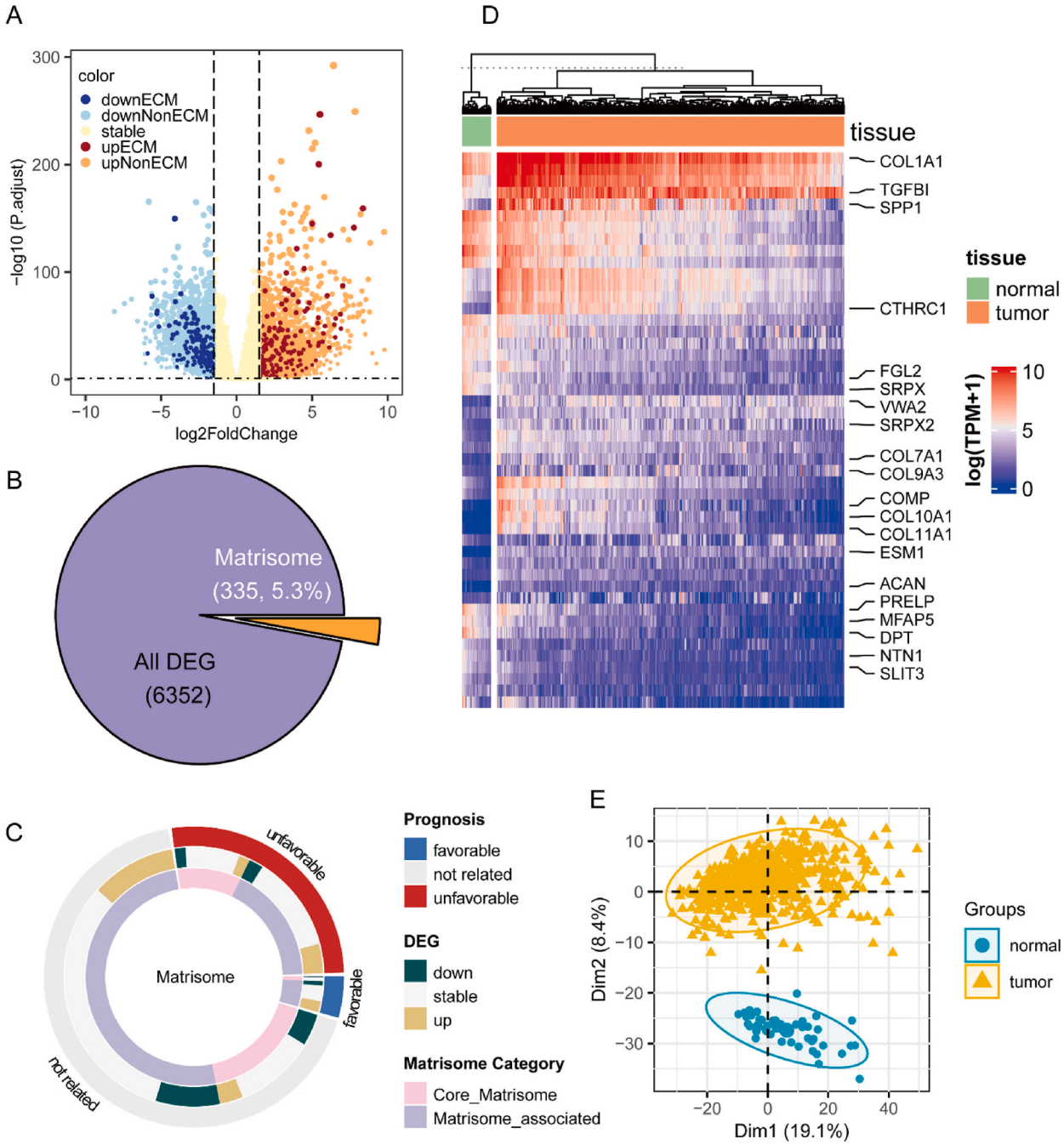


Fig. 1. Distinct ECM patterns in tumor tissues compared with normal tissues. **(A)** Volcano plot illustrating DEGs in CRC. up-regulated genes were marked by red dots, while down-regulated genes were colored blue. ECM genes were represented with triangles, while the non-ECM genes were marked with circles. **(B)** Distribution of DEGs between tumor and non-tumor. A total of 6352 DEGs were identified and 5.3 % of them were ECM genes, also referred to as "Matrisome". **(C)** Overlay of ECM categories (inner ring), expression changes (middle ring), and prognosis relevance (outer ring). **(D)** Gene expression heatmap of core ECM genes with log2 (TPM+1)>1.5 and abs (logFC) > 1.5. The top 20 DEGs were labeled along the right of the heatmap. **(E)** Principal component analysis of ECM genes expression in tumor compared with normal tissue.

predicts transcription factors based on their enrichment in gene sets derived from various databases. The algorithm calculated the integrated mean rank for each transcription factor, which reflects its relative importance and significance in regulating the core ECM genes.

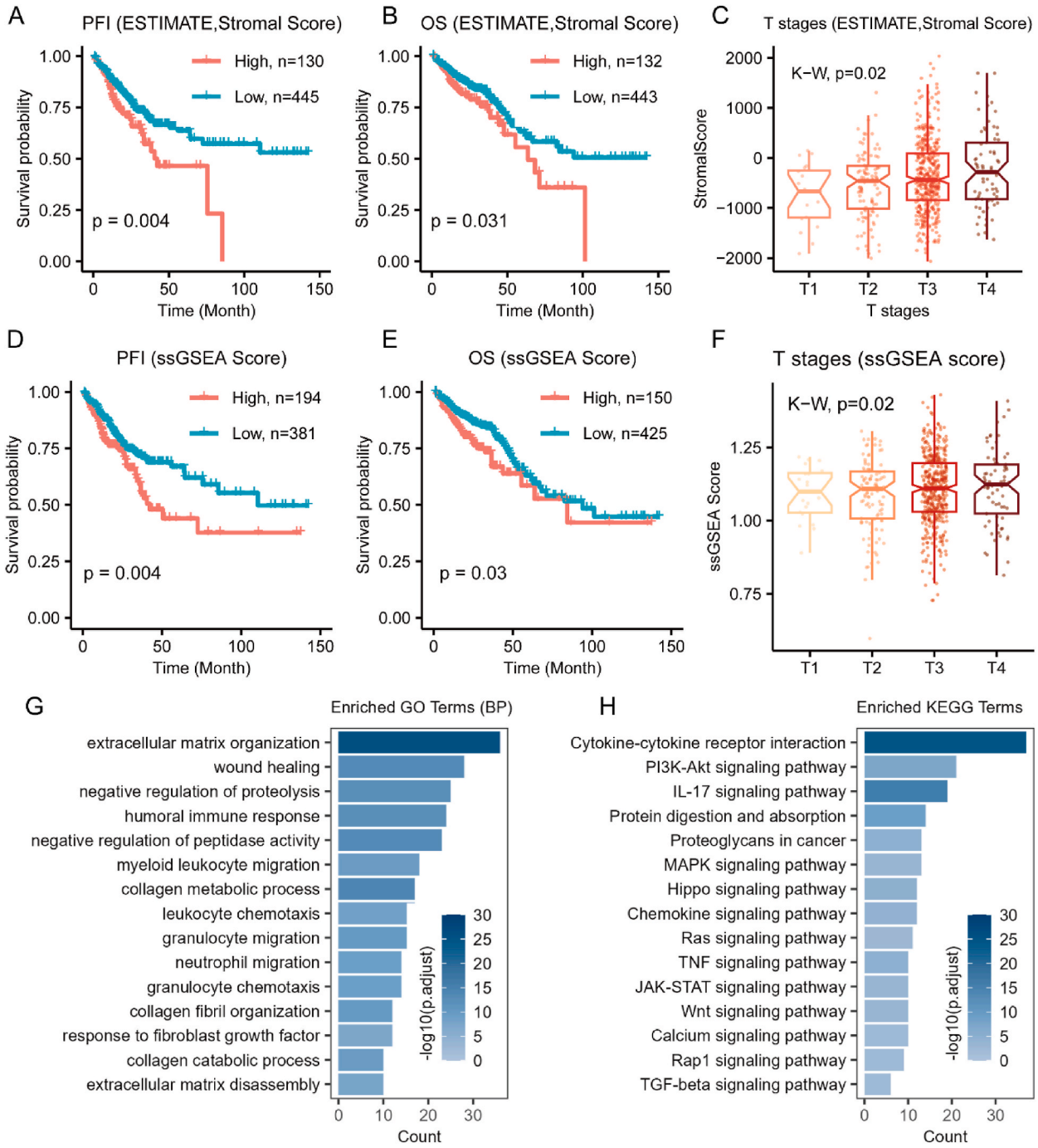


Fig. 2. Prognostic Impact of ECM Components and Associated Pathways in Colorectal Cancer. Kaplan-Meier plots showing (A) progression-free interval and (B) overall survival of patients stratified by high (red line) and low (green line) stromal scores. (C) Stromal scores across T stages. Kaplan-Meier plots representing (D) progression-free interval and (E) overall survival of patients grouped by high (red line) and low (green line) ssGSEA scores. (F) The ssGSEA scores across T stages. Top 20 notably enriched (G) GO terms and (H) KEGG pathways among differentially expressed extracellular matrix genes.

2.12. Development and validation of the ECM related prognostic model

We constructed a risk signature based on the core ECM genes that were highly expressed and significantly correlated with the progression-free interval (PFI) in colorectal cancer (CRC) patients. The TCGA-CRC cohort was split into a training group and an internal validation group, following a 6:4 distribution. We used the least absolute shrinkage and selection operator (LASSO) Cox regression with 5-fold cross-validation to select the most strongly PFI-associated prognostic signature from the core ECM genes with

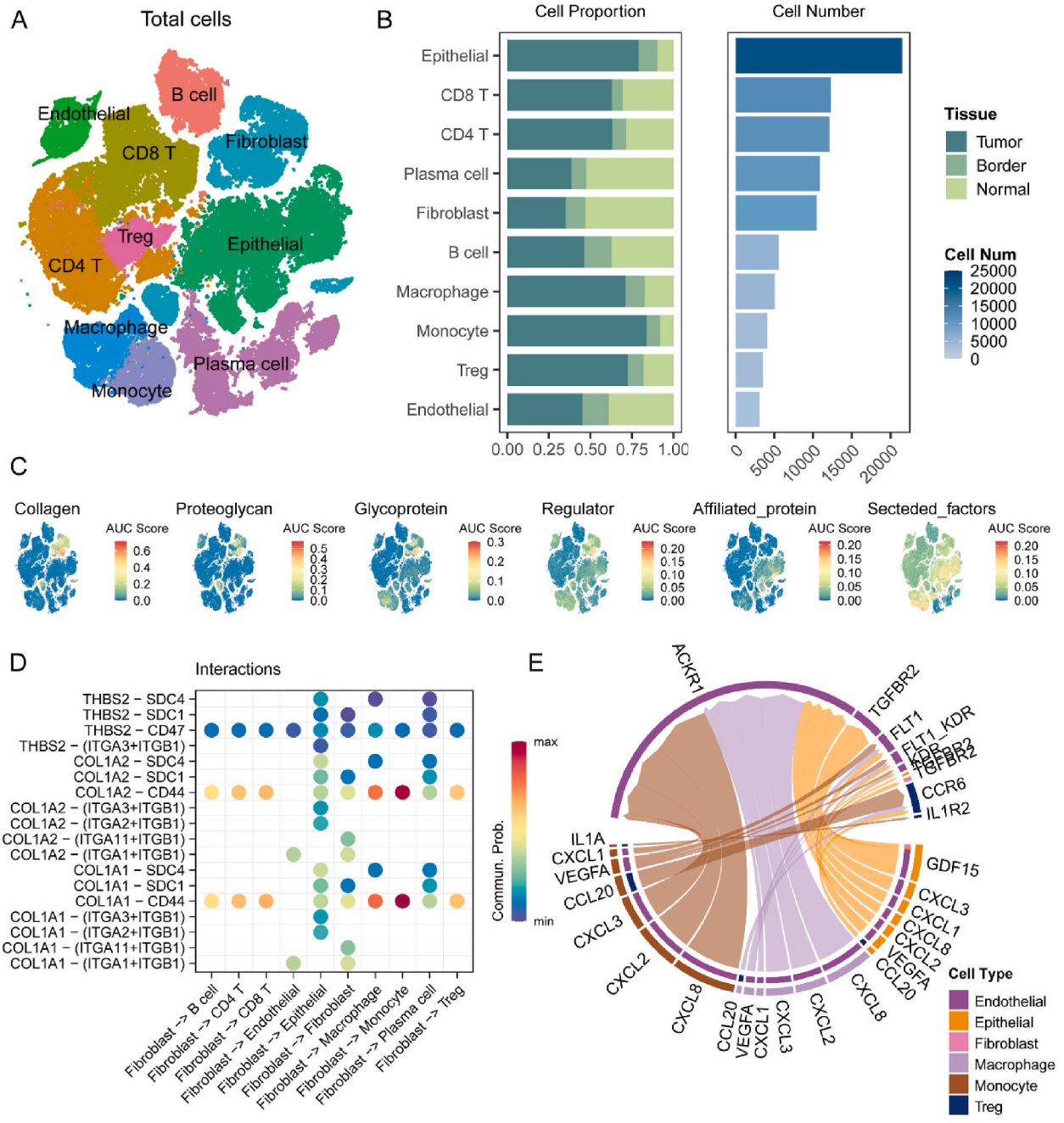


Fig. 3. Single-cell RNA sequencing atlas revealing ECM gene expression across cell types in colorectal cancer tumors. (A) Visualization of 51 colorectal cancer samples with t-distributed stochastic neighbor embedding (t-SNE) plot, which identified 10 identified cell types. (B) Distribution and abundance of the 10 cell types. (C) t-SNE plot with samples color-coded based on signature scores (blue to yellow) for up-regulated gene sets corresponding to collagens, proteoglycans, glycoproteins, extracellular matrix regulators, extracellular matrix affiliated proteins, and secreted factors. (D) Ligand-receptor signaling interactions originating from fibroblasts associated with up-regulated core extracellular matrix genes. (E) Chord diagram depicting significant ligand-receptor interactions associated with up-regulated secreted factors among different cell types.

glmnet R package (version 4.1–4) [27]. We then used a multivariate Cox regression model to determine the coefficient for each core ECM gene in the prognostic model. For each individual, a risk score was calculated by summing the products of gene expression and their corresponding coefficients. The sensitivity and specificity of the risk scores in predicting patient survival were measured with receiver operating characteristic (ROC) curves and the area under the ROC curves (AUC values).

3. Results

3.1. Distinct ECM patterns were identified in CRC samples

To comprehensively profile the ECM landscape in CRC, we conducted a comparative analysis of gene expression between CRC tumor samples and non-tumor samples using bulk transcriptomic data from the TCGA database. We identified 6352 DEGs using the DESeq2 algorithm (Fig. 1A). About 5.3 % (335/6352) of DEGs are categorized as ECM genes. ECM genes can be classified as core ECM genes and ECM-associated genes. Core ECM genes code insoluble proteins that form structural components of the ECM, while the ECM-associated genes code proteins that bind to ECM or regulate the formation of the ECM [12]. DEGs include 93 core ECM genes (see Supplementary Table 1) and 242 ECM-associated genes (Fig. 1B and C). From another aspect, more than one-third (335/941, 35.6 %) of all the ECM genes are included in the DEGs (see Supplementary Table 2). Additionally, a proportion test revealed that the ratio of DEGs in the ECM gene set was significantly higher than in other genes, indicating a potentially important role for ECM genes in tumorigenesis (Supplementary Table 3). We further performed survival analysis, which indicated that 46 ECM genes were notably up-regulated and correlated with poor prognosis (see Supplementary Table 2). SFTA2, SERPINA4, COMP, IBSP, and ESM1 were the top five of these genes, as illustrated in Supplementary Table 2.

To identify the top dysregulated genes in core ECM genes, we analyzed those with log-transformed transcripts per million (TPM) $>$ 1.5 and ignored those with relatively low expression levels (Fig. 1D). The genes with the top 20 log fold changes were marked beside the heatmap. The most upregulated collagens were COL10A1, COL11A1, COL7A1, COL9A3, and COL1A1, while the most upregulated ECM glycoproteins were COMP, VWA2, SPP1, CTHRC1, and TGFB1. In addition, we performed PCA analysis with the 941 ECM genes. ECM genes could clearly distinguish the tumor samples from normal samples by themselves (Fig. 1E), suggesting distinct ECM patterns in the tumor tissue.

Collectively, these results indicated the distinct expressed ECM genes in CRC tissues and were likely to contribute to the progress of colorectal cancer.

3.2. ECM genes showed prognostic value and participated in the progress of CRC

To explore the clinical relevance of ECM genes, we applied two algorithms: the ESTIMATE and ssGSEA algorithms. The ESTIMATE algorithm generates stromal and immune scores by utilizing predefined gene signatures specific to stromal and immune cells, respectively. These scores could be combined to calculate the tumor purity score, which indicates the relative abundance of cancer cells compared to non-cancerous immune and stromal cells within the tumor tissue. Additionally, we applied the ssGSEA method to evaluate the enrichment score of the core ECM signature, providing further insights into the stromal components.

Kaplan–Meier plots illustrated that patients with higher stromal scores exhibited significantly shorter OS and PFI compared to those with lower stromal scores (Fig. 2A and B). These trends were consistently observed across patients with varying core ECM signature scores, as demonstrated in Fig. 2C and D. Additionally, stromal scores significantly increased with the process of advancing T stage ($P = 0.02$) (Fig. 2C). Similarly, core ECM signature scores also showed a significant increase with higher T stage ($P = 0.02$) (Fig. 2C and D).

To identify the potential role of ECM genes in tumorigenesis and disease progression, we conducted GO and KEGG enrichment analysis on the up-regulated ECM genes in tumor tissues. The GO analysis indicated that these differential expressed ECM genes were primarily associated with extracellular matrix organization, collagen metabolism, leukocyte chemotaxis, and myeloid leukocyte migration (Fig. 2G). These processes are intimately linked to extracellular remodeling and immune cell infiltration, implying their potential influence on tumor microenvironment dynamics. Meanwhile, the KEGG analysis, illustrated in Fig. 2H, highlighted that the up-regulated ECM genes were notably enriched in pathways like cytokine-cytokine receptor interaction and various cancer-related pathways. These findings collectively underscore the pivotal role that ECM genes may play in the initiation and progression of tumor immune and tumorigenesis.

3.3. Signature ECM genes are expressed in different cell types from a single-cell perspective

To investigate the origins and potential impacts of ECM genes within the TME of CRC, we utilized a publicly available CRC single-cell dataset [28]. This dataset comprises 88,441 cells from 51 colorectal samples, encompassing 29 tumor samples, 6 tumor border samples, and 16 normal samples. These cells were identified as 10 major groups according to their expression of specific marker genes (Fig. 3A and B, and Supplementary Fig. 1). Additionally, we assessed for batch effects and confirmed that no significant batch effects were present (Supplementary Fig. 2).

Upon analyzing the signature scores across these ten cell clusters, we observed a distinct pattern. As depicted in Fig. 3C and Supplementary Fig. 3A, fibroblasts are the major cell type secreting collagens and glycoproteins. While several immune cells, endothelial cells, and epithelial cells expressed high levels of regulator, affiliated protein and secreted factor genes. This finding suggests the critical fiber-secreting role of fibroblasts and the regulatory role of non-fibroblast cells within tumor tissue. Consistently, as shown in

Supplementary Figs. 3B and C, we compared the signature score of collagens and glycoproteins in tumor, border, and normal tissues. These data indicated that the collagen and glycoprotein signature score increased continuously in border and tumor tissues. These findings strongly suggest the involvement of up-regulated collagen, glycoprotein and multi-cell regulation in the complex process of

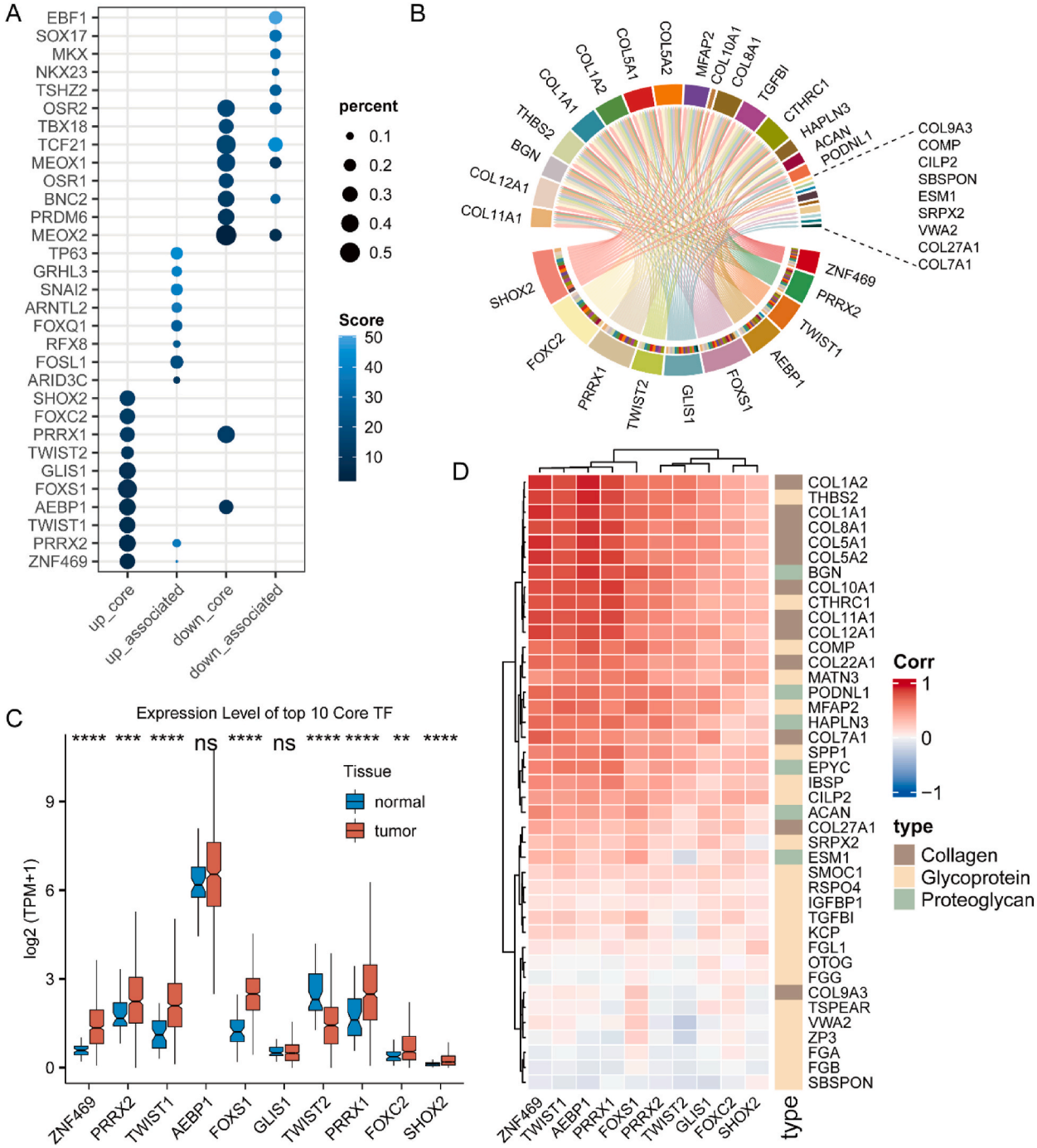


Fig. 4. Analysis of transcription factor enrichment of top dysregulated ECM genes. **(A)** Transcription factor enrichment analysis of top 10 dysregulated core ECM genes, and ECM associated genes. Dot size indicates the percentage of genes in that cluster mapping to that transcription factor, and color indicates the rank of the chEA score (low number = high enrichment strength). **(B)** Circos plot showing gene regulatory relationships, with TFs represented in the lower part and target genes in the upper part. **(C)** Boxplot of expression levels of transcription factors regulating the top 10 up-regulated core ECM genes in tumor (red) and normal (blue) tissues. **(D)** Pearson correlation coefficient heatmap of up-regulated 10 transcriptional factors and 41 corresponding core ECM genes.

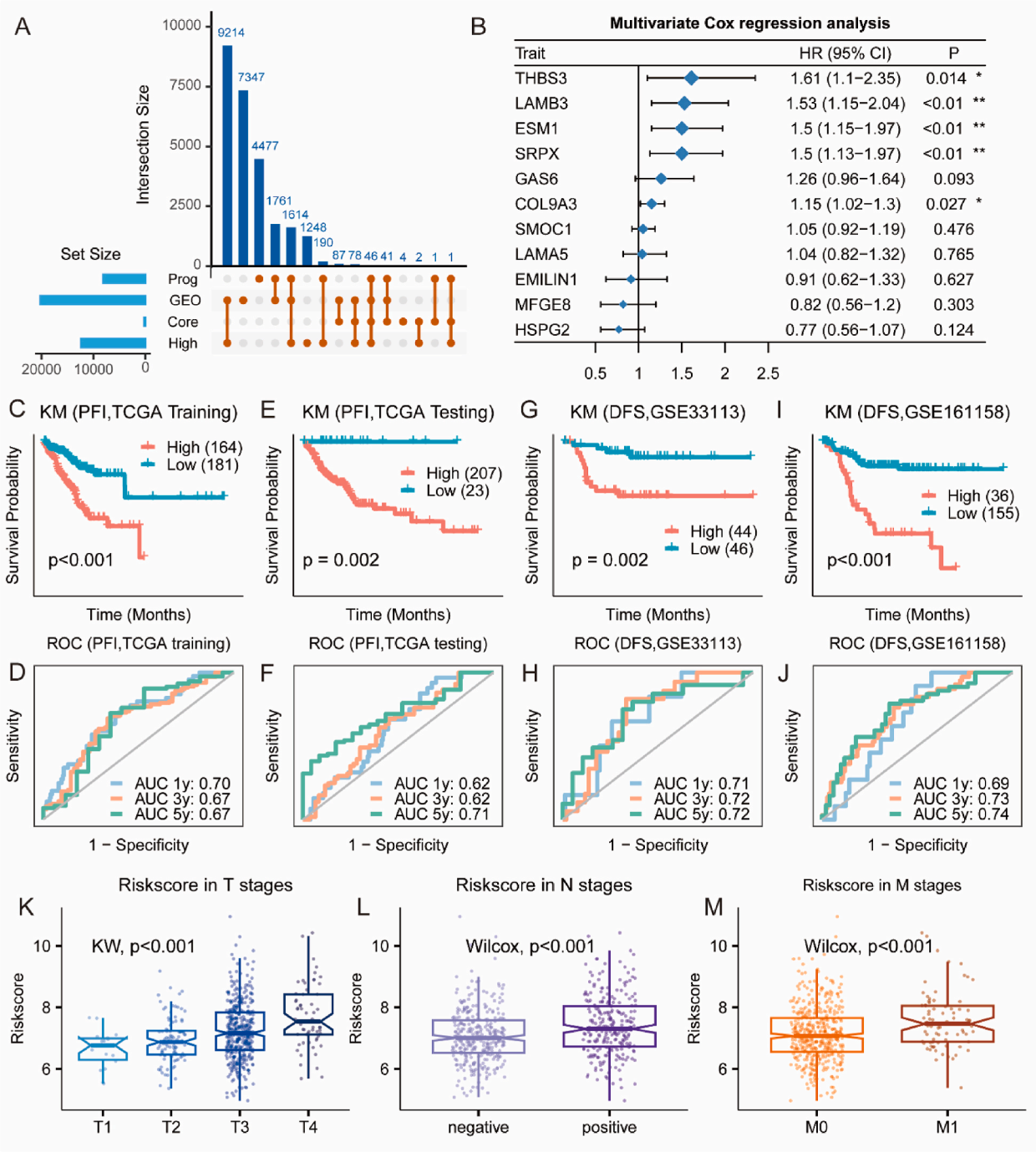


Fig. 5. Development and validation of prognostic model utilizing core ECM genes. **(A)** Venn diagram of included genes to construct the prognostic signature. **(B)** The hazard ratio of model genes. **(C, E, G, I)** Individuals with higher risks have worse prognosis in the TCGA training set, TCGA testing set, and GSE33113, GSE16158. **(D, F, H, J)** The ROC curve of the model for 1, 3, and 5 years. **(K)** Signature risk scores in four T stages of CRC patients, with *p*-value calculated using the Kruskal-Wallis test. **(L, M)** Signature risk scores in N and M stages, with the p-value calculated using the Wilcoxon rank sum test.

tumorigenesis.

3.4. Potential cell-cell interactions involved with ECM components and CRC progression

To elucidate the intricate cell-cell communication that was related to up-regulated ECM components during CRC, we use CellChat R packages to infer the cell-cell interactions. As represented in Supplementary Figs. 4A and B, fibroblasts were closely engaged in interactions with immune cells, endothelial cells, and epithelial cells. Subsequently, we analyzed communication pathways linked to the up-regulated collagens and glycoproteins, our findings indicated that COLLAGEN and TBHS pathways potentially played pivotal roles in these cell-cell interactions. These pathways are secreted from fibroblasts and are closely associated with up-regulated core ECM components (see in Supplementary Table 4). Moreover, our analysis pinpointed the COL1A1 signaling pathway as the most significant contributor to communication between fibroblasts and macrophages and monocytes (Fig. 3D). This was further validated by exploring the correlation between THBS2 and COL1A1 with markers associated with regulatory T cells and macrophages in the TCGA-CRC cohort (Supplementary Fig. 4C), revealing the potential proliferation and activation of immunosuppressive cells by the collagen and THBS pathway.

ECM-related secreted factors are recognized or presumed to bind to structural elements of the ECM [12]. As indicated in Supplementary Fig. 3A, secreted factors were notably elevated in non-fibroblasts of CRC tumor tissue, we wonder whether these ECM-related secreted factors play a role in CRC progression. Consequently, we analyzed signaling pathways originating from these non-fibroblast cells and associated with the up-regulated secreted factors (Fig. 3D, Supplementary Fig. 4D). Intriguingly, most of these pathways targeted regulatory T cells and endothelial cells. Particularly noteworthy was the notable association between the expression of CXCL8 and higher levels of endothelial markers such as PECAM1, FLI1, and ERG in the TCGA-CRC cohort, suggesting a potential role for secreted factors CXCL8 in tumor angiogenesis (Supplementary Fig. 4E).

Collectively, these data underscore the potential tumor-promoting effect of up-regulated ECM components within the colorectal cancer microenvironment. They appear to render this effect through the promotion of immunosuppressive cells, including macrophages and regulatory T cells, as well as by influencing angiogenesis.

3.5. Transcriptional regulation network of up-regulated ECM components in tumor tissues

To explore the molecular mechanisms that drive the changes in ECM patterns, we predicted the transcription factors (TFs) that may be associated with the observed ECM gene expression alterations. We conducted a transcription factor analysis using the chEA3 package. In this study, we identified the top 10 prioritized TFs and related target genes for each module (Fig. 4A and B). Consistently, most of the TFs were markedly up-regulated and positively associated with the target ECM gene expression in the TCGA-CRC cohort (Fig. 4C and D), which further confirmed the TF relationships with the ECM targets.

These results indicated that the TFs may modulate the ECM gene expression and contribute to the ECM remodeling in CRC.

Table 1
Univariate and multivariate Cox analysis of risk score and clinical features.

Characteristic	Univariate analysis			Multivariate analysis		
	HR	95 % CI	p-value	HR	95 % CI	p-value
Riskscore	1.6	1.37, 1.87	<0.001	1.49	1.23, 1.80	<0.001
Age	1	0.98, 1.01	0.54	1.01	0.99, 1.03	0.2
Gender						
Female	–	–		–	–	
Male	1.35	0.97, 1.88	0.073	1.12	0.79, 1.60	0.5
pT						
T1	–	–		–	–	
T2	0.81	0.17, 3.80	0.79	0.55	0.11, 2.88	0.5
T3	2.61	0.64, 10.6	0.18	0.8	0.10, 6.10	0.8
T4	8.12	1.94, 34.0	0.004	1.74	0.22, 13.7	0.6
pN						
Negative	–	–		–	–	
Positive	2.76	1.98, 3.85	<0.001	0.76	0.29, 1.98	0.6
pM						
M0	–	–		–	–	
M1	5.78	4.02, 8.31	<0.001	8.63	1.78, 41.9	0.007
AJCC_stage						
I	–	–		–	–	
II	2.43	1.09, 5.41	0.029	1.53	0.34, 6.98	0.6
III	3.81	1.71, 8.46	0.001	2.47	0.49, 12.5	0.3
IV	14.7	6.66, 32.5	<0.001			
Location						
Colon	–	–		–	–	
Rectum	0.94	0.65, 1.37	0.74	0.96	0.64, 1.45	0.8

3.6. Development and evaluation of the prognostic model based on tumor-associated ECM

To further explore the prognostic availability of the ECM genes in CRC, we developed a prognostic model of CRC patients based on up-regulated core ECM components in the TCGA-CRC cohort. We integrated clinical information with gene expression profiles. The TCGA-CRC cohort was partitioned as a training cohort and an internal validation cohort with a 6:4 distribution. Venn diagram indicated the filtering strategy of 46 up-regulated core ECM genes included in the model construction (Fig. 5A). Then, we performed a LASSO regression analysis 5-fold cross-validation to identify the best ECMs for creating a prognostic model (Supplementary Table 5). Lastly, we applied a multivariate Cox regression analysis to create a prognostic signature and estimate the corresponding coefficients of the ECM genes selected through the LASSO analysis (Fig. 5B).

The final model comprises THBS3, LAMB3, ESM1, SRPX, and COL9A3, with the risk score calculated as the aggregate of the products of gene expression values and their corresponding coefficients. The specific formula was as follows: RiskScore = $\sum \beta i \times Expi$, where i indicates the $\log_2$ (TPM+1) value of the five selected genes, and β corresponds to their Cox regression coefficients. This is the final formula for determining risk scores using the 5-gene signature:

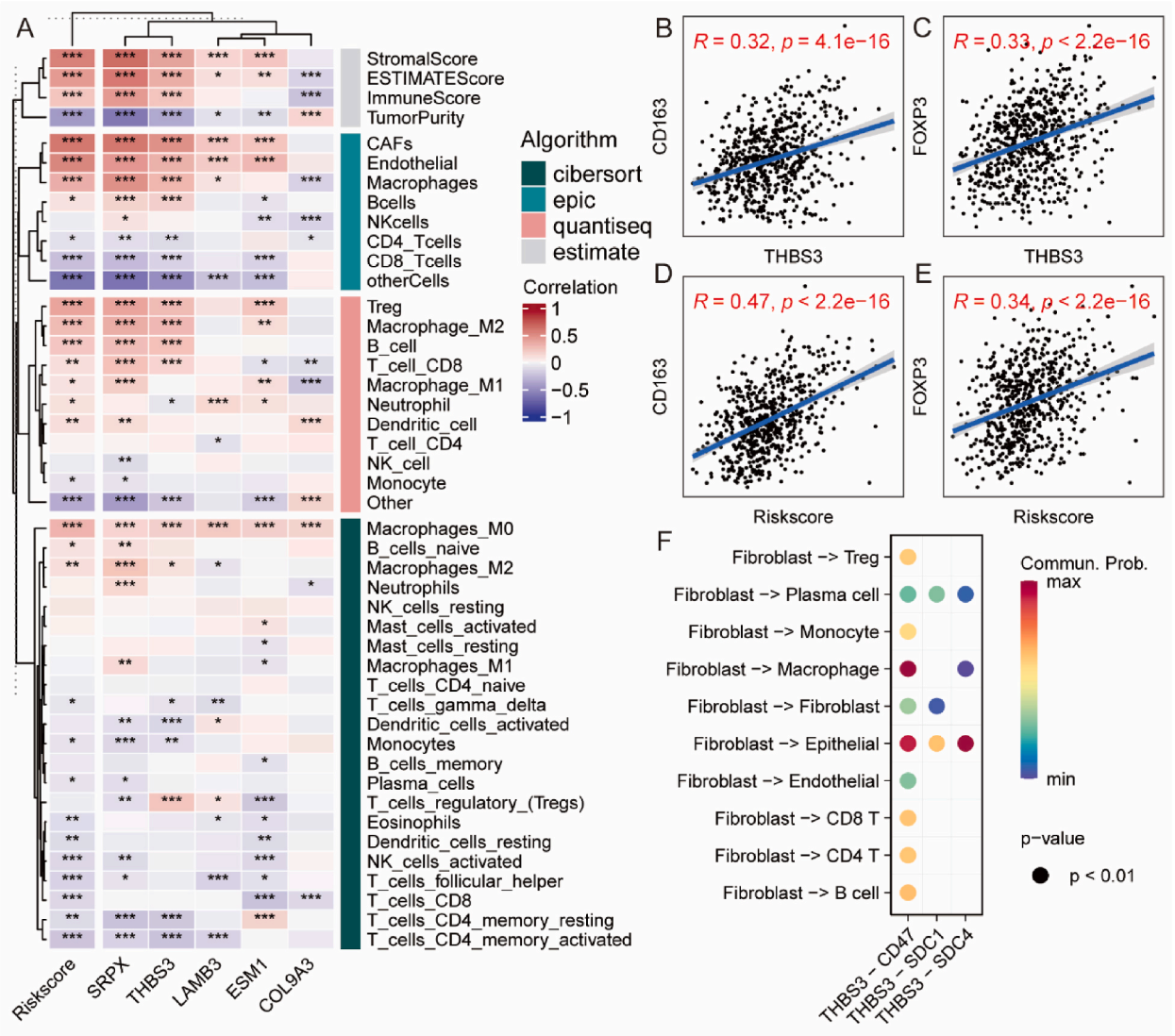


Fig. 6. Correlations between immune cell infiltration levels and signature scores. (**A**) Heatmap of Pearson correlation coefficients between signature scores and estimated immune cell infiltration levels calculated using four algorithms: CIBERSORT, EPIC, QUANTISEQ, and ESTIMATE. Statistical significance is marked as follows: * for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$. (**B**, **C**, **D**, **E**) Pearson's correlation of the expression level of THBS3 and overall signature risk score with markers for macrophages and Tregs, including CD163 and FOXP3. (**F**) Ligand-receptor signalings sent from fibroblasts that were associated with THBS3. The scale color represents the level of communication probability from minimum (blue) to maximum (red).

RiskScore = 0.476 * THBS3 + 0.428 * LAMB3 + 0.409 * ESM1 + 0.403 * SRPX +0.137 * COL9A3.

Then we assessed the performance of this model using the TCGA-CRC training and validation sets, as well as in two external GEO datasets, GSE33113 and GSE161158. The Kaplan–Meier plots revealed that patients in the high-risk category had a notably shorter PFI ($p < 0.001$ for TCGA training set and $p = 0.002$ for TCGA validation set) as well as a shorter DFS in the GEO datasets ($p = 0.002$ for GSE33113 and $p < 0.001$ for GSE161158) (Fig. 5C–E, G, I). We also plotted the time-dependent ROCs for the model. The model showed 1-, 3-, and 5-year AUCs of 0.70, 0.67, and 0.67 for the training group, while in the validation group, the ACUs were 0.62, 0.62, and 0.71 (Fig. 5D and F). The external validation for the model with the two GEO datasets are presented in Fig. 5H and **J.** Furthermore, the model was validated with additional GEO datasets, GSE17536, GSE17537, and GSE38832 (Supplementary Fig. 5), demonstrating similar prognostic value.

In addition, we applied the model to the whole TCGA-CRC cohort and examined the association of risk scores with clinical features. We found that the risk score escalated in correlation with the progression of the T, N, and M stages (Fig. 5K, L, M), indicating a strong association between the risk score and tumor development. We also carried out both univariate and multivariate Cox regression analysis on risk score and clinical characteristics. Our results indicated that the risk score was a significant predictor of survival (Table 1). These findings imply that the model based on upregulated core ECM genes could be a reliable prognostic indicator.

Collectively, these data indicate that the model, which was built from upregulated core ECM genes, could serve as a valuable prognostic predictor for CRC. In addition, upregulated ECM gene expression suggested a negative effect on CRC patient survival.

3.7. The relevance of the prognostic model and tumor immune landscape

To further investigate the internal link between the prognostic model and TME, a correlation analysis was performed to assess the relationship between the risk score of the prognostic model and the immune cell infiltration level using four methods, namely CIBERSORT, ESTIMATE, quanTIseq, and EPIC (Fig. 6A). The risk score and the signature gene expression levels were significantly associated with the relative abundance of regulatory T cells, M2 macrophages, and endothelial. Compared to single gene expression, the risk score showed more significant correlations with immune cells. Moreover, the THBS3 expression level and the risk score were significantly related to the Treg and M2 macrophage markers (Fig. 6B, C, D, E). Cell-cell interaction analysis, CellChat, showed that the THBS3 produced by fibroblasts might participate in several pathways that communicate with macrophages, Tregs, epithelial, and endothelial (Fig. 6F). These findings suggested that the central gene of the signature score might play a tumor-promoting role by interacting with immune suppressive cells and tumor angiogenesis.

4. Discussion

Colorectal cancer ranks as the third most common cancer and the second leading cause worldwide [1]. However, treatment options and efficacy remain limited [2]. As a key component in TME, ECM has drawn more and more attention in recent years [29]. The ECM not only provides structural support for cancer cells but also modulates cell-cell or cell-ECM interactions [8]. Cancer cells and tumor-associated stromal cells may employ the ECM remodeling process to create a tumor-prompting environment, thus facilitating tumorigenesis, tumor invasion, and immune evasion [30]. Therefore, the ECM could be considered a potential target for innovative cancer therapy. Unfortunately, we still do not fully understand the role of ECM in colorectal cancer, as well as the molecular mechanism through which ECM exerts its tumor-prompting effect. In this study, we analyzed the clinical implications of ECM deposition, and we explored the potential molecular mechanism of ECM exerting its tumor-prompting effect. In addition, we constructed a prognostic model that demonstrated reliable stability and precision across both training and validation sets.

PCA analysis indicated distinct ECM patterns in tumor tissues compared with normal tissues. In addition, the abundance of ECM was expected to be closely associated with patients' survival, and the prognostic model based on core ECM genes showed relatively high predictive values. Consistent with previous studies [31], amount as well as the composition of ECM, which might affect the biochemical and mechanical ECM properties, could be fundamental in tumorigenesis and tumor progression of CRC patients. Besides, we explored the top dysregulated ECM components, including collagens such as COL1A1, COL7A1, COL9A3, COL10A1, COL11A1, and TGFB1. Collagens influence the physical and biochemical properties of the TME, which in turn affects the migration, polarity, and signaling of cancer cells [32]. Collagen type I alpha 1 (COL1A1), a primary constituent of type I collagen, is reported that increased COL1A1 expression was significantly correlated with serosal invasion, lymph node metastasis, and distant metastasis status in patients with CRC [33]. However, the roles of different collagens involved in cancers have not been thoroughly determined. This study provides a comprehensive view of collagen composition changes in CRC, which might facilitate new drug development.

We further explored the molecular mechanism involved with ECM components, integrating both the single-cell and bulk RNA sequencing datasets. We found that the ECM components were closely related to various cancer-associated signaling, cell migration, and extracellular remodeling processes, which were vital in tumorigenesis and tumor invasion [34]. Moreover, we demonstrated that most up-regulated ECM components are derived from fibroblasts, which are also termed cancer-associated fibroblasts (CAF) [35], in the single-cell resolution. These data highlight the importance of CAF in the tumorigenesis process of colorectal, as in pancreatic cancer, breast cancer, and other solid tumors [36]. Furthermore, we have identified that THBS2 and collagen type I (including COL1A1 and COL1A2) frequently interact with epithelial cells and immune suppressive cells including regulatory T cells and macrophages. It was reported that THBS2 was a biomarker for a unique subset of CAF which plays a crucial role in driving the aggressiveness of early-stage lung cancer [37], and THBS2 might be crucial in tumorigenesis in colorectal as well, which needs further investigation. In addition to genes expressed by fibroblasts, there are secreted factors, such as CXCL8 and other chemokines, that highly expressed by non-fibroblast cells in CRC TME. CXCL8 has been demonstrated to promote immune evasion in gastric cancer [38]. Collagen genes and

collagen-related regulatory factors may display a broader spectrum of effects in the progression of CRC.

Transcriptional factors (TFs) play a critical role as intracellular regulators in numerous vital biological processes. The association of TFs and up-regulated core ECM genes were visualized with a network, which was consistent with the correlated expression levels between TFs and ECM genes. Above all, SHOX2 was reported to be involved with breast cancer metastasis [39], and PRRX1 functions as a key transcription factor driving the differentiation of stromal fibroblasts into myofibroblastic lineage [40]. These results suggest another opportunity for cancer treatment by targeting central TFs.

While this study has yielded clinically actionable insights from ECM profiling of tumors, it has some limitations. First, all the results were based on RNA-seq analysis, and further experimental verification is required. Second, our study cohort for constructing the prognostic model was collected from different sequencing platforms, the intratumor or intrapatient tumor heterogeneity was inevitable. Nonetheless, our prognostic model demonstrated that the ECM components were significantly related to poor clinical outcomes in patients with CRC.

5. Conclusion

In summary, this study analyzed ECM components in tumor and normal samples and identified several ECM components that were dysregulated in colorectal cancer. These ECM components were involved in various cell-cell interactions and correlated with tumor development and poor survival outcomes. The genes encoding these ECM components could be prognostic markers and potential targets for novel therapeutic interventions in colorectal cancer.

Funding statement

This study received funding from the Grant of National Key Technologies R&D Program (No. 2015BAI13B09), National Key Technologies R&D Program (No. 2017YFC0110904), Clinical Center for Colorectal Cancer, Capital Medical University (No. 1192070313), Distinguished Young Scholars from Beijing Friendship Hospital (yyqcjh2022-4), and Beijing Municipal Administration of Hospitals' Youth Programme (No.QML20230106).

Data availability statement

The datasets supporting our study are available at https://www.ncbi.nlm.nih.gov/geo with accession numbers GSE132465, GSE144735, GSE17536, GSE17537, GSE33113, GSE38832 and GSE161158, as well as in https://portal.gdc.cancer.gov/with identifiers of TCGA-COAD and TCGA-READ.

CRediT authorship contribution statement

Xiaobao Yang: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jiale Gao:** Writing – original draft, Methodology, Data curation. **Tianzhen Zhang:** Visualization, Methodology, Data curation. **Lu Yang:** Validation, Methodology, Conceptualization. **Chao Jing:** Methodology, Data curation, Conceptualization. **Zhongtao Zhang:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Dan Tian:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36164.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, et al., Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 71 (3) (2021) 209–249, https://doi.org/10.3322/caac.21660.
[2] L.H. Biller, D. Schrag, Diagnosis and treatment of metastatic colorectal cancer: a review, JAMA 325 (7) (2021) 669–685, https://doi.org/10.1001/jama.2021.0106.
[3] Y. Xiao, D. Yu, Tumor microenvironment as a therapeutic target in cancer, Pharmacol. Ther. 2212021) 107753, https://doi.org/10.1016/j.pharmthera.2020.107753.
[4] Y. Chhabra, A.T. Weeraratna, Fibroblasts in cancer: unity in heterogeneity, Cell 186 (8) (2023) 1580–1609, https://doi.org/10.1016/j.cell.2023.03.016.
[5] M.Z. Jin, W.L. Jin, The updated landscape of tumor microenvironment and drug repurposing, Signal Transduct. Targeted Ther. 5 (1) (2020) 166, https://doi.org/10.1038/s41392-020-00280-x.

[6] O. Elhanani, R. Ben-Uri, L. Keren, Spatial profiling technologies illuminate the tumor microenvironment, Cancer Cell 41 (3) (2023) 404–420, https://doi.org/10.1016/j.ccell.2023.01.010.

[7] X. Mao, J. Xu, W. Wang, et al., Crosstalk between cancer-associated fibroblasts and immune cells in the tumor microenvironment: new findings and future perspectives, Mol. Cancer 20 (1) (2021) 131, https://doi.org/10.1186/s12943-021-01428-1.

[8] S. Brassart-Pasco, S. Brézillon, B. Brassart, L. Ramont, J.B. Oudart, J.C. Monboisse, Tumor microenvironment: extracellular matrix alterations influence tumor progression, Front. Oncol. 102020) 397, https://doi.org/10.3389/fonc.2020.00397.

[9] H. Liu, H. Zhao, Y. Sun, Tumor microenvironment and cellular senescence: understanding therapeutic resistance and harnessing strategies, Semin. Cancer Biol. 86 (Pt 3) (2022) 769–781, https://doi.org/10.1016/j.semcancer.2021.11.004.

[10] X. Zhou, P. Zhang, N. Liu, et al., Enhancing chemotherapy for pancreatic cancer through efficient and sustained tumor microenvironment remodeling with a fibroblast-targeted nanosystem, J. Contr. Release 3612023) 161-177, https://doi.org/10.1016/j.jconrel.2023.07.061.

[11] M. Mounir, M. Lucchetta, T.C. Silva, et al., New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex, PLoS Comput. Biol. 15 (3) (2019) e1006701, https://doi.org/10.1371/journal.pcbi.1006701.

[12] X. Shao, C.D. Gomez, N. Kapoor, et al., Matrisomedb 2.0: 2023 updates to the ecm-protein knowledge database, Nucleic Acids Res. 51 (D1) (2023) D1519–D1530, https://doi.org/10.1093/nar/gkac1009.

[13] L. Kumar, F.M. E, Mfuzz: a software package for soft clustering of microarray data, Bioinformation 2 (1) (2007) 5–7, https://doi.org/10.6026/97320630002005.

[14] S. Lê, J. Josse, F. Husson, Factominer: an r package for multivariate analysis, J. Stat. Software 25 (1) (2008) 1–18, https://doi.org/10.18637/jss.v025.i01.

[15] J. Racle, K. de Jonge, P. Baumgaertner, D.E. Speiser, D. Gfeller, Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data, Elife 62017), https://doi.org/10.7554/eLife.26476.

[16] A.M. Newman, C.L. Liu, M.R. Green, et al., Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (5) (2015) 453–457, https://doi.org/10.1038/nmeth.3337.

[17] K. Yoshihara, M. Shahmoradgoli, E. Martínez, et al., Inferring tumour purity and stromal and immune cell admixture from expression data, Nat. Commun. 42013) 2612, https://doi.org/10.1038/ncomms3612.

[18] F. Finotello, C. Mayer, C. Plattner, et al., Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of rna-seq data, Genome Med. 11 (1) (2019) 34, https://doi.org/10.1186/s13073-019-0638-6.

[19] D. Zeng, Z. Ye, R. Shen, et al., Iobr: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures, Front. Immunol. 122021) 687975, https://doi.org/10.3389/fimmu.2021.687975.

[20] Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data, Bioinformatics 32 (18) (2016) 2847–2849, https://doi.org/10.1093/bioinformatics/btw313.

[21] S. Hänzelmann, R. Castelo, J. Guinney, Gsva: gene set variation analysis for microarray and rna-seq data, BMC Bioinf. 142013) 7, https://doi.org/10.1186/1471-2105-14-7.

[22] T. Wu, E. Hu, S. Xu, et al., Clusterprofiler 4.0: a universal enrichment tool for interpreting omics data, Innovation 2 (3) (2021) 100141, https://doi.org/10.1016/j.xinn.2021.100141.

[23] Y. Hao, S. Hao, E. Andersen-Nissen, et al., Integrated analysis of multimodal single-cell data, Cell 184 (13) (2021) 3573–3587, https://doi.org/10.1016/j.cell.2021.04.048.

[24] M. Andreatta, S.J. Carmona, Ucell: robust and scalable single-cell gene signature scoring, Comput. Struct. Biotechnol. J. 192021) 3796-3798, https://doi.org/10.1016/j.csbj.2021.06.043.

[25] S. Jin, C.F. Guerrero-Juarez, L. Zhang, et al., Inference and analysis of cell-cell communication using cellchat, Nat. Commun. 12 (1) (2021) 1088, https://doi.org/10.1038/s41467-021-21246-9.

[26] A.B. Keenan, D. Torre, A. Lachmann, et al., Chea3: transcription factor enrichment analysis by orthogonal omics integration, Nucleic Acids Res. 47 (W1) (2019) W212–W224, https://doi.org/10.1093/nar/gkz446.

[27] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Software 33 (1) (2010) 1–22.

[28] H.O. Lee, Y. Hong, H.E. Etlioglu, et al., Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer, Nat. Genet. 52 (6) (2020) 594–603, https://doi.org/10.1038/s41588-020-0636-z.

[29] J. Huang, L. Zhang, D. Wan, et al., Extracellular matrix and its therapeutic potential for cancer treatment, Signal Transduct. Targeted Ther. 6 (1) (2021) 153, https://doi.org/10.1038/s41392-021-00544-0.

[30] K. Khalaf, D. Hana, J.T. Chou, C. Singh, A. Mackiewicz, M. Kaczmarek, Aspects of the tumor microenvironment involved in immune resistance and drug resistance, Front. Immunol. 122021) 656364, https://doi.org/10.3389/fimmu.2021.656364.

[31] J.F. Hastings, J.N. Skhinas, D. Fey, D.R. Croucher, T.R. Cox, The extracellular matrix as a key regulator of intracellular signalling networks, Br. J. Pharmacol. 176 (1) (2019) 82–92, https://doi.org/10.1111/bph.14195.

[32] M.W. Pickup, J.K. Mouw, V.M. Weaver, The extracellular matrix modulates the hallmarks of cancer, EMBO Rep. 15 (12) (2014) 1243–1253, https://doi.org/10.15252/embr.201439246.

[33] Z. Zhang, C. Fang, Y. Wang, et al., Col1a1: a potential therapeutic target for colorectal cancer expressing wild-type or mutant kras, Int. J. Oncol. 53 (5) (2018) 1869–1880, https://doi.org/10.3892/ijo.2018.4536.

[34] J. Winkler, A. Abisoye-Ogunniyan, K.J. Metcalf, Z. Werb, Concepts of extracellular matrix remodelling in tumour progression and metastasis, Nat. Commun. 11 (1) (2020) 5120, https://doi.org/10.1038/s41467-020-18794-x.

[35] D. Lavie, A. Ben-Shmuel, N. Erez, R. Scherz-Shouval, Cancer-associated fibroblasts in the single-cell era, Nat. Can. (Ott.) 3 (7) (2022) 793–807, https://doi.org/10.1038/s43018-022-00411-z.

[36] G. Biffi, D.A. Tuveson, Diversity and biology of cancer-associated fibroblasts, Physiol. Rev. 101 (1) (2021) 147–176, https://doi.org/10.1152/physrev.00048.2019.

[37] H. Yang, B. Sun, L. Fan, et al., Multi-scale integrative analyses identify thbs2(+) cancer-associated fibroblasts as a key orchestrator promoting aggressiveness in early-stage lung adenocarcinoma, Theranostics 12 (7) (2022) 3104–3130, https://doi.org/10.7150/thno.69590.

[38] C. Lin, H. He, H. Liu, et al., Tumour-associated macrophages-derived cxcl8 determines immune evasion through autonomous pd-l1 expression in gastric cancer, Gut 68 (10) (2019) 1764–1773, https://doi.org/10.1136/gutjnl-2018-316324.

[39] Y. Teng, R. Loveless, E.M. Benson, L. Sun, A.Y. Shull, C. Shay, Shox2 cooperates with stat3 to promote breast cancer metastasis through the transcriptional activation of wasf3, J. Exp. Clin. Cancer Res. 40 (1) (2021) 274, https://doi.org/10.1186/s13046-021-02083-6.

[40] K.W. Lee, S.Y. Yeo, J.R. Gong, et al., Prrx1 is a master transcription factor of stromal fibroblasts for myofibroblastic lineage progression, Nat. Commun. 13 (1) (2022) 2793, https://doi.org/10.1038/s41467-022-30484-4.