TIMP1, CLCA1 and KLF4 are the Prognostic Biomarkers for the Development of Colon Cancer

Guoliang Zhang (124236332@qq.com)
Affiliated Hospital of Putian University
Bing Li
Affiliated Hospital of Putian University
Xuejie Xu
Affiliated Hospital of Putian University
Weidan Wu
Affiliated Hospital of Putian University
Linlin Zheng
Affiliated Hospital of Putian University

Research Article

Keywords: colon cancer, diagnosis, prognosis, biomarker, hub genes

Posted Date: February 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1306817/v1

Abstract

Objective: Colon cancer (CC) is one of the most common malignant diseases that poses a serious threat to human health. The aim of this study was to identify critical biomarkers or features that may be used as potential targets for early diagnosis and treatment monitoring of CC.

Methods: Two microarray datasets of colorectal cancer were downloaded from the Gene Expression Omnibus (GEO) database. The transcriptome and clinical data were downloaded from The Cancer Genome Atlas (TCGA) database. The method of robust rank aggregation was adopted to integrate differentially expressed genes (DEGs). The protein-protein interaction (PPI) network of the DEGs was constructed using the STRING platform, and hub genes were identified using the Cytoscape plugin cytoHubba. The gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment analysis were performed using DAVID online tool. Kaplan Meier survival analysis were used to evaluate the prognostic value of the hub genes. Receiver operating characteristic (ROC) analysis was performed to assess the diagnostic value of the genes. Pearson's correlation analysis was utilized to detect the relationship between the prognostic genes and infiltrated immune cells. The human protein atlas (HPA) database was used to verify the protein expression of the prognostic genes.

Results: A total of 501 robust DEGs were obtained by integrating datasets from GEO and TCGA. Fifty genes were screened as hub genes using Cytoscape. Functional analysis revealed that the hub genes were mainly involved in the chemical carcinogenesis, ECM-receptor interaction, focal adhesion, tight junction, PI3K-Akt signaling pathway and cytokine-cytokine receptor interaction. Six of the hub genes were associated with prognosis in CC, including TIMP1, VEGFA, KLF4, GCG, CXCL8, and CLCA1. All the 6 genes have high diagnostic value for CC identification, and were involved in some immune cells infiltration. Among the above 6 genes, the expression of TIMP1, KLF4, and CLCA1 in HPA database were mostly corresponding with gene expression profile.

Conclusions: TIMP1, CLCA1 and KLF4 were the key genes that associated with the development of CC, and may act as effective biomarkers that indicate the prognosis of CC.

Introduction

Colorectal cancer (CRC) has increased rapidly in recent years. It has become the third most common cancer and the second leading cause of cancer death worldwide [1] . In 2018, about 1.7 million patients were diagnosed with CRC every year, CC account for more than 1 million [1]. It was reported that about 20% of patients diagnosed at stage IV [1]. Early intervention is an effective measure to improve the survival rate of CC. Patients with poor prognostic factors often need more aggressive treatment. We still lack of effective treatments for patients with metastatic CC. Therefore, identifying an efficient biomarker for early diagnosis and prognostic prediction is urgently needed for CC.

Genetic alterations, such as abnormal gene expression and mutation, play a key role in cancer initiation and progression [2-5]. Analysis of abnormal expression genes can give us guidance about the mechanisms of tumorigenesis. The screening of differentially expressed genes (DEGs) in microarrays can provide us potential biomarkers for early diagnosis and treatment strategy developing.

In this study, we downloaded the mRNA microarray datasets GSE5206, GSE9348 from the Gene Expression Omnibus (GEO) database, and CC expression profile from TCGA database. These three datasets were analyzed integratedly to get common DEGs. Then comprehensive bioinformatics analysis was performed, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional and pathway analysis, protein–protein interaction (PPI) network analysis. Kaplan Meier survival analysis were used to evaluate the prognostic value of these genes. The human protein atlas (HPA) database was used to verify the protein expression of the prognostic hub genes.

Materials And Methods

Microarray data and TCGA gene expression data processing.

The microarray datasets of colorectal cancer were downloaded from GEO (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/). The included microarray data should meet the following conditions: Homo species; the minimum sample size >80; both normal samples and tumor samples were available. The raw data of expression profile were preprocessed by background correction and standardization. The probes were transformed to the corresponding genes according to the annotation information in the platforms. “limma” package in R4.1.0 was used for the identification of DEGs between colorectal cancer and normal tissue.

The transcriptome data and corresponding clinical information of patients with COAD were downloaded from the TCGA database (https://portal. gdc.cancer.gov/). DEGs was screened with “edgeR” package in R4.1.0.

DEGs identification

Criteria were set as |log2(fold change) FC|>1 and adjusted P value < 0.05. Log FC >1 was considered up-regulated while log FC < -1 was considered down-regulated. DEGs from GEO datasets and TCGA dataset were then processed integratedly to get an intersection.

GO and KEGG Pathway Analysis of DEGs.

DAVID (Database for Annotation, Visualization and Integrated Discovery, https://david.ncifcrf. gov/) is a comprehensive database that provide functional annotation tools for large lists of genes, which is often

used to identify enriched biological themes and signaling pathways. To clarify potential gene functional annotation and pathway enrichment behind those DEGs, we conducted biological processes (BP), molecular functions (MF), cellular components (CC), and KEGG pathways analysis by DAVID.P value < 0.05 was set as the significant criterion.

PPI Network and Modules Analysis.

STRING (the Retrieval of Interacting Genes, https://string-db.org/) is a powerful online web tool used for the interaction network between proteins and discovering the core regulatory genes. We imported these DEGs to STRING to detect the potential interactions among those genes, a confidence score of ≥0.4 was set as the screening threshold. Cytoscape v3.9.0 software was used to visualize the results. CytoHubba plugin was applied to screen the hub genes among these DEGs. MCODE plugin was utilized to search the functional modules inside PPI network, and node score cutoff = 0.2, k-core = 3, max. depth = 100, degree cutoff = 2, and nodules number > 4 were set as the threshold.

Survival analysis

Kaplan-Meier survival analysis was employed to explore the correlation between the prognostic genes and overall survival (OS) and progression free survival (PFS). R packages “survival” and “survminer” were used for survival analysis.

Diagnostic value assessment of the prognostic genes in colon cancer

The prognostic genes were selected based on survival analysis. Then receiver operating characteristic (ROC) analysis was performed to assess the diagnostic value of the genes using TCGA colon cancer dataset. The sensitivity, specificity, area under the curve (AUC), 95% confidence interval (CI), and the best cutoff value were recorded in ROC analysis. “pROC” package of R was used for the analysis.

Correlation of the prognostic genes with Immune cell infiltration

TCGA dataset of colon cancer was used to acquire the expression matrix of 22 types of immune cells. The CIBERSORT algorithm and R were employed for the calculation and analysis. Correlations between the prognostic genes and infiltrated immune cells was analyzed using “ggplot2” and “ggpubr” packages of R.

Protein levels of the hub genes in the HPA database

The human protein atlas (HPA) database is an open access resource for human proteins expression maps. The protein levels of hub genes in cancerous and normal tissues were detected with immunohistochemistry using the HPA database (https://www.proteinatlas.org/).

Statistical methods

R software (version 4.1.0) was utilized for statistical analysis and plot. Kaplan-Meier survival analysis was used for survival analysis. Pearson correlation analysis was used to detect the correlation confidence of the prognostic genes with Immune cell infiltration. $P<0.05$ was considered statistically significant.

Results

Identification of DEGs

Two microarray datasets with 170 colorectal cancer tissues and 17 normal tissues, and TCGA dataset with 482 CC tissue and 42 normal tissues were included in this study. According to our screen criteria, 651, 3047, and 13048 DEGs were pick out from GSE5206, GSE9348, and TCGA datasets, respectively. Subsequently, we applied the VennDiagram package of R to acquire an intersection for the DEGs. As a result, 501 DEGs were identified, including 139 up-regulated and 362 down-regulated genes (Figure 1).

Functional enrichment analysis of DEGs

All the DEGs were imported to the online tool DAVID, and GO categories and KEGG pathways were analyzed to identify potential gene functional annotation and pathway enrichment. The top 20 GO terms and pathways were present as bubble graphics using R package of ggplot2. The KEGG pathways were shown in Figure 2A, some signaling transduction and cancer developing related signaling pathways were involved, including cytokine-cytokine receptor interaction, ECM-receptor interaction, chemical carcinogenesis and PI3K-Akt signaling pathways ($P<0.05$). GO terms were shown in Figure 2B, the results showed that DEGs were mainly enriched in collagen catabolic process, extracellular matrix organization and negative regulation of growth in biological processes (BP); extracellular space, proteinaceous extracellular matrix and extracellular region in cell component (CC); chemokine activity, extracellular matrix structural constituent, and extracellular matrix binding in molecular function (MF).

PPIN construction and hub gene identification.

We utilized the STRING database to explore the interactions among the 501 DEGs. The interaction score ≥ 0.4 was set as the threshold for the network. The results showed that 494 nodes (genes) and 1586 edges (interactions) were included in the constructed PPI network, with an average node degree 6.42,

enrichment p-value< 1.0E-16. (Figure 3). Network analysis using cytoscape indicated that VEGFA has an interaction degree of 53, the highest of all. Besides, we enrolled the genes with connectivity degree > 15 as hub genes. As a result, a total of 50 genes were included. The detailed data about these hub genes were list in Table 1. Furthermore, top 3 significant functional modules with score >6 were screened using MOCDE (Figure 4A- Figure 4C). KEGG analysis of the three modules revealed that the significantly enriched pathways were mainly correlated with ECM-receptor interaction, protein digestion and absorption, chemokine signaling pathway, bladder cancer, focal adhesion, PI3K-Akt signaling pathway, tight junction, leukocyte transendothelial migration, hepatitis C (Figure 4D- Figure 4F).

Table 1

Description of the 50 hub genes

Gene	Connective degree	Expression	Full name
VEGFA	53	Up	Vascular endothelial growth factor A
MYC	47	Up	Myc proto-oncogene
CXCL8	44	Up	Chemokine (C-X-C motif) ligand 8
COL1A1	43	Up	Collagen alpha-1
CXCL12	41	Down	Chemokine (C-X-C motif) ligand 12
CCND1	39	Up	G1/S-specific cyclin-D1
TIMP1	37	Up	Metalloproteinase inhibitor 1
MMP2	35	Down	Matrix metalloproteinase-2
MMP3	34	Up	Matrix metalloproteinase-3
SPP1	34	Up	Secreted phosphoprotein 1
SOX9	33	Up	SRY (sex determining region Y)-box 9
COL1A2	33	Up	Collagen alpha-2
BGN	30	Up	Biglycan
THY1	29	Up	Thy-1 cell surface antigen
BMP2	29	Down	Bone morphogenetic protein 2
MMP1	29	Up	Matrix metalloproteinase-1
KIT	26	Down	Mast/stem cell growth factor receptor Kit
CXCL1	26	Up	Chemokine (C-X-C motif) ligand 1
SPARC	26	Up	Secreted protein, acidic, cysteine-rich (osteonectin)
CD19	24	Down	CD19 molecule
VCAN	24	Up	Versican
THBS2	24	Up	Thrombospondin 2
HGF	23	Down	Hepatocyte growth factor
SLC26A3	21	Down	Solute carrier family 26, member 3
MMP7	21	Up	Matrix metalloproteinase-7
HPGDS	20	Down	Hematopoietic prostaglandin D synthase
COL11A1	20	Up	Collagen, type XI, alpha 1
COL5A2	20	Up	Collagen, type V, alpha 2

FABP1	19	Down	Fatty acid binding protein 1, liver
GNAI1	19	Down	Guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 1
CDH11	19	Up	Cadherin 11, type 2, OB-cadherin (osteoblast)
COL4A1	19	Up	Collagen, type IV, alpha 1
COL5A1	18	Up	Collagen, type V, alpha 1
CXCL13	18	Down	Chemokine (C-X-C motif) ligand 13
CXCL5	18	Up	Chemokine (C-X-C motif) ligand 5
CCL20	18	Up	Chemokine (C-C motif) ligand 20
NR1H4	17	Down	Nuclear receptor subfamily 1, group H, member 4
ABCG2	17	Down	ATP-binding cassette, sub-family G (WHITE), member 2
KLF4	17	Down	Kruppel-like factor 4
UGT2A3	17	Down	UDP glucuronosyltransferase 2 family, polypeptide A3
ANPEP	17	Down	Alanyl (membrane) aminopeptidase
GUCA2A	17	Down	Guanylate cyclase activator 2A
COL12A1	17	Up	Collagen, type XII, alpha 1
GCG	17	Down	Glucagon
MET	16	Up	MET proto-oncogene, receptor tyrosine kinase
CD79A	16	Down	CD79a molecule, immunoglobulin-associated alpha
CLCA1	16	Down	Chloride channel accessory 1
TGFBI	16	Up	Transforming growth factor, beta-induced
TWIST1	16	Up	Twist family bhlh transcription factor 1
PLAU	16	Up	Plasminogen activator, urokinase

Survival analysis of the hub genes

We applied Kaplan Meier survival analysis to assess the prognostic value of these hub genes. The results showed that 6 genes including TIMP1, VEGFA, KLF4, GCG, CXCL8, CLCA1, were significantly associated with prognosis in CC. Among the up-regulated DEGs, high expression of TIMP1 were correlated with worse OS (Figure 5A, P=0.042) and PFS (Figure 5B, P=0.046) remarkably. High expression of VEGFA was significantly associated with worse PFS (Figure 5D, P=0.019), but not with OS (Figure 5C, P=0.054).

However, high expression of CXCL8 had a positive correlation with OS (Figure 5E, P=0.0058), no significant correlation was detected with PFS (Figure 5F, P=0.11). Among the down-regulated DEGs, low expression of CLCA1 was associated with poor OS (Figure 5K, P=0.004) and PFS (Figure 5L, P=0.0037). Low expression of KLF4 was associated with poor PFS (Figure 5J, P=0.022), but not with OS (Figure 5I, P=0.69). Low expression of GCG was associated with poor OS (Figure 5G, P=0.026), but not with PFS (Figure 5H, P=0.77).

The diagnostic value of the prognostic genes in CC

The ROC curve analysis was shown in Figure 6, all the 6 prognostic genes presented high diagnostic value for CC identification. The diagnostic parameters were AUC 95.143, 95%CI(91.872-98.414),cutoff value 246.396, sensitivity 90.87 and specificity 90.244 with TIMP1; AUC 97.292, 95%CI(95.601-98.982), cutoff value18.943,sensitivity 90.658 and specificity 95.122 with VEGFA, AUC 97.561, 95%CI(95.055-100),cutoff value 162.207, sensitivity 95.117 and specificity 95.122 with KLF4; AUC 98.674, 95%CI(97.696-99.652),cutoff value 6.046, sensitivity 94.904 and specificity 97.561 with GCG;AUC 91, 95%CI(85.809-96.191),cutoff value 8.388, sensitivity 92.357 and specificity 78.049 with CXCL8;and AUC 93.931, 95%CI(91.235-96.627),cutoff value 429.4, sensitivity 84.926 and specificity 95.122 with CLCA1,respectively(Table 2).

Table 2

Parameters of receiver operating characteristic analysis for the prognostic genes.

Gene	AUC	95% CI	Cutoff value	Sensitivity	Specificity
TIMP1	95.143	91.872-98.414	246.396	90.87	90.244
VEGFA	97.292	95.601-98.982	18.943	90.658	95.122
KLF4	97.561	95.055-100	162.207	95.117	95.122
GCG	98.674	97.696-99.652	6.046	94.904	97.561
CXCL8	91	85.809-96.191	8.388	92.357	78.049
CLCA1	93.931	91.235-96.627	429.4	84.926	95.122
AUC: area under the curve; CI: confidence interval.					

Correlation between prognostic genes and Immune cell infiltration

The top correlation between the prognostic genes and infiltrated immune cells were shown in Figure 7. TIMP1 presented a moderate positive correlation with macrophages M0; and a low negative correlation

with B cells memory, CD4 memory resting, CD4 memory activated, eosinophils, mast cell resting and plasma cells. VEGFA presented a moderate negative correlation with mast cell resting; a low negative correlation with macrophages M2 and monocytes; and a low positive correlation with macrophages M0, and NK cell resting. CXCL8 presented a moderate positive correlation with neutrophils; a low positive correlation with mast cell activated; and a low negative correlation with B cells memory, mast cell resting, monocyte, and Tregs.KLF4 exhibited a low positive correlation with macrophages M2, mast cell resting, and plasma cells; a low negative correlation with macrophages M0. GCG exhibited a moderate positive correlation with mast cell resting; a low positive correlation with B cell memory, macrophages M2 and plasma cells; and a low negative correlation with macrophages M0 and NK cell resting. CLCA1 exhibited a low positive correlation with CD4 memory resting, macrophages M2, mast cell resting and plasma cells; and a low negative correlation with macrophages M0.

Expression of the 6 prognostic genes in normal and colon cancer tissue using HPA database

In HPA database, TIMP1 was negative in normal tissues, but positively stained in some of the cancerous tissues（Figure 8A）. VEGFA was positively expressed in both normal and tumor tissues, but much stronger in some of the cancerous tissues（Figure 8B）. KLF4 was highly expressed in normal tissues, but negative or low expressed in cancerous tissues（Figure 8C）. GCG and CXCL8 were negative or low expressed in both normal and tumor tissues（Figure 8D, Figure 8E）. CLCA1 was greatly expressed in normal tissues, but negative in cancer tissues（Figure 8F）. Among the above 6 genes, the expression of TIMP1, KLF4, and CLCA1 in HPA database were mostly corresponding with gene expression profile.

Discussion

Bioinformatics analysis was an efficient method that frequently used to identify key driver genes and screen the molecular mechanisms involved in disease development[6 7,8]. To seek key genes and candidate biomarkers in the development of CC, we conducted comprehensive bioinformatics analysis based on TCGA and GEO databases. In this study, a total of 50 genes, including TIMP1, VEGFA, CXCL8, GCG, KLF4, CLCA1, were identified as hub genes, and TIMP1, CLCA1 and KLF4 was discovered to be a potential prognostic biomarker that associated with OS and/or PFS in CC.

To explore potential signaling pathways involved in CC development, we performed KEGG pathway analysis for the identified DEGs. Our findings showed that some cancer developing related signaling pathways, such as cytokine-cytokine receptor interaction, ECM-receptor interaction, chemical carcinogenesis and PI3K-Akt signaling pathways, were significantly enriched. We conducted module analysis for the PPI network using MOCDE. The results revealed that ECM-receptor interaction, PI3K-Akt signaling pathway, focal adhesion and tight junction were involved in the three modules. ECM (extracellular matrix) is a network formed by macromolecular as proteins, polysaccharides and proteoglycans secreted by cells distributed in the extracellular space. It is produced mainly by cancer

associated fibroblasts that includes connective materials of specialized fibrotic proteins, which are overexpressed in most of the cancers[9]. ECM is essential for the development of a cancer malignant phenotype[10]. The adhesion of integrins to ECM enhance the capacity of invasion for tumor cell[11]. PI3K-Akt pathway is one of the crucial signaling pathways that take part in multiple intracellular functions[12]. The activation of PI3K-Akt regulates cell survival, cell cycle progression and cellular growth, which was commonly occurred in human cancers[13-16]. Tight junctions maintain the stability of cell structure and polarity constitute through the formation of junctional complex on the apical side between epithelial cells[17]. The loss of tight junction proteins and E-cadherin are important characteristics of epithelial-mesenchymal transition (EMT)[18],which is an essential process for tumor progression. In our study, PI3K-Akt pathway and ECM-receptor interaction were enriched but tight junction pathway was inactivated with 5 members including CLDN8, CLDN23, CLDN7, CGN and TJP3, were down-regulated. Based on these signaling pathways, we can find some clues to the development of CC.

We performed Kaplan-Meier survival analysis to identify prognostic genes, and 6 genes including TIMP1, VEGFA, CXCL8, GCG, KLF4 and CLCA1 were significantly associated with OS and/or PFS. TIMP1, VEGFA and CXCL8 was up-regulated in CC, High expression of TIMP1 were associated with worse OS and PFS, while high expression of VEGFA was significantly associated with worse PFS. But high expression of CXCL8 was correlated with a better OS. CLCA1, GCG and KLF4 were down-regulated in CC. Low expression of CLCA1 was associated with worse OS and PFS, while low expression of KLF4 was associated with worse PFS, low expression of GCG was associated with worse OS. ROC analysis indicated that all the 6 genes showed high diagnostic value for CC identification, especially TIMP1, VEGFA, GCG and KLF4, with sensitivity and specificity greater than 90%.

To survey the relationship between the prognostic genes and immune cells infiltration, we obtained 22 types of infiltrated immune cells from TCGA dataset using the CIBERSORT algorithm. Correlations analysis between prognostic genes and the distribution of the infiltrated immune cells suggested that TIMP1 was positively correlated with macrophages M0. It is reported that the infiltration of M0 macrophage cells connected with an unfavorable overall survival in several cancer types[19-21]. Conversely, CLCA1 and KLF4 exhibited a negative correlation with macrophages M0. In TCGA colon cancer cohort, low expression of CLCA1 and KLF4 indicated more macrophages M0 infiltration, which was in concordance with the prognostic role of CLCA1 and KLF4.

We examed the HPA database to detect the protein levels of these prognostic genes for verification. TIMP1 protein expression was positively expressed in CC but negatively expressed in normal tissues. CLCA1 was strongly expressed in normal tissues but negative in CC. KLF4 was low expressed in CC but significantly highly expressed in normal tissue. The protein expression difference of the 3 genes between tumor and normal tissues were consistent with GEO and TCGA datasets. TIMP1, also called tissue inhibitor matrix metalloproteinase 1, has been shown to have the role of inhibiting the proteolytic activity of matrix metalloproteinases (MMPs) and adjust the balance of matrix remodeling during degradation of extracellular matrix[22-24].It was reported that TIMP1 was associated with chemotherapy drug resistance

and promote tumor progression in pancreatic carcinoma[25]. High expression of TIMP1 is able to induce cell survival via PI3-kinase pathway, and is associated with poor prognosis in melanoma[26,27]. Calcium-activated chloride channel regulator 1(CLCA1) is an activator of calcium-dependent chloride channels[28]. This gene participate in many important intracellular activities and signaling transmission, involving cancer-related proliferation, apoptosis, migration and angiogenesis[29]. KLF4 is a type of transcription factor with an evolutionarily conserved zinc finger domain, and is reported having regulatory effects on diverse cellular processes including cell proliferation, apoptosis, and differentiation[30,31]. Overexpression of KLF4 causes growth arrest in several cell lines[32,33], and leads to a reduction of the tumorigenic ability of colonic and gastric cancer cells in vivo[33,34]. The expression of KLF4 is found down-regulated in multiple human cancers[35-39].In this study, we discovered that TIMP1 was greatly over expressed, KLF4 and CLCA1 were significantly low expressed in CC. High expression of TIMP1 and low expression of CLCA1were associated with worse OS and PFS. Low expression of KLF4 was correlated with worse PFS. Thus, we speculated that TIMP1, CLCA1 and KLF4 may be potential biomarkers for CC.

In conclusion, our study identified a set of DEGs that may be involved in the progression of CC using a multi-database-based bioinformatics analysis. The most relevant signaling pathways and functional modules were revealed. The study also identified a useful molecular that TIMP1, CLCA1 and KLF4 can be used as a candidate prognostic biomarker for CC. However, further experiments are needed for molecular biological research to testify the mechanisms involved in CC development.

Declarations

Funding

The work was supported by Science and Technology Program of Putian（Grant No.2020S3F009）and Science and Technology Program of Fujian Province (Grant No.2020J011258).

Competing Interests

All the authors declare that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria, educational grants, participation in speakers' bureaus, membership, employment, consultancies, stock ownership, or other equity interest, and expert testimony or patent-licensing arrangements), or nonfinancial interest (such as personal or professional relationships, affiliations, knowledge, or beliefs) in the subject matter or materials discussed in this manuscript.

Author Contributions

Bing Li contributed to the research design and manuscript drafting; Guoliang Zhang contributed to data analysis and graphics; Xuejie Xu contributed to making important suggestions and manuscript revision Weidan Wu and Linlin Zheng participated in data acquisition. All authors have given final approval of the version to be published.

Data Availability

The datasets generated during and/or analyzed during the current study are available in the TCGA database (https://portal. gdc.cancer.gov/) and GEO database (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/).

Ethics approval

This is an observational study. The Research Ethics Committee of the affiliated hospital of Putian university has confirmed that no ethical approval is required.

References

1. Bray, F.*, et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424 (2018).
2. Zhang, L. & Shay, J.W. Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer. *J Natl Cancer Inst* **109**(2017).
3. Wolff, R.K.*, et al.* Mutation analysis of adenomas and carcinomas of the colon: Early and late drivers. *Genes Chromosomes Cancer* **57**, 366-376 (2018).
4. Wang, Y.*, et al.* Dysregulation of miR-6868-5p/FOXM1 circuit contributes to colorectal cancer angiogenesis. *J Exp Clin Cancer Res* **37**, 292 (2018).
5. Chaney, K.E.*, et al.* Cdkn2a Loss in a Model of Neurofibroma Demonstrates Stepwise Tumor Progression to Atypical Neurofibroma and MPNST. *Cancer Res* **80**, 4720-4730 (2020).
6. Hua, X., Chen, J. & Wu, L. Identification of candidate biomarkers associated with apoptosis in melanosis coli: GNG5, LPAR3, MAPK8, and PSMC6. *Biosci Rep* **39**(2019).
7. Huang, D.*, et al.* Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer Metastasis Rev* **37**, 173-187 (2018).
8. Liang, B., Li, C. & Zhao, J. Identification of key pathways and genes in colorectal cancer using bioinformatics analysis. *Med Oncol* **33**, 111 (2016).
9. Najafı, M., Farhood, B. & Mortezaee, K. Extracellular matrix (ECM) stiffness and degradation as cancer drivers. *J Cell Biochem* **120**, 2782-2790 (2019).
10. Chen, J.*, et al.* Promotion of Tumor Growth by ADAMTS4 in Colorectal Cancer: Focused on Macrophages. *Cell Physiol Biochem* **46**, 1693-1703 (2018).

11. Desgrosellier, J.S. & Cheresh, D.A. Integrins in cancer: biological implications and therapeutic opportunities. *Nat Rev Cancer* **10**, 9-22 (2010).
12. Alzahrani, A.S. PI3K/Akt/mTOR inhibitors in cancer: At the bench and bedside. *Semin Cancer Biol* **59**, 125-132 (2019).
13. Fresno Vara, J.A.*, et al.* PI3K/Akt signalling pathway and cancer. *Cancer Treat Rev* **30**, 193-204 (2004).
14. Noorolyai, S., Shajari, N., Baghbani, E., Sadreddini, S. & Baradaran, B. The relation between PI3K/AKT signalling pathway and cancer. *Gene* **698**, 120-128 (2019).
15. Ediriweera, M.K., Tennekoon, K.H. & Samarakoon, S.R. Role of the PI3K/AKT/mTOR signaling pathway in ovarian cancer: Biological and therapeutic significance. *Semin Cancer Biol* **59**, 147-160 (2019).
16. Tan, A.C. Targeting the PI3K/Akt/mTOR pathway in non-small cell lung cancer (NSCLC). *Thorac Cancer* **11**, 511-518 (2020).
17. Kyuno, D.*, et al.* Role of tight junctions in the epithelial-to-mesenchymal transition of cancer cells. *Biochim Biophys Acta Biomembr* **1863**, 183503 (2021).
18. Ikenouchi, J., Matsuda, M., Furuse, M. & Tsukita, S. Regulation of tight junctions during the epithelium-mesenchyme transition: direct repression of the gene expression of claudins/occludin by Snail. *J Cell Sci* **116**, 1959-1967 (2003).
19. Zheng, Y.*, et al.* A Novel Immune-Related Prognostic Model for Response to Immunotherapy and Survival in Patients With Lung Adenocarcinoma. *Front Cell Dev Biol* **9**, 651406 (2021).
20. Le, T., Su, S. & Shahriyari, L. Immune classification of osteosarcoma. *Math Biosci Eng* **18**, 1879-1897 (2021).
21. Huang, R.*, et al.* Identification of Hub Genes and Their Correlation With Immune Infiltration Cells in Hepatocellular Carcinoma Based on GEO and TCGA Databases. *Front Genet* **12**, 647353 (2021).
22. Song, G.*, et al.* TIMP1 is a prognostic marker for the progression and metastasis of colon cancer through FAK-PI3K/AKT and MAPK pathway. *J Exp Clin Cancer Res* **35**, 148 (2016).
23. Wang, J.C. Importance of plasma matrix metalloproteinases (MMP) and tissue inhibitors of metalloproteinase (TIMP) in development of fibrosis in agnogenic myeloid metaplasia. *Leuk Lymphoma* **46**, 1261-1268 (2005).
24. Tan, Y.*, et al.* TIMP1 down-regulation enhances gemcitabine sensitivity and reverses chemoresistance in pancreatic cancer. *Biochem Pharmacol* **189**, 114085 (2021).
25. D'Costa, Z.*, et al.* Gemcitabine-Induced TIMP1 Attenuates Therapy Response and Promotes Tumor Growth and Liver Metastasis in Pancreatic Cancer. *Cancer Res* **77**, 5952-5962 (2017).
26. Toricelli, M.*, et al.* Timp1 Promotes Cell Survival by Activating the PDK1 Signaling Pathway in Melanoma. *Cancers (Basel)* **9**(2017).
27. Palmieri, G.*, et al.* Multiple Molecular Pathways in Melanomagenesis: Characterization of Therapeutic Targets. *Front Oncol* **5**, 183 (2015).

28. Hu, D., Ansari, D., Bauden, M., Zhou, Q. & Andersson, R. The Emerging Role of Calcium-activated Chloride Channel Regulator 1 in Cancer. *Anticancer Res* **39**, 1661-1666 (2019).
29. Prevarskaya, N., Skryma, R. & Shuba, Y. Ion channels and the hallmarks of cancer. *Trends Mol Med* **16**, 107-121 (2010).
30. Ghaleb, A.M. & Yang, V.W. Kruppel-like factor 4 (KLF4): What we currently know. *Gene* **611**, 27-37 (2017).
31. Ma, Y.*, et al.* KLF4 inhibits colorectal cancer cell proliferation dependent on NDRG2 signaling. *Oncol Rep* **38**, 975-984 (2017).
32. Chen, X.*, et al.* Kruppel-like factor 4 (gut-enriched Kruppel-like factor) inhibits cell proliferation by blocking G1/S progression of the cell cycle. *J Biol Chem* **276**, 30423-30428 (2001).
33. Dang, D.T.*, et al.* Overexpression of Kruppel-like factor 4 in the human colon cancer cell line RKO leads to reduced tumorigenecity. *Oncogene* **22**, 3424-3430 (2003).
34. Zhang, N.*, et al.* Kruppel-like factor 4 negatively regulates beta-catenin expression and inhibits the proliferation, invasion and metastasis of gastric cancer. *Int J Oncol* **40**, 2038-2048 (2012).
35. Wei, D.*, et al.* Drastic down-regulation of Kruppel-like factor 4 expression is critical in human gastric cancer development and progression. *Cancer Res* **65**, 2746-2754 (2005).
36. Zhao, W.*, et al.* Identification of Kruppel-like factor 4 as a potential tumor suppressor gene in colorectal cancer. *Oncogene* **23**, 395-402 (2004).
37. Schulz, W.A. & Hatina, J. Epigenetics of prostate cancer: beyond DNA methylation. *J Cell Mol Med* **10**, 100-125 (2006).
38. Ohnishi, S.*, et al.* Downregulation and growth inhibitory effect of epithelial-type Kruppel-like transcription factor KLF4, but not KLF5, in bladder cancer. *Biochem Biophys Res Commun* **308**, 251-256 (2003).
39. Fadous-Khalife, M.C.*, et al.* Kruppel-like factor 4: A new potential biomarker of lung cancer. *Mol Clin Oncol* **5**, 35-40 (2016).

Figures

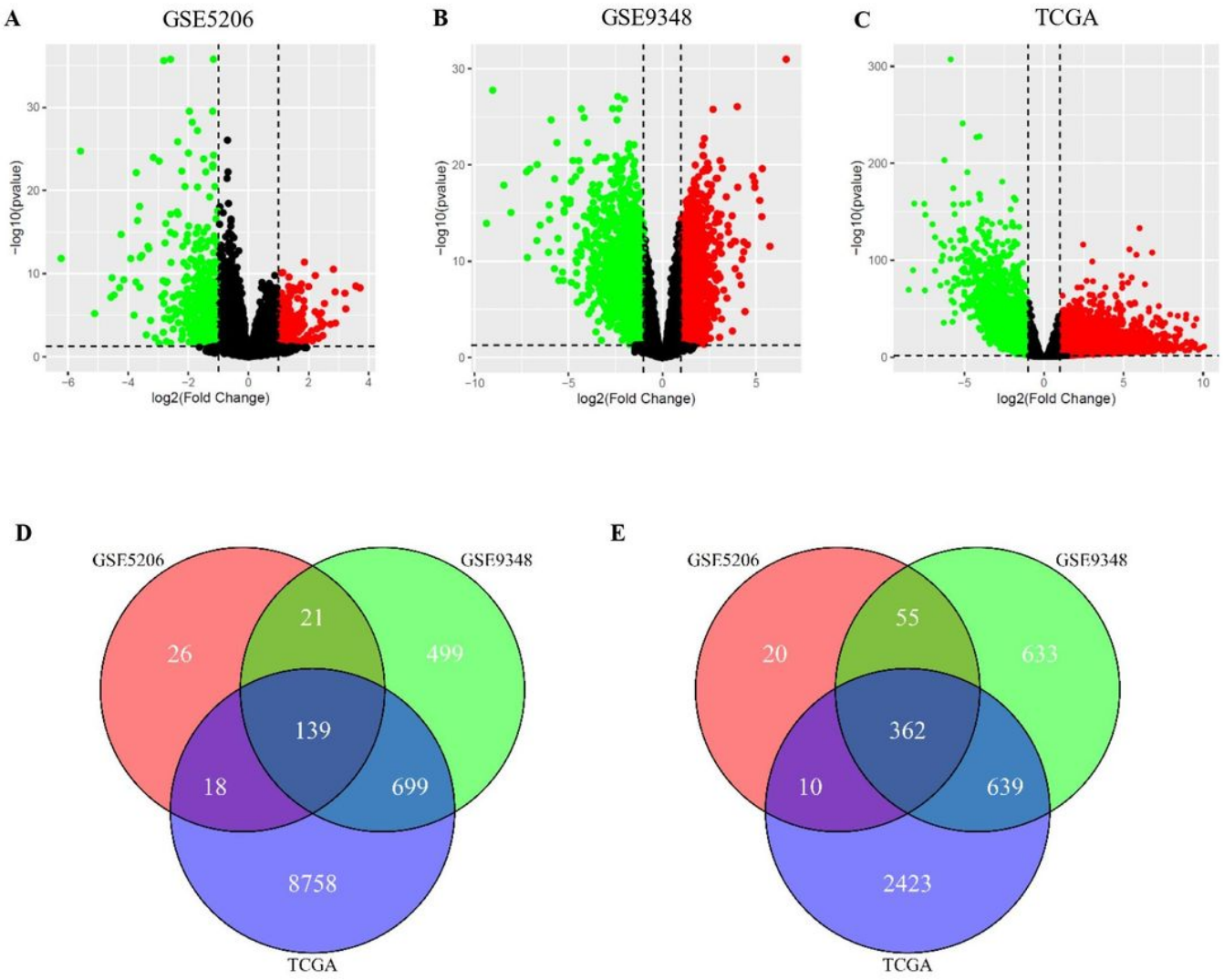


Figure 1

Identification of DEGs. (A-C) Volcano plot for DEGs from GSE5206, GSE9348, and TCGA datasets. Red points represent up-regulated DEGs, green points represent down-regulated DEGs. (D) Intersection of up-regulated DEGs from the three datasets. (E) Intersection of down-regulated DEGs from the three datasets.

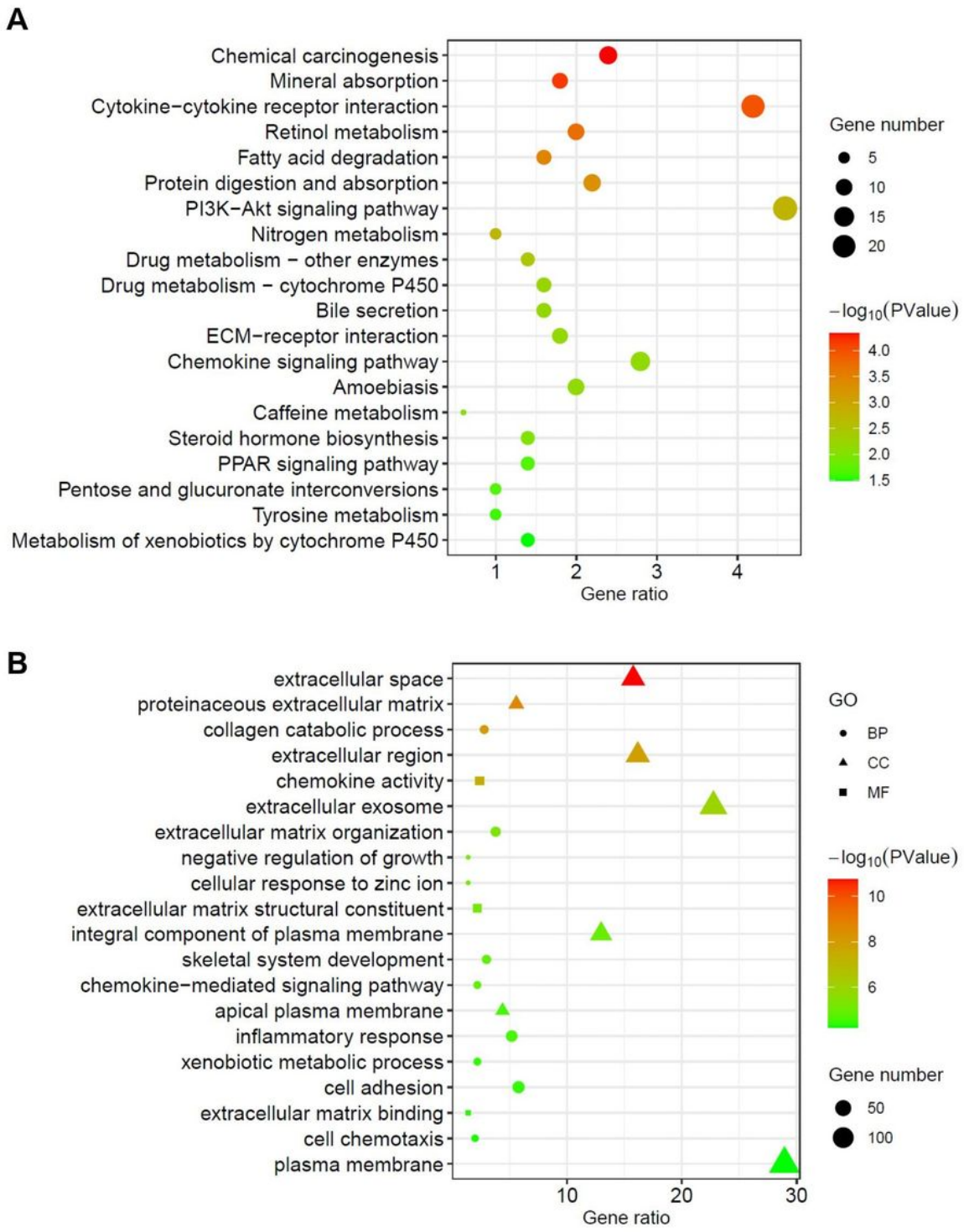


Figure 2

KEGG pathway enrichment and GO annotation analysis of DEGs. (A)Top 20 of the enriched KEGG pathways. (B) Top 20 of the enriched GO terms.

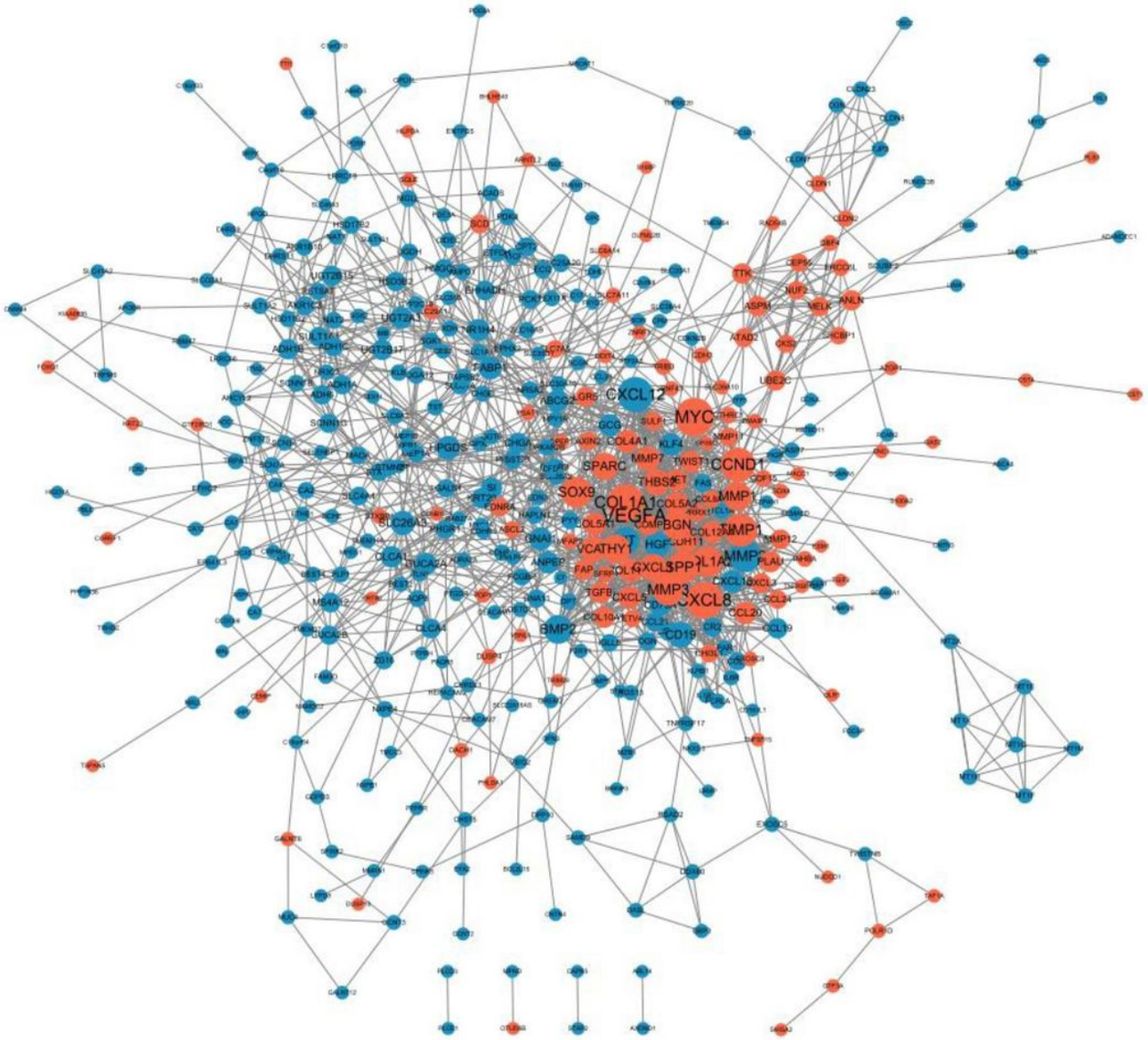
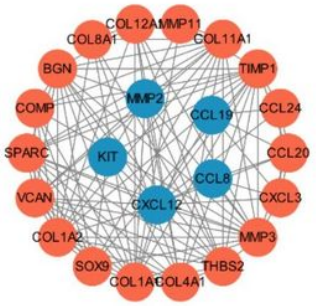


Figure 3

PPI network analysis of the DEGs. The network was made up of 494 nodes and 1586 edges. Red node represents high up-regulated DEGs, blue node represents down-regulated DEGs.

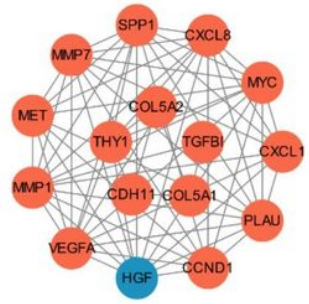
A



D

Term	Count	P-Value	FDR	Genes
ECM-receptor interaction	6	2.17E-06	5.86E-05	COL1A1, COMP, COL1A2, COL4A1, COL11A1, THBS2
Protein digestion and absorption	5	6.67E-05	7.93E-04	COL1A1, COL1A2, COL4A1, COL11A1, COL12A1
Chemokine signaling pathway	6	8.81E-05	7.93E-04	CCL24, CXCL12, CCL8, CCL20, CCL19, CXCL3

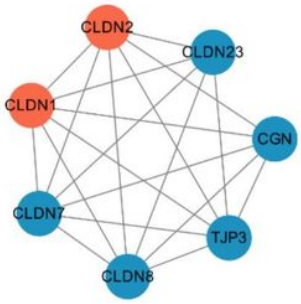
B



E

Term	Count	P-Value	FDR	Genes
Bladder cancer	5	7.47E-07	2.02E-05	CXCL8, CCND1, MMP1, MYC, VEGFA
Focal adhesion	7	9.66E-07	2.02E-05	CCND1, COL5A1, HGF, COL5A2, SPP1, MET, VEGFA
PI3K-Akt signaling pathway	8	9.93E-07	2.02E-05	CCND1, COL5A1, MYC, HGF, COL5A2, SPP1, MET, VEGFA

C



F

Term	Count	P-Value	FDR	Genes
Tight junction	7	3.44E-12	6.88E-12	CLDN8, CLDN23, CLDN7, CGN, CLDN2, CLDN1, TJP3
Leukocyte transendothelial migration	5	1.08E-06	1.08E-06	CLDN8, CLDN23, CLDN7, CLDN2, CLDN1
Hepatitis C	5	1.94E-06	1.94E-06	CLDN8, CLDN23, CLDN7, CLDN2, CLDN1

Figure 4

Top 3 significant modules were screened with MOCDE plugin of Cytoscape. Red node represents high up-regulated DEGs, blue node represents down-regulated DEGs. (A) Module 1, (D) the enriched pathways of module 1. (B) Module 2, (E) the enriched pathways of module 2. (C) Module 3, (F) the enriched pathways of module 3.

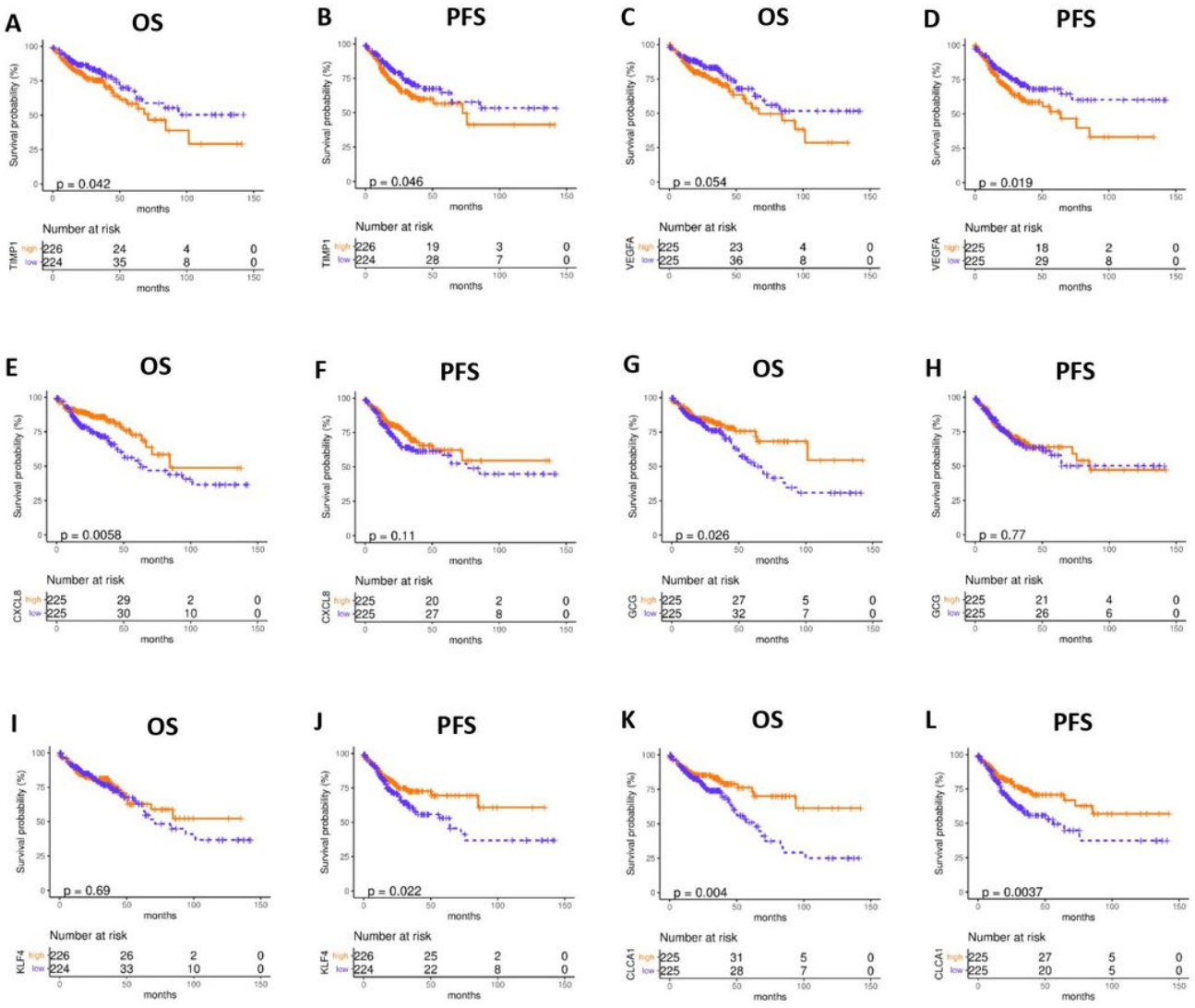


Figure 5

Kaplan Meier survival analysis of the hub genes using TCGA data. Survivorship curves of overall survival and progression free survival analysis with the 6 significant hub genes, includingTIMP1(A)(B), VEGFA(C)(D), KLF4(E)(F), GCG(G)(H), CXCL8(I)(J), CLCA1(K)(L). OS: overall survival; PFS: progression free survival.

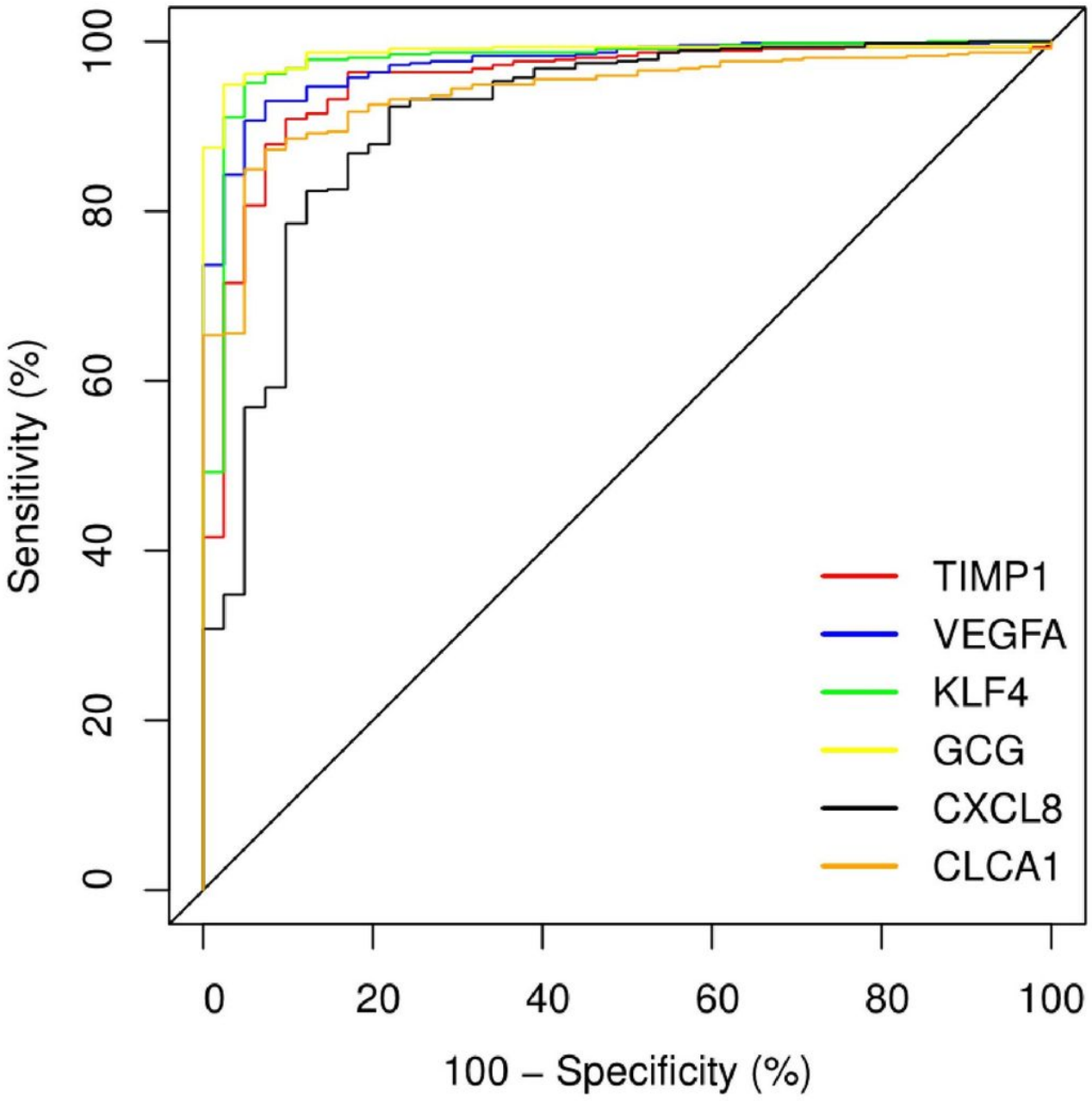


Figure 6

Diagnostic value assessment using receiver operating characteristic analysis for the prognostic genes of TIMP1, VEGFA, CXCL8, KLF4, GCG, and CLCA1.

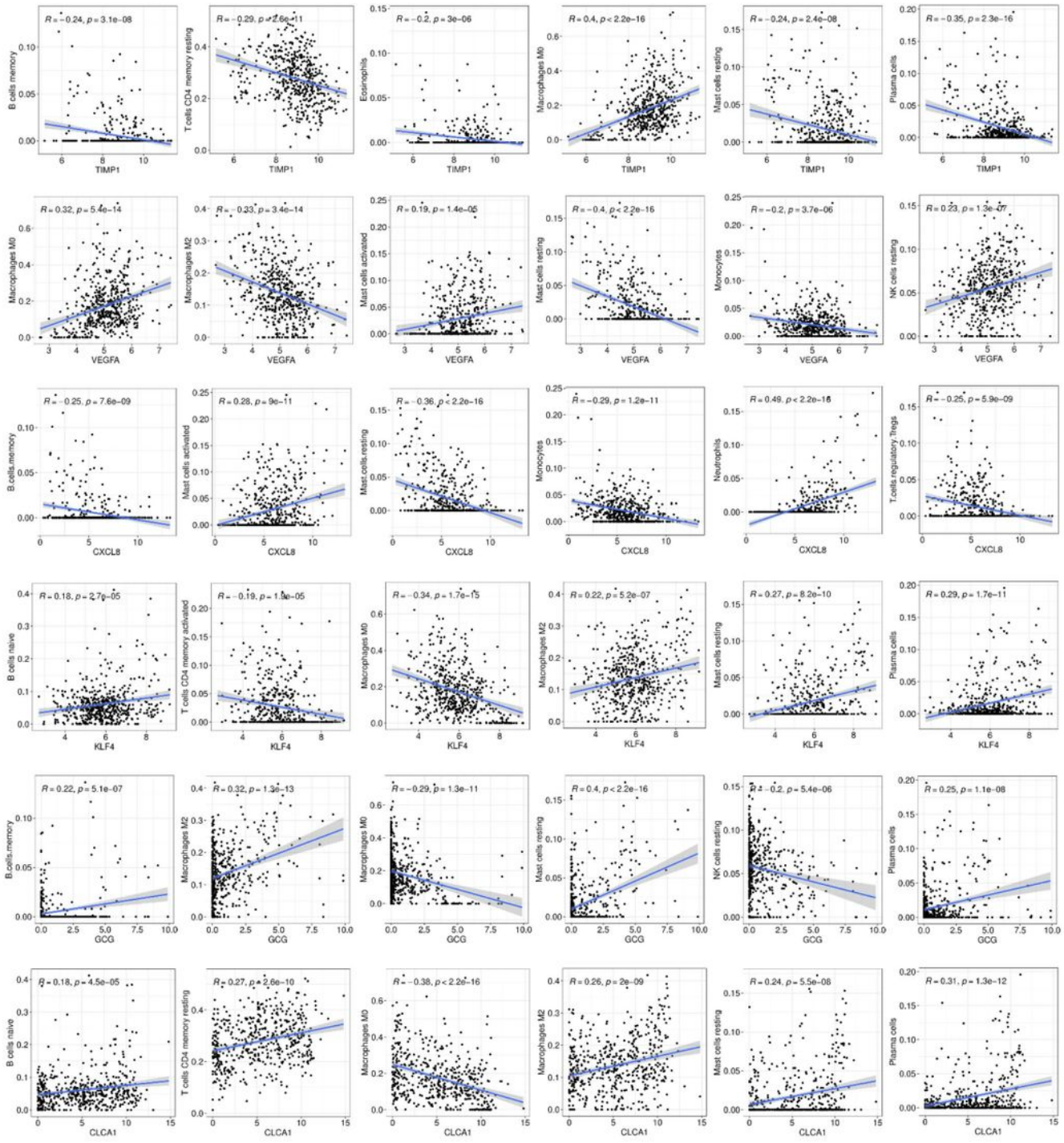


Figure 7

Correlation between immune cell infiltration and the prognostic genes of TIMP1, VEGFA, CXCL8, KLF4, GCG, and CLCA1.

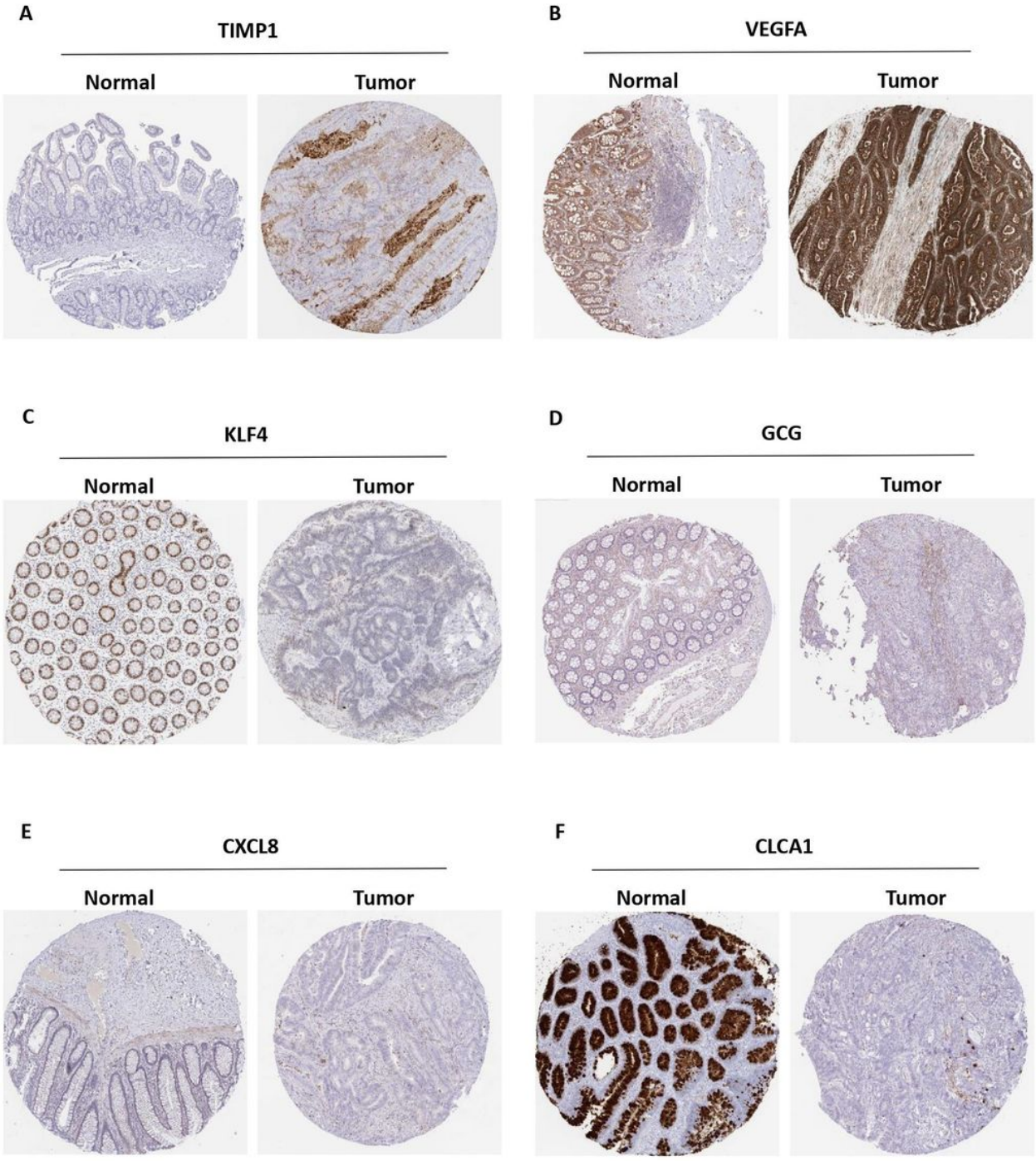


Figure 8

Expression of the 6 prognostic genes with immunochemistry assay in normal and colon cancer tissue using HPA database. (A)TIMP1, (B)VEGFA, (C)KLF4, (D) GCG, (E)CXCL8, (F)CLCA1 in tissue microarray of normal and cancer tissue.