BIOENGINEERED
2021, VOL. 12, NO. 1, 2820–2835
https://doi.org/10.1080/21655979.2021.1940030

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

Constructing a new prognostic signature of gastric cancer based on multiple data sets

Liqiang Zhou, Hao Lu, Fei Zeng, Qi Zhou, Shihao Li, You Wu, Yiwu Yuan, and Lin Xin

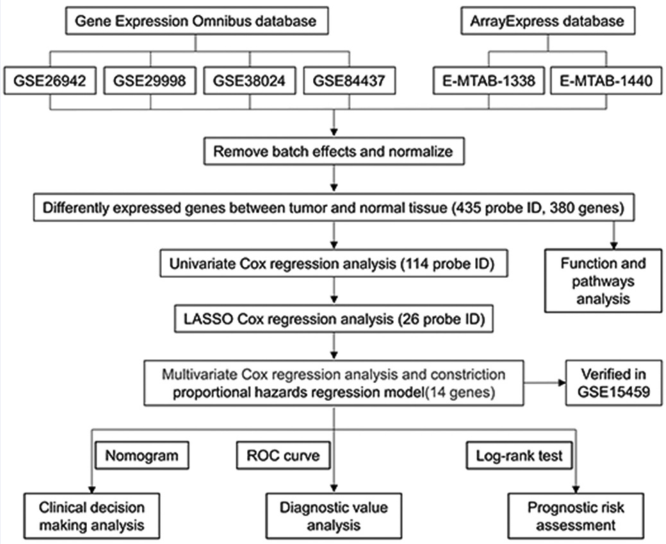
Department of General Surgery, The Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, P.R China

ABSTRACT

In order to explore new prediction methods and key genes for gastric cancer. Firstly, we downloaded the 6 original sequencing data of gastric cancer on the Illumina HumanHT-12 platform from Array Expression and Gene Expression Omnibus, and used bioinformatics methods to identify 109 up-regulated genes and 271 down-regulated genes. Further, we performed univariate Cox regression analysis of prognostic-related genes, then used Lasso regression to remove collinearity, and finally used multivariate Cox regression to analyze independent prognostic genes (MT1M, AKR1C2, HEYL, KLK11, EEF1A2, MMP7, THBS1, KRT17, RPESP, CMTM4, UGT2B17, CGNL1, TNFRSF17, REG1A). Based on these, we constructed a prognostic risk proportion signature, and found that patients with high-risk gastric cancer have a high degree of malignancy. Subsequently, we used the GSE15459 data set to verify the signature. By calculating the area under the recipient operator characteristic curve of 5-year survival rate, the test set and verification set are 0.739 and 0.681, respectively, suggesting that the prognostic signature has a moderate prognostic ability. The nomogram is used to visualize the prognostic sig-nature, and the calibration curve verification showed that the prediction accuracy is higher. Finally, we verified the expression and prognosis of the hub gene, and suggested that HEYL, MMP7, THBS1, and KRT17 may be potential prognostic biomarkers.

ARTICLE HISTORY
Received 30 March 2021
Revised 28 May 2021
Accepted 29 May 2021

KEYWORDS
Gastric cancer; prognostic signature; bioinformatics; nomogram


Gene Expression Omnibus database
ArrayExpress database
GSE26942
GSE29998
GSE38024
GSE84437
E-MTAB-1338
E-MTAB-1440
Remove batch effects and normalize
Differently expressed genes between tumor and normal tissue (435 probe ID, 380 genes)
Univariate Cox regression analysis (114 probe ID)
Function and pathways analysis
LASSO Cox regression analysis (26 probe ID)
Multivariate Cox regression analysis and constriction proportional hazards regression model(14 genes)
Verified in GSE15459
Nomogram
ROC curve
Log-rank test
Clinical decision making analysis
Diagnostic value analysis
Prognostic risk assessment

Introduction

Gastric cancer (GC) is one of the most common tumors in the world, and many factors are involved in its pathogenesis, such as low fruit and vegetables, high salt intake, Epstein-Barr virus and Helicobacter pylori infection [1]. The latest statistics showed that globally, GC deaths

CONTACT Lin Xin xlyxbs@126.com Department of General Surgery, The Second Affiliated Hospital of Nanchang University, Nanchang Jiangxi, 330006, P.R China

[#]These authors contributed equally to this work.

Supplemental data for this article can be accessed here.

reached 770,000, and new cases reached 1.09 million [2]. There are usually no obvious clinical symptoms in early gastric cancer, the patient was already with advanced gastric cancer at the time of treatment. In China, due to the extensive metastasis of advanced gastric cancer and strong drug resistance, the 5-year survival rate is less than 30%[3]. However, the diagnosis method of gastric cancer has the disadvantage of not being able to detect and diagnose early. Traditional molecular biomarkers such as CEA and CA199 lack specificity and cannot detect early gastric cancer. Although there are a variety of new diagnostic techniques and new treatment methods, they have not yet been widely used [4,5]. Therefore, it is great significant to find new and effective biomarkers on the basis of removing individual differences.

At present, sequencing technology can analyze the potential changes in the occurrence and development of diseases from the entire genome level, and has become one of the important means to find the mechanism of disease occurrence and development. However, due to the extensive heterogeneity of tumors, different patients show different specificities and prognosis. Therefore, if a large sample analysis can be used to obtain key candidate genes related to the prognosis of tumors, it will lay a solid foundation for tumor prevention and treatment. In addition, reports indicated that the combination of multiple prognostic factors is stronger than a single prognostic factor, and has more accurate performance [6,7].

With the development and improvement of public databases, the use of databases to mine prognostic genes and construct prognostic models for different tumors can accurately predict the survival of tumor patients, which is a promising method [8–10]. Compared with traditional tumor markers, gene-based prognostic models can provide more accurate prediction capabilities. The purpose of our research is to construct a hazard ratio model to predict the survival rate of gastric cancer patients. On this basis, we integrated and analyzed 6 types of gastric cancer sequencing chips on the same platform, and obtained 14 central genes through systematic statistical and bioinformatics methods to construct a prognostic signature and analyze them through external data sets. Since our prediction model directly reflects the progression of the tumor at the gene level, it can provide more accurate survival predictions for gastric cancer and more personalized treatment methods based on the gene level, which ultimately improve the survival rate of gastric cancer patients.

Materials and methods

Data collection and data processing

We downloaded Illumina HumanHT-12 V3 and Illumina HumanHT-12 V4.0 platform gastric cancer tissue microarray raw data from GEO (https://www.ncbi.nlm.nih.gov/gds/) and ArrayExpression (https://www.ebi.ac.uk/arrayexpress/), the microchip data set included GSE26942, GSE29998, GSE38042, GSE84437, E-MTAB-1338, E-MTAB-1440. We first used the 'lumi' R package to extract the expression-based data of each microarray [11]. Subsequently, the 'sva' R package normalized each microarray data, removed batch effects, and combined 6 microarray data for analysis [12]. Using the 'limma' R package to analyze differentially expressed genes for the combined data set [13], the screening threshold is $|\text{logFC}|>1$, $P < 0.05$, and the difference genes obtained were processed in the next step.

Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) enrichment analysis

GO annotation project has carried out a consistent description of gene functions, developed a controllable vocabulary, and has no species specificity. It includes Cellular Component (CC), Molecular Function (MF) and Biological Process (BP). KEGG is a comprehensive database that integrates information on genomes, chemistry, and system functions. We used the 'org.Hs.eg.db' R package to perform enrichment analysis on these differential genes, and set the filtering conditions to $P < 0.05$ and $FDR<0.05$ [14]. The obtained items are visualized using the 'enrichplot' and 'ggplot2' R packages. In order to systematically understand the potential pathways of these differential genes regulating gastric cancer.

Construction of prognosis model

In order to screen out the differential genes related to overall survival, use the 'survival analysis' package to perform univariate Cox regression analysis, and screen out prognostic-related genes with $P < 0.05$. In addition, we used Lasso regression analysis to screen out genes related to the prognosis of gastric cancer that are more significantly related to the prognosis of gastric cancer. According to the results of lasso regression analysis, the multivariate Cox regression analysis was used to find the independent predictive gastric cancer hub gene and construct a risk ratio model, respectively, to calculate the HR value, 95% confidence interval of HR and the corresponding regression coefficient (β). We calculated the risk score of each gastric cancer patient based on the model and the hub gene expression level (Exp), the formula is as follows: Risk = β1*Exp1+ β2*Exp2 + βi*Expi.

Therefore, gastric cancer patients were divided into high-risk groups and low-risk groups. Kaplan-Meier was used to draw survival curves and Log-Rank test was used to analyze the difference in OS between the two groups of patients. The 'Survival ROC' package was used to draw a 5-year ROC curve and calculate the AUC value to evaluate the predictive ability of the predictive model [15]. Univariate and multiple Cox regression analysis were used to evaluate the risk score and clinicopathological characteristics of HR and P values to determine independent prognostic factors for patients with gastric cancer. In addition, we also use the Chi-square test to analyze the relationship between high and low risk and the expression of each gene and each clinicopathological feature.

Gene set enrichment analyses (GSEA)

GSEA can fully enrich the biological differences between samples of different classifications as a - whole[16]. In this study, GSEA4.0 based on the molecular signature database (MSigDB) was used, hallmark7.1 was used as the comparison gene set, and the Number of permutations was set to 1000 for enrichment. Use NES>1 and FDR<0.001 as the screening conditions to identify the similarities and differences between the occurrence and development of gastric cancer in the high-risk group and the low-risk group.

Construction and verification of nomogram

We drew a nomogram according to the prognostic risk model, obtained the corresponding score by analyzing the hub gene expression level, and added the points of all hub genes to obtain the corresponding total points [17]. By drawing a vertical line, we can predict the probability of survival for patients with gastric cancer in 1 to 5 years. In addition, in order to test the prediction ability of the nomogram, we draw a 5-year calibration curve by analyzing the survival probabilities of the predicted value and the actual value at the quartile of all gastric cancer patients. If the actual value is close to the predicted value, the nomogram has good predictive performance.

Hub gene expression and prognostic verification

The human protein atlas (HPA, https://www.proteinatlas.org/) database contains the tissue and cell distribution information of 24,000 human proteins [18]. This database was used to verify the hub genes. Kaplan-Meier Plotter website (http://kmplot.com/) includes the expression and survival information of the gastric cancer GPL570 platform chip in the GEO database, which can quickly analyze the relationship between genes and overall prognosis[19]. We use this tool to verify the prognostic ability of hub genes.

Results

Gastric cancer is a common disease with high mortality worldwide. The clinical symptoms of early gastric cancer are not obvious, resulting in poor prognosis and high recurrence rate. Therefore, some technical means are needed to improve the accuracy of diagnosis and prognosis. Because gastric cancer has a high degree of heterogeneity, it is more suitable to study gastric cancer at the genetic level. Through the analysis of the database, we have obtained 14 core genes to construct the prediction model of gastric cancer, and the verification of external data showed that it has good predictive performance. We finally constructed a nomogram to visually represent the survival rate of 1–5 years, which helps individualized and precise treatment. For the obtained hub gene, the expression and prognosis were verified in the HPA database and Kaplan-Meier Plotter. It was finally confirmed that HEYL, MMP7, THBS1, and KRT17 may be potential biomarkers of gastric cancer.

Identify differentially expressed genes (DEGs)

Our entire research process has been shown in Figure 1. We downloaded 4 data sets (GSE26942, GSE29998, GSE38024, GSE84437) from the GEO database and 2 data sets (E-MTAB-1338, E-MTAB-1440) on ArrayExpress. We extracted the original expression data and removed the batch effect. After standardization, we obtained a total of 118 normal gastric tissue samples and 827 gastric cancer samples. After using the 'limma' package analysis, a total of 380 genes meets the screening conditions, of which 109 genes are up-regulated, and 271 genes are down-regulated **Table S1**). We used R to draw heat maps and volcano maps of these differential genes (Figure 2 (a, b)).

GO and KEGG function enrichment analysis

In order to study the mechanism of differential genes affecting gastric cancer, we used GO and KEGG to further analyze the potential mechanisms of these DEGs in regulating gastric cancer. Among them, GO enrichment analysis showed that BP is enriched in extracellular matrix organization, extracellular structure organization, collagen fibril organization, detoxification, digestion, detoxification of copper ion, stress response to copper ion, response to toxic substance, detoxification of inorganic compound, stress response to metal ion; CC is enriched in

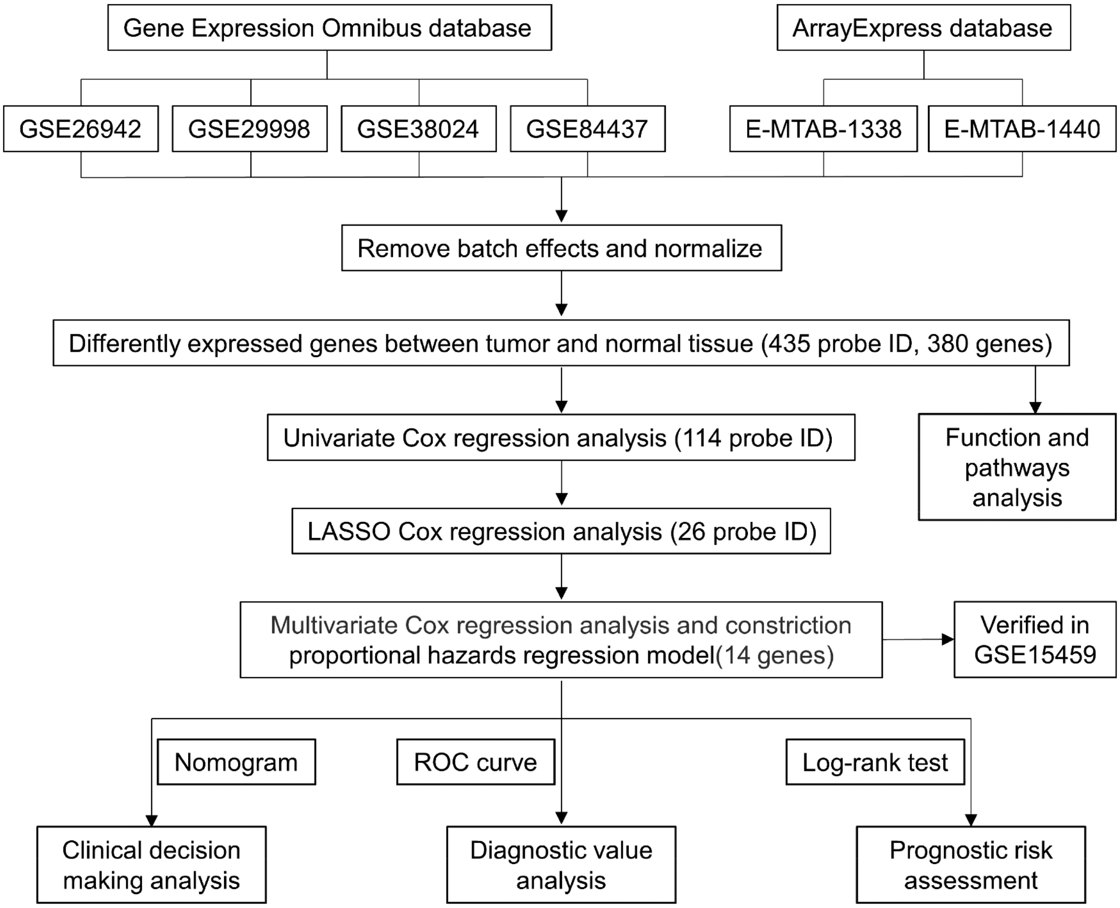

Gene Expression Omnibus database
ArrayExpress database
GSE26942
GSE29998
GSE38024
GSE84437
E-MTAB-1338
E-MTAB-1440
Remove batch effects and normalize
Differently expressed genes between tumor and normal tissue (435 probe ID, 380 genes)
Univariate Cox regression analysis (114 probe ID)
Function and pathways analysis
LASSO Cox regression analysis (26 probe ID)
Multivariate Cox regression analysis and constriction proportional hazards regression model(14 genes)
Verified in GSE15459
Nomogram
ROC curve
Log-rank test
Clinical decision making analysis
Diagnostic value analysis
Prognostic risk assessment

Figure 1. Flow chart of this research.

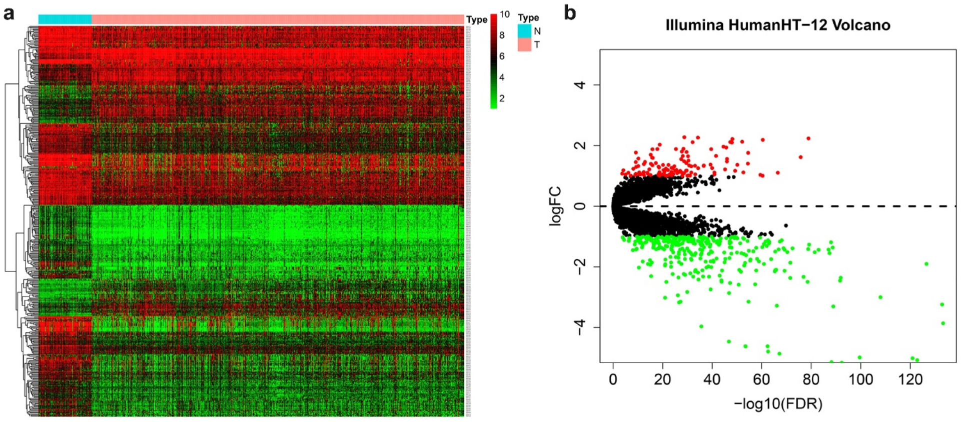


Figure 2. The differentially expressed genes in gastric cancer.
(A) Heat map; (B) Volcano plot.

collagen-containing extracellular matrix, complex of collagen trimers, collagen trimer, endoplasmic reticulum lumen, apical part of cell fibrillar collagen trimer, banded collagen fibril, apical plasma membrane, basolateral plasma membrane, basement membrane; MF is enriched in extracellular matrix structural constituent, extracellular matrix structural constituent conferring tensile strength, glycosaminoglycan binding, aldo-keto reductase (NADP) activity, oxidoreductase activity, platelet-derived growth factor binding, oxidoreductase activity, alcohol dehydrogenase (NADP+) activity, heparin binding (Figure 3(a-c)). The results of KEGG indicated that the DEGs are enriched in Chemical carcinogenesis, Glycolysis/Gluconeogenesis, ECM-receptor interaction, etc (Figure 3d).

Identify prognostic-related hub genes

By extracting clinical data from each data set, we obtained a total of 635 gastric cancer samples with clinical data. Using univariate Cox regression to analyze the relationship between 435 differential gene probes and prognosis, 114 probes related to prognosis were subsequently obtained, of which 77 probes with HR>1 and 37 probes with HR<1. In addition, we used lasso regression to perform further 114 prognostic-related probes, and we obtained 26 gene probes significantly related to prognosis (Figure 4(a, b)). We also used multivariate Cox regression analysis to test the results obtained in the previous step to find independent prognostic factors among the 26 probe IDs. The results showed (Figure 4c) that a total of 14 hub genes are independent prognostic factors for

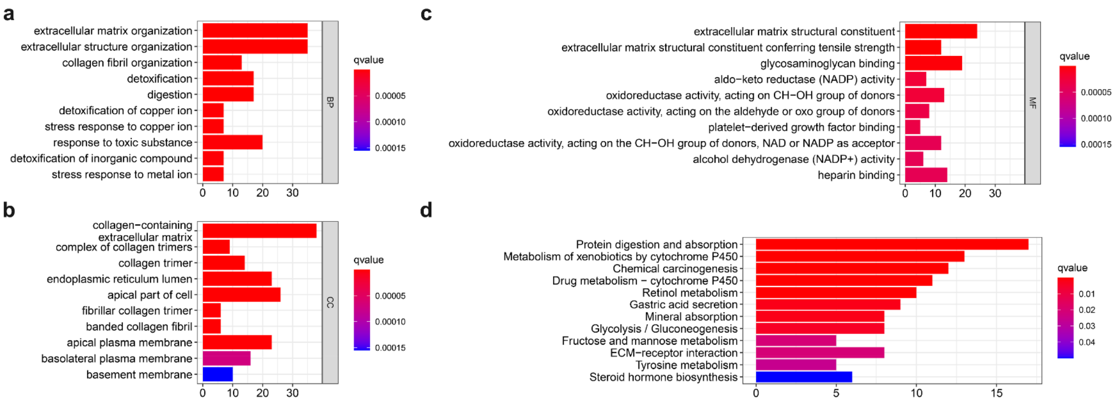


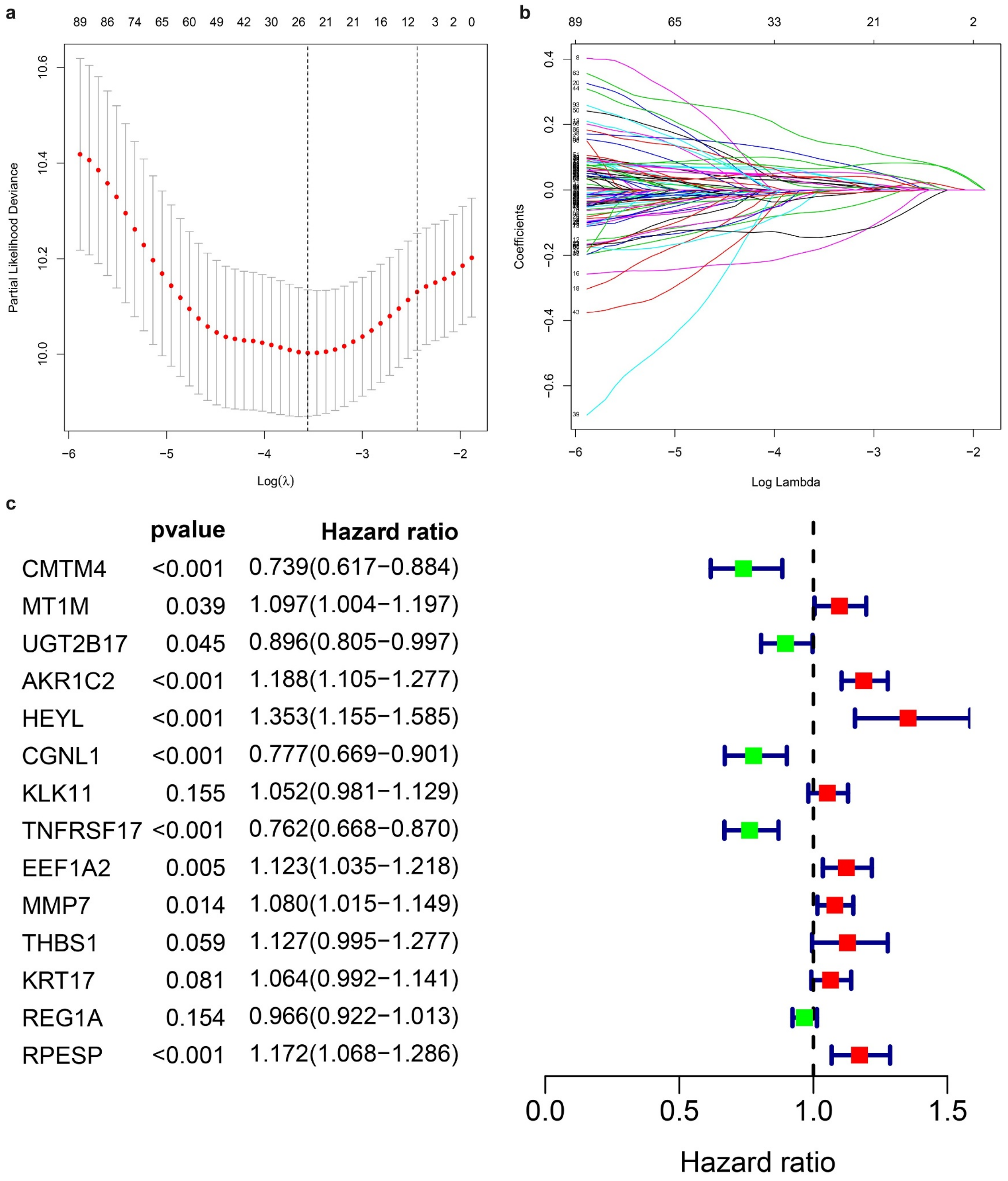
Figure 3. Functional enrichment analysis of the DEGs.
(A) The biological process enrichment results of GO with different genes. (B) The cellular components enrichment results of GO with different genes. (C) The molecular function enrichment results of GO with different genes. (D) KEGG enrichment results of differential genes.

gastric cancer patients. Among them, those with HR>1 include MT1M, AKR1C2, HEYL, KLK11, EEF1A2, MMP7, THBS1, KRT17, RPESP, HR< 1 Including CMTM4, UGT2B17, CGNL1, TNFRSF17, REG1A.

Prognostic model construction and verification

Based on multivariate Cox regression, we constructed a 14-gene risk ratio model to predict the prognosis of gastric cancer patients. By assigning the corresponding regression coefficient to each gene, the risk-score of each gastric cancer patient can be calculated. The risk score formula of this model is as follows:

Risk score = (−0.303 * ExpCMTM4) + (0.092 * ExpMT1M)+ (−0.110 * ExpUGT2B17) + (0.172 * ExpAKR1C2)+ (0.302 * ExpHEYL)+ (−0.253 * ExpCGNL1)+ (−0.051 * ExpKLK11)+ (−0.272 * ExpTNFRSF17)+ (−0.116 *



	pvalue	Hazard ratio
CMTM4	<0.001	0.739(0.617−0.884)
MT1M	0.039	1.097(1.004−1.197)
UGT2B17	0.045	0.896(0.805−0.997)
AKR1C2	<0.001	1.188(1.105−1.277)
HEYL	<0.001	1.353(1.155−1.585)
CGNL1	<0.001	0.777(0.669−0.901)
KLK11	0.155	1.052(0.981−1.129)
TNFRSF17	<0.001	0.762(0.668−0.870)
EEF1A2	0.005	1.123(1.035−1.218)
MMP7	0.014	1.080(1.015−1.149)
THBS1	0.059	1.127(0.995−1.277)
KRT17	0.081	1.064(0.992−1.141)
REG1A	0.154	0.966(0.922−1.013)
RPESP	<0.001	1.172(1.068−1.286)

Figure 4. Prognosis-related gene screening.

(A, B) Selecting the best parameters for gastric cancer in LASSO regression analysis, and 26 gene probes significantly related to prognosis; (C) Multivariate Cox regression analysis to get prognosis related 14 genes.

ExpEEF1A2)+ (−0.077 * ExpMMP7)+ (- 0.120 * ExpTHBS1)+ (−0.062 * ExpKRT17)+ (−0.034 * ExpREG1A)+ (−0.159 * ExpRPESP).

According to the obtained risk scores, the median is divided into high and low risk groups. The dot plot of survival status revealed (Figure 5a) that the number of deaths of gastric cancer patients in the high-risk group is greater and their survival time is less than that in the low-risk group. We drew the K-M curve (Figure 5b) and used log-Rank to test. The results suggested that the high-risk group has a worse survival rate than the low-risk group ($P < 0.05$). Then, we drew a 5-year ROC curve to judge the predictive ability of the prognostic signature, and calculated the AUC to be 0.739, showed that our prediction model has a medium ability to predict (Figure 5c). In addition, we also combined the clinicopathological characteristics to analyze whether the prognostic signature is an independent prognostic factor. Univariate Cox regression analysis showed Age [HR = 1.020, 95%CI (1.007–1.032)), P = 0.002], N stage [HR = 1.676, 95%CI (1.429–1.967), P < 0.001], T stage [HR = 1.740, 95%CI (1.378–2.198), P < 0.001], risk score [HR = 1.700, 95%CI (1.539–1.878), P < 0.001] are related to prognosis (Figure 5d). Multivariate Cox regression analysis showed Age [HR = 1.020, 95%CI (1.008–1.032), P = 0.001], N stage [HR = 1.476, 95%CI (1.256–1.734), P < 0.001], T stage [HR = 1.408, 95%CI (1.100–1.801), P = 0.006], risk score [HR = 1.619, 95%CI (1.453–1.804), P < 0.001] are independent prognostic factors for gastric cancer patients and showed that the risk score has the best predictive ability (Figure 5e).

We downloaded the original data and clinical information of the GSE15459 dataset from the GEO database and used the RMA algorithm to extract the expression of each gene to verify the 14-gene prediction model. The survival status chart (Figure 6a) showed that the high-risk group (n = 79) has a higher number of deaths and a lower survival time than the low-risk group (n = 113). The K-M curve (Figure 6b) confirmed that the high-risk group has a lower survival rate than the first-risk group. We also drew a 5-year ROC curve, and the AUC of the GS415459 cohort is 0.681 (Figure 6c). Combined with clinical case parameters, univariate Cox regression analysis indicated that AJCC stage [HR = 2.790, 95%CI (2.141–3.635), P < 0.001] and risk score [HR = 1.849, 95%CI (1.429–2.392), P < 0.001] has prognostic ability (Figure 6d). Multivariate Cox regression analysis showed that AJCC stage [HR = 3.050, 95%CI (2.292–4.059), P < 0.001] and risk score [HR = 1.953, 95%CI (1.476–2.584), P = 0.006] are independent prognostic factors (Figure 6e). The results of the analysis in the validation data set are similar with the model we built.

Bioinformatics analysis

We used chi-square test to explore whether there is a relationship between each clinicopathological feature and high and low risk groups. The results showed that the risk grouping is related to T stage (Figure 7a). The high-risk group gastric cancer patients were mostly in T3 and T4 stages, and more deaths. We also analyzed the relationship between 14 hub genes and clinicopathological characteristics. Analysis suggested (Figure 7b) that almost all hub genes are related to risk scores. In addition, CMTM4, MT1M, KLK11, RPESP, HEYL, THBS1 are related to T stage. To identify the difference between high and low risk groups, we used GSEA for analysis. The enrichment indicated that the samples of the high-risk group were enriched in Epithelial Mesenchymal Transition, Myogenesis, NF-KB /TNF-α via, TGF-β signal, Goagulation, Apical junction, Angiogenesis, Hypoxia, Hedgehog signal, UV response down (Figure 8).

Nomogram construction and verification

Based on the constructed 14-gene prognostic model, we drew nomograms for better helping clinicians to make precise treatment decisions, thereby improving the survival time and quality of life of patients with gastric cancer. By detecting the expression levels of 14 hub genes and assigning corresponding scores, the 1–5 years survival rate of gastric cancer patients can be judged after adding the total scores (Figure 9a). In order to test the predictive ability of the nomogram, we tested the actual survival rate and predicted survival rate of 635 gastric cancer samples respectively and drew a 5-year calibration curve. The results showed

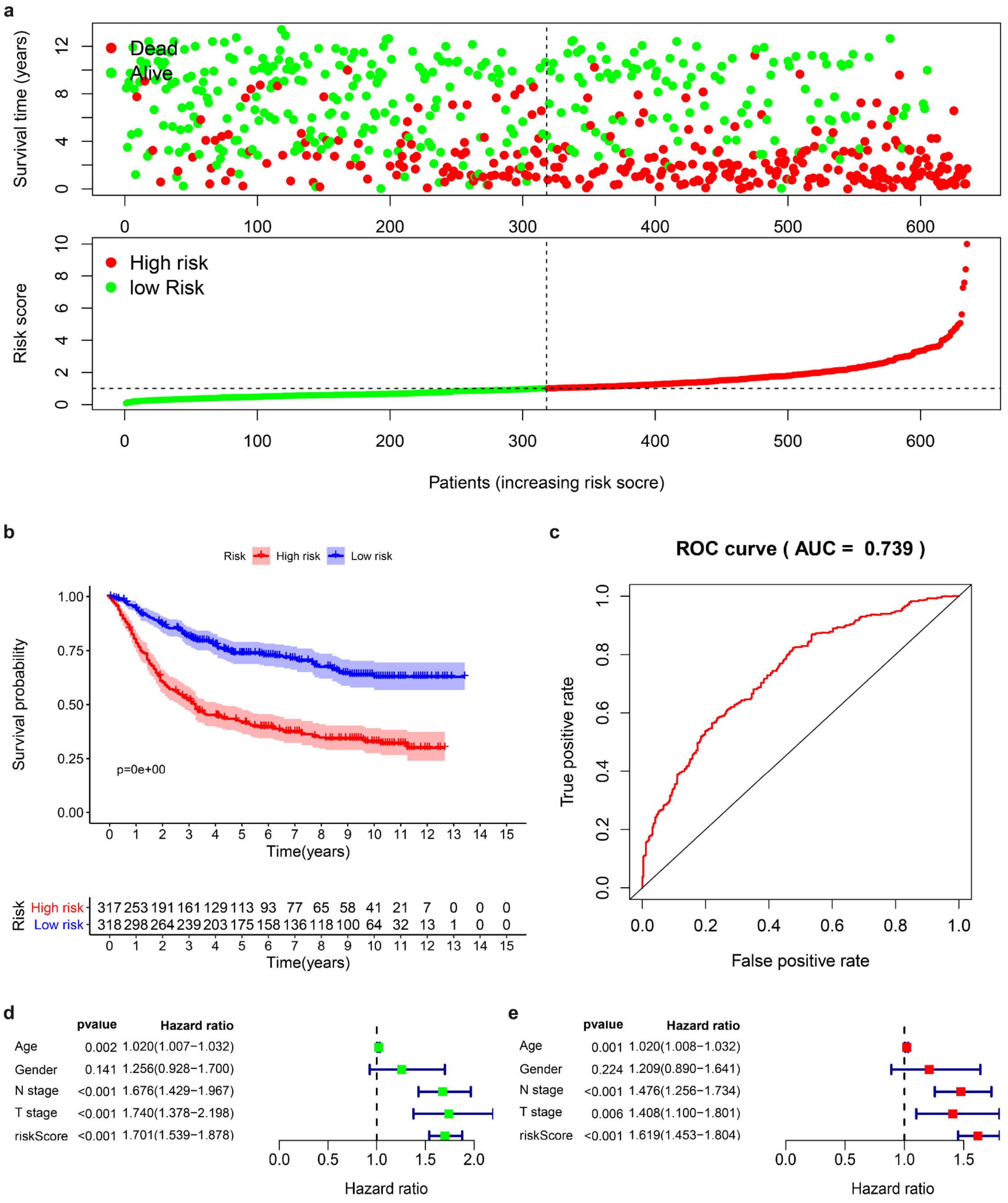


Figure 5. Prognostic analysis of 14-genes signature in the train cohort.

(A) Scatterplots of GC patients with different survival status in training group; Risk score distribution of GC patients with different risks (low, green; high, red) in the testing group; (B) Kaplan–Meier Survival curve of low-risk and high-risk subgroups; (C) 5-year time-dependent ROC for survival prediction models; (D) Univariate Cox regression analysis on the prognosis of clinicopathological characteristics and risk scores in patients with GC; (E) Multivariate Cox regression analysis on the prognosis of clinicopathological characteristics and risk scores in patients with GC.

that the calibration curve is almost in line with the standard line, suggesting that our connection diagram has very accurate capabilities (Figure 9b).

Hub genes expression and prognostic verification

First, we searched the HPA database for the immunohistochemical data of CMTM4,

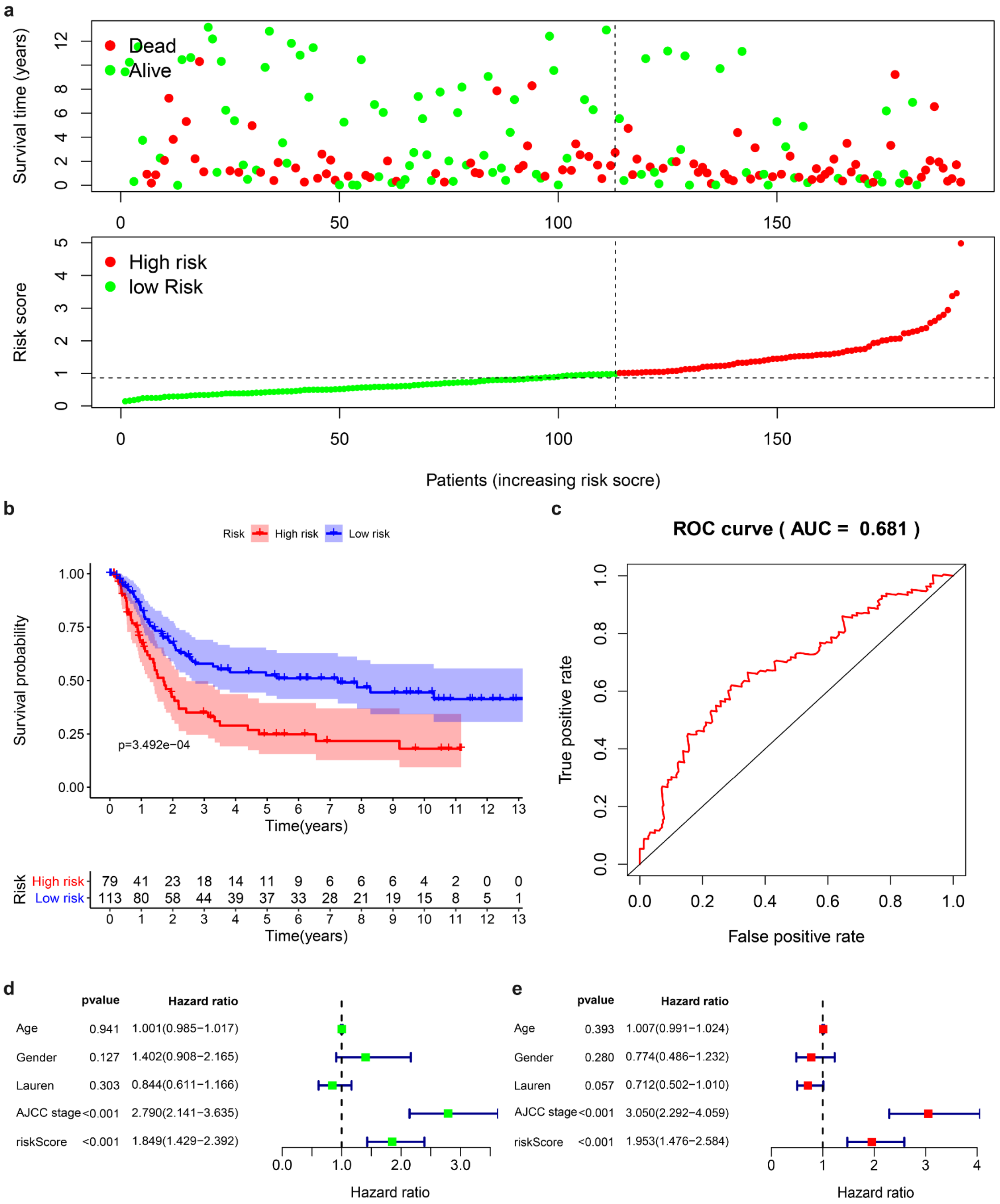


Figure 6. Prognostic analysis of 14 genes signature in the GSE15459 data set.

(A) Scatterplots of GC patients with different survival status in training group; Risk score distribution of GC patients with different risks (low, green; high, red) in the testing group; (B) Kaplan–Meier Survival curve of low-risk and high-risk subgroups; (C) 5-year time-dependent ROC for survival prediction models; (D) Univariate Cox regression analysis on the prognosis of clinicopathological characteristics and risk scores in patients with GC; (E) Multivariate Cox regression analysis on the prognosis of clinicopathological characteristics and risk scores in patients with GC.

AKR1C2, HEYL, CGNL1, KLK11, EEF1A2, MMP7, KRT17, REG1A, RPESP, and indicated that they were consistent with the mRNA level expression we analyzed (Figure 10). Secondly, for prognostic verification of the 14 hub genes, we used the Kaplan-Meier Plotter website. The results showed (Figure 11) that AKR1C2, HEYL, KLK11, EEF1A2, MMP7, THBS1, KRT17, C8ORF84 are associated with poor prognosis. The expression of CMTM4, MT1E, UGT2B17,

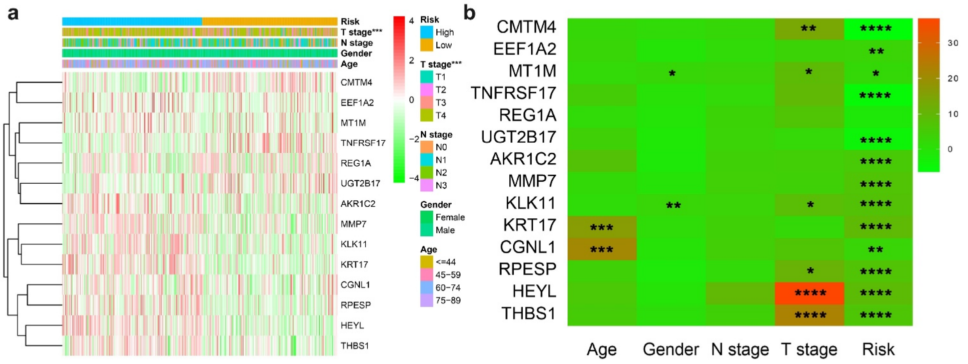


Figure 7. Risk and clinicopathological characteristics of 14 genes.

(A) The relationship between clinicopathological characteristics and risk group; (B) The relationship between 14 genes expression level and clinicopathological characteristics and risk value.

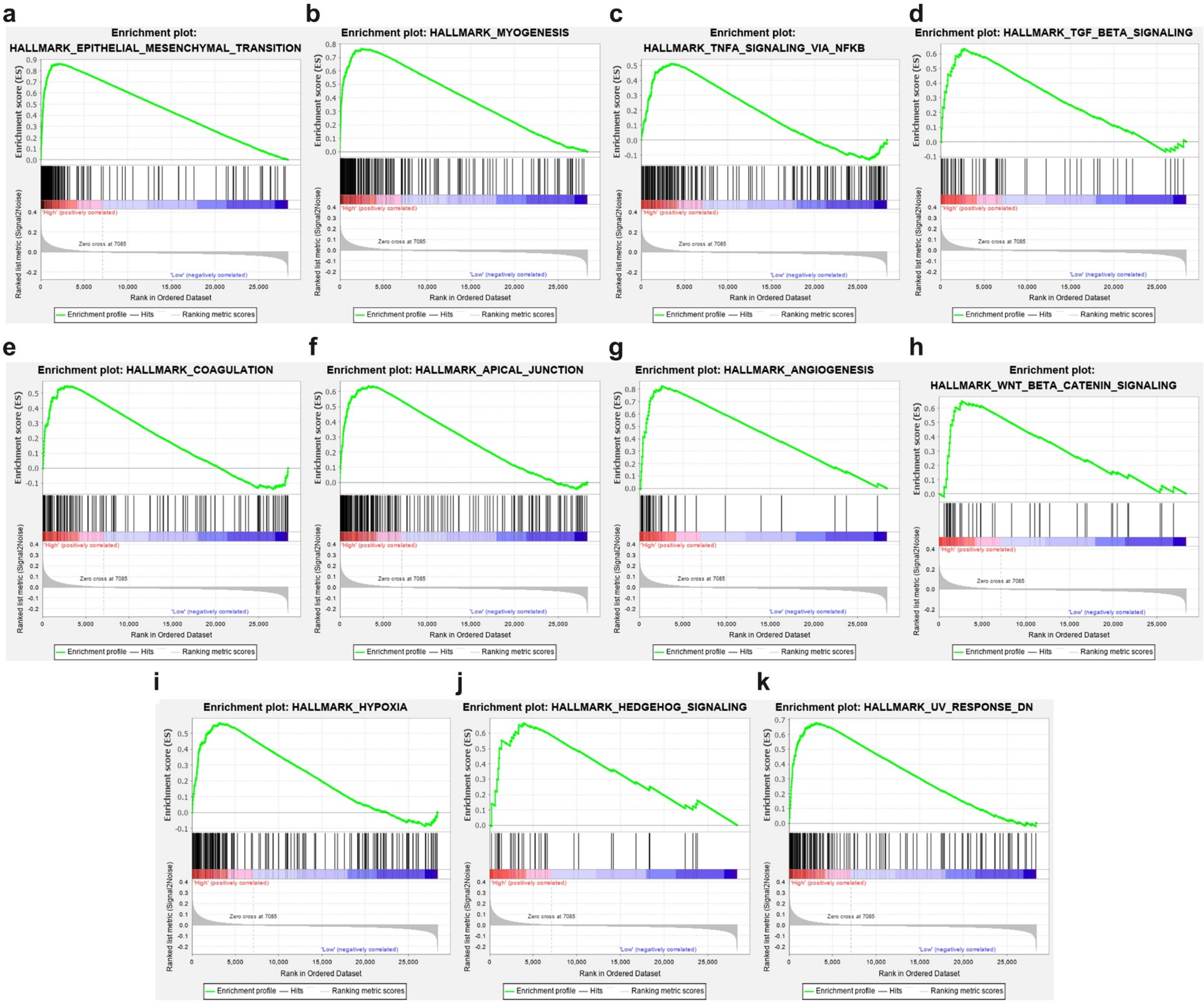


Figure 8. High-risk group conducts GSEA enrichment pathway analysis.

CGNL1, BCMA, and REG1A are associated with better prognosis. In summary, HEYL, MMP7, THBS1, and KRT17 are not only highly expressed in gastric cancer, but also associated with poor prognosis and may be potential prognostic biomarkers.

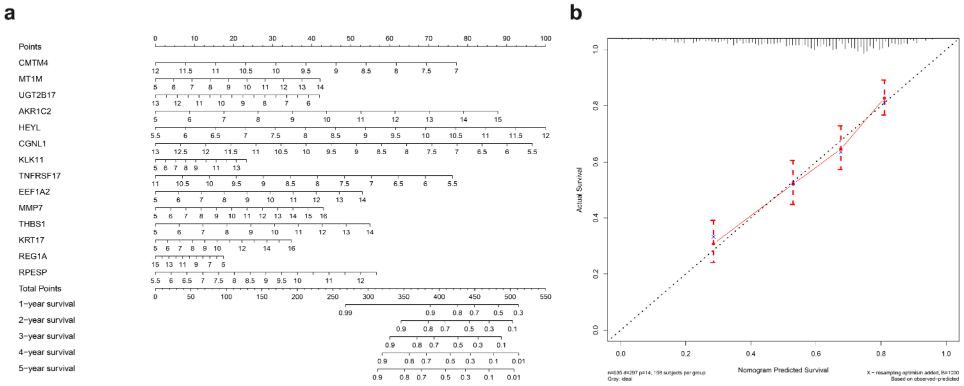


Figure 9. Establishment and validation of Nomogram (a) Nomogram for predicting 1–5 years OS of GC patients. (b) calibration chart for nomogram accuracy.

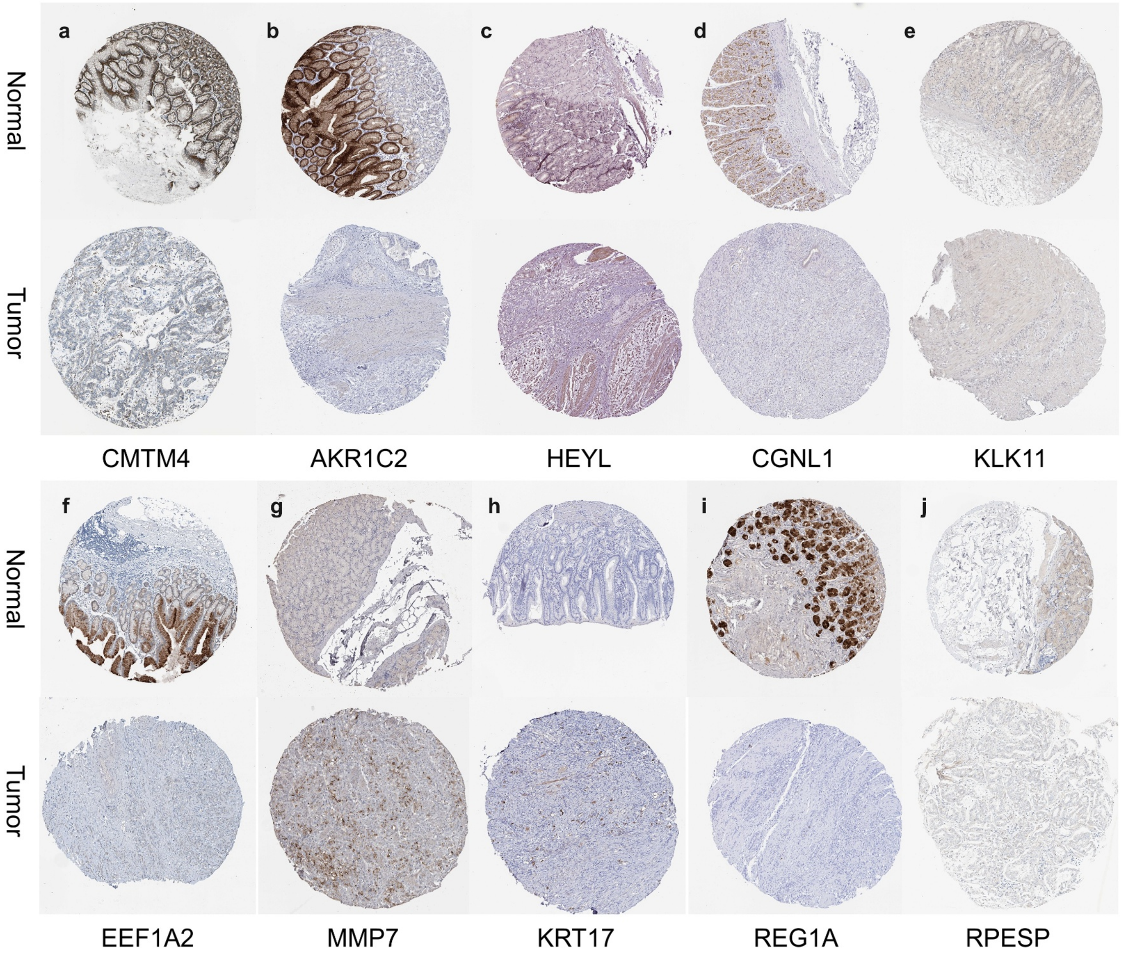


Figure 10. Verification of 14 genes expression in GC and normal gastric tissue using the HPA database.

Discussion

Gastric cancer is a highly heterogeneous malignant tumor. According to research, about 10% of gastric cancer patients showed familial aggregation, and 1–3% of gastric cancer patients will have germline mutations [20], Prognosis and treatment should be judged and selected from the genetic level. Gastric cancer of the same pathological type and stage could show different prognosis in different cases.

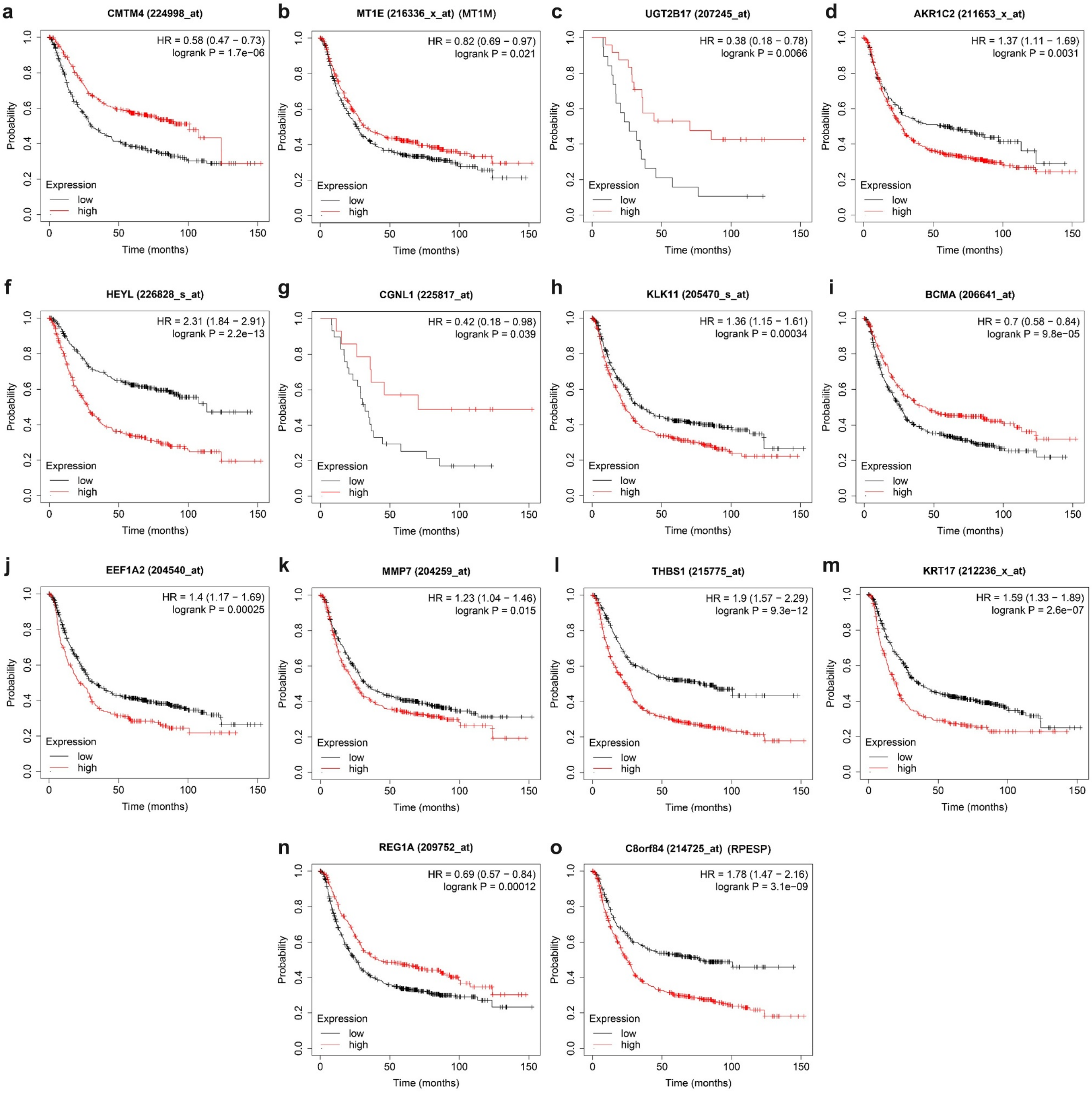


Figure 11. Validation the prognostic value of 14 genes in GC by Kaplan Meier-plotter.

Individualized treatment of gastric cancer patients at the genetic level also reflects the concept of precision medicine and improve the quality of life of patients. For example, HER2 oncogene amplification and HER2 protein overexpression occur in approximately 17–20% of gastric cancer patients, and it is more common in intestinal gastric cancer and cancers in the proximal stomach or gastroesophageal junction [21], according to HER2 gene Positive expression, the use of trastuzumab prolonged the survival rate of patients [22,23]. All the above indicated that due to the genetic instability and extensive heterogeneity of tumors, a single factor cannot accurately predict the occurrence, development and prognosis of gastric cancer.

In this study, we systematically analyzed the differential genes between gastric cancer and normal tissues, and obtained key genes after screening. GO enrich the above genes to show that they are mainly enriched in extracellular matrix organization, collagen fibril organization, etc. which means that the above genes are inseparable from the microenvironment of the tumor. Recent studies have shown that the tumor microenvironment not only significantly affects tumor growth, angiogenesis, chemotherapy resistance and immune regulation, but also plays an important role in tumor cell immune regulation, chemotherapy resistance, and recurrence and metastasis. It is also the target of emerging tumor-targeted therapeutic drugs [24]. KEGG is mainly enriched in Chemical carcinogenesis, Glycolysis/Gluconeogenesis, ECM-receptor interaction, etc. Glycolysis in tumor cells is the main way to obtain energy, which can promote the proliferation and metastasis of tumor cells [25]. The treatment of energy metabolism of tumors is also an important option. Such as, Wei et al. used the compound DT-13 to inhibit the expression of glucose transporter 1, thereby inhibiting glucose absorption and aerobic glycolysis, thereby inhibiting the proliferation of colon cancer [26].

We used systematic statistical analysis to identify independent prognostic genes and constructed a risk ratio model. According to the expression levels and statistical regression coefficients of hub genes, they are divided into low and high-risk groups. Among them, gastric cancer patients in the high-risk group have a lower survival rate and are closely related to the T stage. Combining clinicopathological information, using univariate and multivariate Cox regression analysis, the prognostic model and N stage and T stage are independent prognostic factors for patients with gastric cancer. In addition, we also calculated the 5 sticky AUC value, showing medium predictive power. The external data set GSE15459 also showed that the model has good predictive ability. Based on the prognostic signature, we also constructed a nomogram and verified its prediction accuracy with a calibration curve, indicating that it has good prediction performance.

In order to analyze the potential pathogenic mechanism of high-risk patients with gastric cancer, we conducted pathway analysis on the high-risk and low-risk groups through GSEA, and found that the high-risk group was mainly enriched in epithelial-mesenchymal transition (EMT), TGF-β, Wnt /β-catenin, NF-KB/TNF-α signals, etc. These signals are closely related to tumor progression. Studies have shown that abnormal activation of the Wnt /β-catenin signal pathway plays a vital role in the occurrence and metastasis of gastric cancer [27], and the activated Wnt /β-catenin signal pathway has been found in more than 30% of gastric cancers [28], therefore, Wnt/β-catenin signal may be a potential strategy for targeted therapy of gastric cancer. TGF-β signal transduction is a very important regulator in the human body. It can mainly regulate the growth of tissues and maintain the homeostasis of the internal environment. When this signal is out of regulation, it often leads to a series of diseases including cancer[29]. Studies in breast, lung, and pancreatic cancer have proved that TGF-beta plays a key role [30], and its mechanism may be to induce EMT to promote tumor growth and invasion. In this study, we found that multiple enrichment pathways in the high-risk group are closely related to EMT. Therefore, we speculated that the EMT pathway in gastric cancer is the confluence point of other genes and pathways and plays a vital role in the occurrence and development of the disease. In our study, the EMT pathway is also an obvious enrichment pathway, and the possible mechanism is the N6-methyladenosine modification mediated by METTL3, which regulates the effects of E-cadherin and non-coding RNA. [31,32].

Our prognostic model includes 14 genes including MT1M, AKR1C2, HEYL, KLK11, EEF1A2, MMP7, THBS1, KRT17, RPESP, CMTM4, UGT2B17, CGNL1, TNFRSF17, REG1A. Among them, HEYL, MMP7, THBS1, and KRT17 are not only highly expressed in gastric cancer, but are also independent prognostic risk factors for gastric cancer. HEYL is a member of the division-related family of transcription factors. It not only regulates the differentiation, self-renewal and proliferation of cancer cells, but also promotes tumor angiogenesis, so it plays an important role in tumor progression [33–35]. Studies have shown that the HEYL gene is significantly increased in patients with gastric cancer, usually showing a poor prognosis [36]. The mode of action of HEYL on gastric cancer has not yet been clearly studied, and further proof is needed in the future. MMP7, also known as stroma lysin, is a unique member of the matrix metalloproteinase family and plays a key role in the middle of the family. It is mainly expressed by tumor cells and is different from other family members, which is an important feature of it. Reports showed that the expression of MMP7 is associated with the poor prognosis of gastric cancer and is involved in the epithelial-mesenchymal transition of the tumor to induce metastasis [37]. The mechanism may be that MMP-7 can cleave E-cadherin in gastric cancer cells [38]. THBS1 is a secreted protein. Many studies have confirmed that it is highly expressed in gastric cancer stroma and is closely related to tumor growth and metastasis [39,40]. However, there are also reports showed that THBS1 is abnormally elevated in gastric cancer tissues, which is associated with a poor prognosis and enhances angiogenesis in gastric cancer cells [41–43]. KRT17 has been confirmed to be involved in the progression of a variety of tumors. In gastric cancer studies, it has been shown that KRT17 is closely related to tumor size, depth of invasion, lymph node metastasis, stage of tumor lymph node metastasis, vascular invasion and poor prognosis [44]. After inhibiting gastric cancer cell KRT17, it inhibits proliferation and metastasis and induces apoptosis, and its mechanism of promoting tumor progression may be mainly regulated by AKT/mTOR signal [44,45].

Conclusion

In general, we used bioinformatics methods to identify differential genes and independent prognostic genes in gastric cancer. A prognostic signature is constructed on this basis, and after multiple verifications, it reflects good predictive performance. Furthermore, we have also established a nomogram, which can effectively predict the survival rate of gastric cancer patients and will help clinicians to make accurate judgments. Although our research has problems such as lack of clinical information. However, the shortcomings are not concealed. A huge number of samples are included in the revised prognosis model, which has wide applicability.

Highlights

- 380 differential genes were screened from 118 normal and 827 gastric cancer samples.
- Fourteen independent prognostic genes in gastric cancer were identified, and a prognostic signature was constructed with good predictive performance.
- Constructed a nomogram for forecasting 1-5 years, and verified its accuracy in predicting survival.
- HEYL, MMP7, THBS1, KRT17 may be potential biomarkers of gastric cancer.

Disclosure of potential conflicts of interest

The authors declare that they have no competing interests.

Funding

National Natural Science Foundation of China, [Grant/Award Numbers: 81872480, 81760549, 81560492]; Key R&D Projects in Jiangxi Province, [Grant/Award Numbers: 20203BBG73056]; Science and Technology Research Project of Education Department of Jiangxi Province, [Grant/Award Number: GJJ180024].

Authors' contributions

Liqiang Zhou and Fei Zeng conceived, designed, analyzed the data, and Hao Lu write the manuscript. Qi Zhou, Shihao Li helped to search for some relevant papers for this research. You Wu and Yiwu Yuan generated the figures and tables. Lin Xin guided the research process and review the manuscript. All authors read and approved the final manuscript.

Notes on contributors

Liqiang ZhouM.D., Ph.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is targeted therapy of gastrointestinal tumors.

Hao LuM.D., Ph.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is the clinical treatment of gastrointestinal tumors.

Fei ZengM.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is the clinical treatment of gastrointestinal tumors.

Qi Zhou M.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is the clinical treatment of gastrointestinal tumors.

Shihao LiM.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is the clinical treatment of gastrointestinal tumors.

You Wu M.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is the clinical treatment of gastrointestinal tumors.

Yiwu Yuan M.D., Ph.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is the clinical treatment of gastrointestinal tumors.

Lin Xin M.D., Ph.D., Department of General Surgery, The Second Affiliated Hospital of Nanchang University, the research direction is gastrointestinal tumor metabolism and targeted therapy.

ORCID

Liqiang Zhou http://orcid.org/0000-0002-3713-258X

References

[1] Choi IJ, Kook MC, Kim YI, et al. Helicobacter pylori therapy for the prevention of metachronous gastric cancer. N Engl J Med. 2018;378(12):1085–1095.

[2] Miller KD, Nogueira L, Mariotto AB, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin. 2019;69(5):363–385.

[3] Gao K, Wu J. National trend of gastric cancer mortality in China (2003-2015): a population-based study. Cancer Commun (Lond). 2019;39(1):24.

[4] Shin Low S, Pan Y, Ji D, et al. Smartphone-based portable electrochemical biosensing system for detection of circulating microRNA-21 in saliva as a proof-of-concept. Sens Actuators B Chem. 2020;308:127718.

[5] How CW, Ong YS, Low SS, et al. How far have we explored fungi to fight cancer? Semin Cancer Biol. 2021. DOI:10.1016/j.semcancer.2021.03.009

[6] Xiang Z, Shen E, Li M, et al. Potential prognostic biomarkers related to immunity in clear cell renal cell carcinoma using bioinformatic strategy. Bioengineered. 2021;12(1):1773–1790.

[7] Liu Z, Sun D, Zhu Q, et al. The screening of immune-related biomarkers for prognosis of lung adenocarcinoma. Bioengineered. 2021;12 (1):1273–1285.

[8] Liu J, Li S, Feng G, et al. Nine glycolysis-related gene signature predicting the survival of patients with endometrial adenocarcinoma. Cancer Cell Int. 2020;20:183.

[9] Deng R, Zuo C, Li Y, et al. The innate immune effector ISG12a promotes cancer immunity by suppressing the canonical Wnt/β-catenin signaling pathway. Cell Mol Immunol. 2020;17(11):1163–1179.

[10] Zhang G, Fan E, Yue G, et al. Five genes as a novel signature for predicting the prognosis of patients with laryngeal cancer. J Cell Biochem. 2019;121(8–9):3804–3813.

[11] Du P, Kibbe W, Lin S. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24 (13):1547–1548.

[12] Leek J, Johnson W, Parker H, et al. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–883.

[13] Ritchie M, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43 (7):e47.

[14] Huang DW, Sherman BT, Tan Q, et al. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9):R183.

[15] Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61(1):92–105.

[16] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Nat Acad Sci. 2005;102(43):15545–15550.

[17] Xiong C, Wang G, Bai D. A novel prognostic models for identifying the risk of hepatocellular carcinoma based on epithelial-mesenchymal transition-associated genes. Bioengineered. 2020;11(1):1034–1046.

[18] Luck K, Kim DK, Lambourne L, et al. A reference map of the human binary protein interactome. Nature. 2020;580(7803):402–408.

[19] Szász A, Lánczky A, Nagy Á, et al. Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. Oncotarget. 2016;7(31):49322–49333.

[20] Oliveira C, Pinheiro H, Figueiredo J, et al. Familial gastric cancer: genetic susceptibility, pathology, and implications for management. Lancet Oncol. 2015;16 (2):e60–70.

[21] Van Cutsem E, Bang YJ, Feng-Yi F, et al. HER2 screening data from ToGA: targeting HER2 in gastric and gastroesophageal junction cancer. Gastric Cancer. 2015;18(3):476–484.

[22] Wang DS, Liu ZX, Lu YX, et al. Liquid biopsies to track trastuzumab resistance in metastatic HER2-positive gastric cancer. Gut. 2019;68(7):1152–1161.

[23] Muro K, Van Cutsem E, Narita Y, et al. Pan-Asian adapted ESMO clinical practice guidelines for the management of patients with metastatic gastric cancer: a

JSMO-ESMO initiative endorsed by CSCO, KSMO, MOS, SSO and TOS. Ann Oncol. 2019;30(1):19–33.

[24] Roma-Rodrigues C, Mendes R, Baptista PV, et al. Targeting tumor microenvironment for cancer therapy. Int J Mol Sci. 2019;20(4):840.

[25] Li X, Gu J, Zhou Q. Review of aerobic glycolysis and its key enzymes - new targets for lung cancer therapy. Thorac Cancer. 2015;6(1):17–24.

[26] Wei X, Mao T, Li S, et al. DT-13 inhibited the proliferation of colorectal cancer via glycolytic metabolism and AMPK/mTOR signaling pathway. Phytomedicine. 2019;54:120–131.

[27] Chiurillo MA. Role of the Wnt/β-catenin pathway in gastric cancer: an in-depth literature review. World J Exp Med. 2015;5(2):84–102.

[28] Mao J, Fan S, Ma W, et al. Roles of Wnt/β-catenin signaling in the gastric cancer stem cells proliferation and salinomycin treatment. Cell Death Dis. 2014;5(1):e1039.

[29] Zi Z. Molecular engineering of the TGF-β signaling pathway. J Mol Biol. 2019;431(15):2644–2654.

[30] Catteau X, Simon P, Noël JC. Myofibroblastic stromal reaction and lymph node status in invasive breast carcinoma: possible role of the TGF-β1/TGF-βR1 pathway. BMC Cancer. 2014;14:499.

[31] Bure IV, Nemtsova MV, Zaletaev DV. Roles of E-cadherin and Noncoding RNAs in the epithelial-mesenchymal transition and progression in gastric cancer. Int J Mol Sci. 2019;20(12):2870.

[32] Yue B, Song C, Yang L, et al. METTL3-mediated N6-methyladenosine modification is critical for epithelial-mesenchymal transition and metastasis of gastric cancer. Mol Cancer. 2019;18(1):142.

[33] Wiese C, Heisig J, Gessler M. Hey bHLH factors in cardiovascular development. Pediatr Cardiol. 2010;31(3):363–370.

[34] Sang L, Roberts JM, Coller HA. Hijacking HES1: how tumors co-opt the anti-differentiation strategies of quiescent cells. Trends Mol Med. 2010;16(1):17–26.

[35] Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial-mesenchymal transition. Nat Rev Mol Cell Biol. 2014;15(3):178–196.

[36] Weng HLSNW. Charactering tumor microenvironment reveals stromal-related transcription factors promote tumor carcinogenesis in gastric cancer. Cancer Med. 2020;9:5247–5257.

[37] Fanelli M, Chinen L, Begnami M, et al. The influence of transforming growth factor-α, cyclooxygenase-2, matrix metalloproteinase (MMP)-7, MMP-9 and CXCR4 proteins involved in epithelial-mesenchymal transition on overall survival of patients with gastric cancer. Histopathology. 2012;61(2):153–161.

[38] Lee K, Choi E, Hyun M, et al. Association of extracellular cleavage of E-cadherin mediated by MMP-7 with HGF-induced in vitro invasion in human stomach cancer cells. Eur Surg Res Europaische Chirurgische Forschung Recherches Chirurgicales Europeennes. 2007;39(4):208–215.

[39] Kashihara H, Shimada M, Yoshikawa K, et al. Correlation between thrombospondin-1 expression in non-cancer tissue and gastric carcinogenesis. Anticancer Res. 2017;37(7):3547–3552.

[40] Lin X, Chen S, Qi Y, et al. Overexpression of thrombospondin-1 in stromal myofibroblasts is associated with tumor growth and nodal metastasis in gastric carcinoma. J Surg Oncol. 2012;106(1):94–100.

[41] Nakao T, Kurita N, Komatsu M, et al. Expression of thrombospondin-1 and Ski are prognostic factors in advanced gastric cancer. Int J Clin Oncol. 2011;16(2):145–152.

[42] Xu Z, Chen J, Shu Y. Gene expression profile towards the prediction of patient survival of gastric cancer. Biomed Pharmacothe. 2010;64(2):133–139.

[43] Zhang J, Ito R, Oue N, et al. Expression of thrombospondin-1 is correlated with microvessel density in gastric carcinoma. Virchows Arch. 2003;442(6):563–568.

[44] Hu H, Xu D, Huang X, et al. Keratin17 promotes tumor growth and is associated with poor prognosis in gastric cancer. J Cancer. 2018;9(2):346–357.

[45] Chivu-Economescu M, Dragu D, Necula L, et al. Knockdown of KRT17 by siRNA induces antitumoral effects on gastric cancer cells. Gastric Cancer. 2017;20(6):948–959.