Computational and Structural Biotechnology Journal 18 (2020) 3217–3229

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Development and validation of metabolism-related gene signature in prognostic prediction of gastric cancer

Tianqi Luo [a,b,1], Yuanfang Li [a,b,1], Runcong Nie [a,b,1], Chengcai Liang [a,b,1], Zekun Liu [b], Zhicheng Xue [a,b], Guoming Chen [a,b], Kaiming Jiang [a,b], Ze-Xian Liu [b], Huan Lin [c,*], Cong Li [b,d,*], Yingbo Chen [a,b,*]

[a] Department of Gastric Surgery, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou 510060, China
[b] State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[c] The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou 510120, China
[d] Department of Colorectal Surgery, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou 510060, China

ARTICLE INFO

Article history:
Received 13 July 2020
Received in revised form 24 September 2020
Accepted 26 September 2020
Available online 17 October 2020
Dataset link: https://www.ncbi.nlm.nih.gov/geo/https://gdc.xenahubs.net

Keywords:
Gastric cancer
Metabolic studies
Nomogram
Prognosis

ABSTRACT

Gastric cancer is one of the most common malignant tumours in the world. As one of the crucial hallmarks of cancer reprogramming of metabolism and the relevant researches have a promising application in the diagnosis treatment and prognostic prediction of malignant tumours. This study aims to identify a group of metabolism-related genes to construct a prediction model for the prognosis of gastric cancer.

A large cohort of gastric cancer cases (1121 cases) from public database was included in our analysis and classified patients into training and testing cohorts at a ratio of 7: 3. After identifying a list of metabolism-related genes having prognostic value, we constructed a risk score based on metabolism-related genes using LASSO-COX method. According to the risk score, patients were divided into high- and low-risk groups. Our results revealed that high-risk patients had a significantly worse prognosis than low-risk patients in both the training (high-risk vs low-risk patients; five years overall survival: 37.2% vs 72.2%; $p < 0.001$) and testing cohorts (high-risk vs low-risk patients; five years overall survival: 42.9% vs 62.9%; $p < 0.001$). This observation was validated in the external validation cohort (high-risk vs. low-risk patients; five years overall survival: 30.2% vs 40.4%; $p = 0.007$).

To reinforce the predictive ability of the model, we integrated risk score, age, adjuvant chemotherapy, and TNM stage into a nomogram. According to the result of receiver operating characteristic curves and decision curves analysis, we found that the nomogram score had a superior predictive ability than conventional factors, indicating that the risk score combined with clinicopathological features can develop a robust prediction for survival and improve the individualized clinical decision making of the patient.

In conclusion, we identified a list of metabolic genes related to survival and developed a metabolism-based predictive model for gastric cancer. Through a series of bioinformatics and statistical analyses, the predictive ability of the model was confirmed.

1. Introduction

Gastric cancer (GC) is the most frequently diagnosed type of malignant tumour and the primary cause of cancer death worldwide [1]. Although significant improvement has been witnessed in the survival of GC due to early diagnosis and comprehensive treatment [2,3], GC's prognosis remains relatively poor. Presently, the American Joint Committee on Cancer (AJCC) tumour-node-metastasis (TNM) staging system has been generally conducted to predict the prognosis of GC [4], whereas several patients with a similar tumour stage finally obtain different clinical outcomes, indicating that the TNM staging system is still incomplete.

In addition, carcinoembryonic antigen (CEA), carbohydrate antigen (CA) 19-9, and CA 72-4 are used in clinical prediction for GC widely. However, these biomarkers have a restricted efficiency in

* Corresponding authors at: State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China.
E-mail addresses: reallinhuan@hotmail.com (H. Lin), licong2@sysucc.org.cn (C. Li), chenyb@sysucc.org.cn (Y. Chen).
[1] These authors contributed equally to this study.

https://doi.org/10.1016/j.csbj.2020.09.037

prediction for prognosis. Currently, gene biomarkers, such as microRNA, circular RNA, and mRNA are becoming important increasingly in the application of GC's prognosis [5,6]. On the other hand, there are many pieces of research developing prognostic classifier that could split GC patients into different risk groups based on the multigene expression [7–10]. Unfortunately, these studies have a small size in the sample and fail to perform an internal validation to estimate the possibility for optimism and overfitting in model performance [11].

Reprogramming of cellular metabolism is an important hallmark of cancer and is strongly associated with the tumorigenesis [12,13]. In the 1920s, Warburg had revealed that an increased amount of glucose was consumed by tumour tissues compared to normal tissues [14]. Additionally, excessive activated anaerobic glycolysis and impaired aerobic respiration are considered features of tumour cells [15], and this observation was also found in GC samples [16]. Aside from the metabolism of glucose, amino acid, lipid, nucleotide, and other metabolite metabolism also present an increased or decreased trend in the development of GC [17]. With the increasing application of bioinformatics analysis in the diagnosis and prognosis prediction of malignant tumours, several investigators have linked the metabolome to the genome, which allows broad and accurate metabolite profiles to be profiled [18].

Therefore, it is essential to combine metabolomics with genomics, and transcriptomics to obtain a more comprehensive understanding of tumour metabolism. Previous studies have utilized metabolism-related genes to generate a predictive model for glioma and achieve an excellent prediction effect [19–21]. To date, however, there are no types of research using the expression profile of metabolism-related genes to assist in the prediction of GC patient outcomes. Therefore, the present study aims to identify a group of metabolism-related genes to construct a predictive model for GC.

2. Method and materials

2.1. Collection of data

We downloaded the gene expression profiling of GC from the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) database. The detailed inclusion criteria for candidate datasets are following: Human gene expression profile; gastric cancer specimen; samples' total count $\geq$ 90; availability of follow-up information (overall survival) and related clinical data. A total of five datasets (GSE84437, GSE62254, GSE26942, GSE29272, and GSE13861) were included in our study.

For the purpose of estimating the power and robustness of the model, The Cancer Genome Atlas stomach adenocarcinoma (TCGA-STAD) cohort, as the external validation cohort, was obtained from UCSC Xena website (https://gdc.xenahubs.net). Furthermore, to evaluate the specificity of the metabolic gene signature for GC, we downloaded the mRNA sequencing data of the remaining 32 TCGA tumours from UCSC Xena website.

In this study, clinical variates involved age, sex, American Joint Committee on Cancer (AJCC) stage, histological grade, Lauren type, adjuvant chemotherapy (ACT), survival status, and survival time. Subsequently, we excluded normal tissues adjacent to cancer, gastric stromal tumours, and cases that lacked survival information from these datasets.

2.2. Data processing

The mRNA microarray data sets had been normalized before being downloaded from the GEO database. The sequencing data were normalized as Fragments Per Kilobase Million (FPKM) value. Probe identifications of gene matrix files were transformed into gene symbols according to the annotation file from the corresponding platform. All gene expression values were processed by log2, and the average value was taken as the final expression value if multiple probes corresponded to the same gene symbol. Because of our data involving a combination of multiple datasets, the Empirical Bayes method ("sva" package) was executed to diminish the batch effect after merging these series [22]. Finally, based on the ratio of 7: 3, these cases were randomly grouped into a training and testing cohorts (internal validation).

2.3. Extraction of metabolism-related genes

In this study, all metabolism-related genes were derived from Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolism-related gene sets ("c2.cp.kegg.v7.0.symbols.gmt"; http://software.broadinstitute.org/gsea/downloads.jsp). After intersecting the whole gene set of samples with the metabolic gene sets, we identified 703 metabolism-related genes in our transcriptome data. The expression level of these genes was also extracted from each case to perform further analysis.

2.4. Construction and validation of the metabolism-related signature

In the training cohort, univariate Cox regression analysis ("survival" package) was performed to screen the metabolic genes correlated with survival. To address the impact of overfitting, the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm ("glmnet" package) was performed to select potential genes to construct metabolic gene signature [23]. LASSO method has many advantages in the development of the model. First of all, the Lasso method is initially formulated for linear regression models, which makes the model more simple and visualized. It also can reduce variance through removing and shrinking coefficients, which provide a good prediction accuracy. Furthermore, it can increase the model interpretability and decrease overfitting by eliminating irrelevant variables.

The values of penalty parameter λ were determined by 200-fold cross-validations. Finally, these genes selected by LASSO were used to develop a formula comprising the gene expression level (expr) weighted by the corresponding coefficient:

$$\begin{aligned} Risk\,score = {} & [Expr\,of gene(1) \times coefficient\,of\,gene(1)] \\ & + [Expr\,of\,gene\,(2) \times coefficient\,of\,gene(2)] + \ldots \\ & + [Expr\,of\,gene\,(n) \times coefficient\,of\,gene(n)] \end{aligned}$$

Based on the optimal cut-off value of the training cohort calculated by "survminer" package, training, and test cohorts were classified into high- and low-risk groups, respectively. Next, we applied the multivariate Cox regression analyses to determine whether the risk score was an independent prognostic factor for GC. Moreover, calibration curve, receiver operating characteristic (ROC) analysis ("survival ROC" package), and decision curve analysis (DCA) was used to estimate the accuracy and clinical utility of the model for prognosis [24].

2.5. Bioinformatic analysis

Based on Hallmarks gene set ("h.all.v7.0.symbols.gmt"), GSEA software (https://www.gsea-msigdb.org/gsea/login.jsp) was applied to identify the significantly enriched pathways between the high- and low-risk groups. Metascape tool (http://metascape.org/) was carried out to achieve the functional annotation of the metabolism-related genes selected by LASSO [25]. Moreover, we investigated the association between the risk score and immune cell infiltration using CIBERSORT algorithm (https://cibersort.

stanford.edu/index.php), and the algorithm can identify 22 types of human immune cell phenotypes according to the gene expression data [26]. Samples with a CIBERSORT result of $p < 0.01$ were considered to be eligible for further analysis.

2.6. Statistical analysis

All statistical analyses were performed using R version 3.6.0 (http://www.r-project.org). Wilcoxon and Kruskal-Wallis tests were used to compare continuous variables. Categorical variables

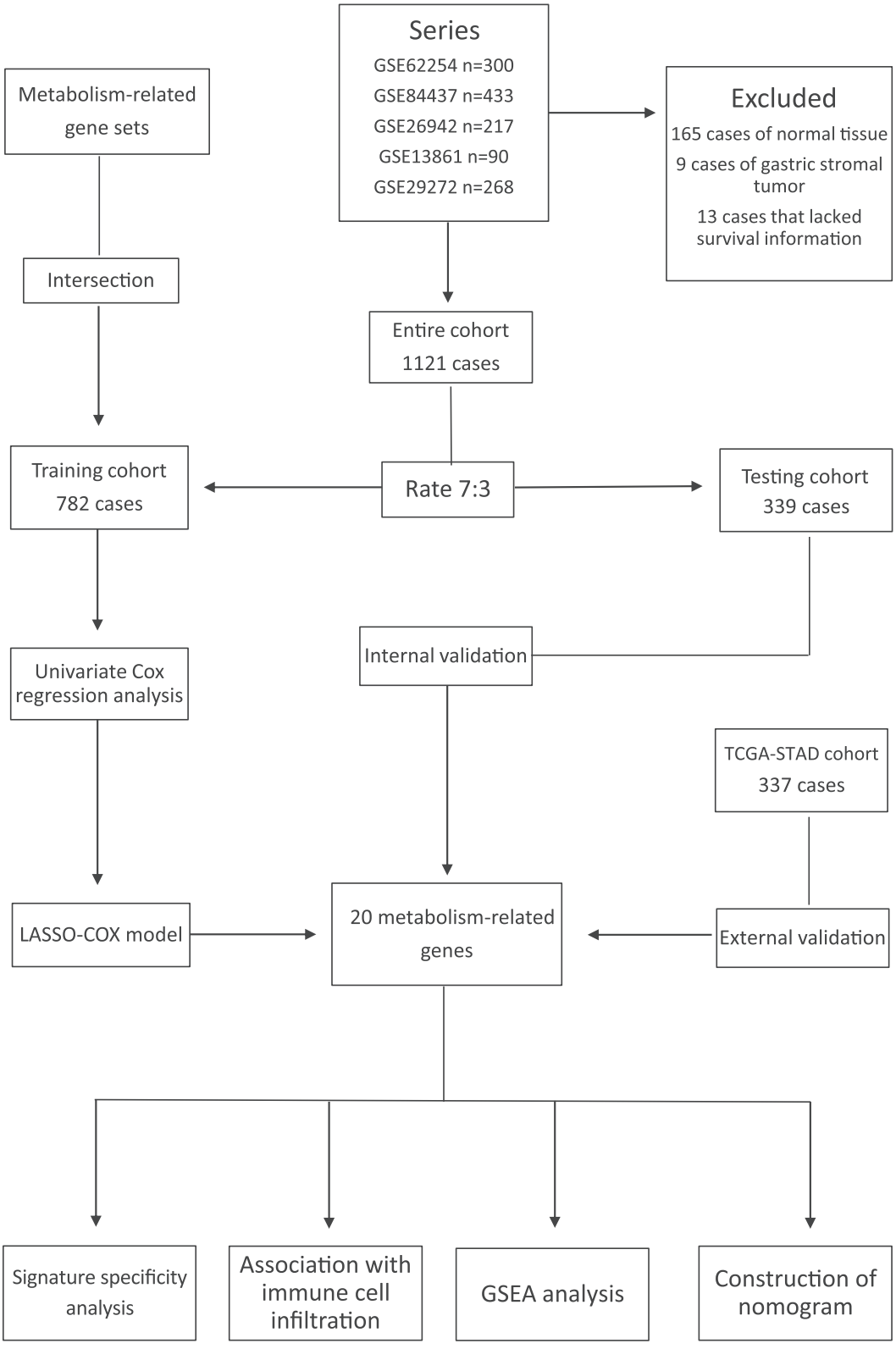

Fig. 1. Flow diagram of the model's Construction and validation. LASSO, least absolute shrinkage and selection operator.

were analyzed by chi-square. Overall survival (OS) was defined as the length of time from the date of diagnosis to death from any cause. The survival curves of the two groups were estimated using the Kaplan–Meier method, and significant differences were examined using the log-rank test. All p values were based on two-sided statistical tests, and $p < 0.05$ was considered statistically significant.

3. Result

3.1. Patient characteristics and the Construction of the risk score

As shown in Fig. 1, we excluded 165 cases of normal tissue adjacent to cancer, 9 cases of gastric stromal tumour, and 13 cases that lacked survival information. Finally, a total of 1121 cases were included in our analysis, and patient characteristics are displayed in Table 1. Subsequently, these cases (1121) were randomly separated into a training cohort (7 8 2) and testing cohort (3 3 9) at a ratio of 7: 3. In the training cohort, the univariate Cox regression analysis was utilized to screen metabolic genes related to survival. As a result, 56 metabolic genes ($p < 0.001$) were selected to achieve further analysis. Nevertheless, Strong correlations among these genes were observed in the training cohort (Fig. 2A). Therefore, we used the LASSO algorithm to reduce overfitting and construct the model.

After performing and the LASSO algorithm, a total of 20 metabolism-related genes were selected to develop the formula (Fig. 2B, C):

$$\begin{aligned}\textit{Risk score} = {} & \textit{GSTZ1} \times (-0.1612) + \textit{ACOX3} \times (-0.0262) + \textit{CYB5R3} \times (0.1392) + \textit{PDE8B} \times (0.0269) + \textit{LTC4S} \times (0.0007) + \textit{ME2} \times (-0.0671) + \textit{AMD1} \times (-0.0003) + \textit{PAFAH2} \times (-0.1058) + \textit{GSTO1} \times (-0.1627) + \textit{METTL2B} \times (-0.3462) + \textit{TYRP1} \times (0.1403) + \textit{ALDH1A3} \times (0.0157) + \textit{DDC} \times (-0.0223) + \textit{DGKI} \times (0.0574) + \textit{GUCY1A2} \times (0.5639) + \textit{SCLY} \times (-0.0965) + \textit{CYP1B1} \times (0.0125) + \textit{CD38} \times (-0.2146) + \textit{COX10} \times (-0.1336) + \textit{HIBCH} \times (-0.1082)\end{aligned}$$

Then, a risk score was calculated for each patient based on the formula. Finally, we calculated the optimal cut-off value (−3.82) that can generate the largest survival difference between the high- and low-risk groups (Fig. S1). In the training cohort, the distribution of risk scores and the survival status of patients are displayed in Fig. 2D.

3.2. Validation and evaluation of the metabolic gene signature

As shown in Fig. 3A, high-risk patients remarkably had a worse prognosis compared to low-risk patients in the training cohort (high-risk vs. low-risk patients; five years OS: 37.2% vs. 72.2%; $p < 0.001$). The results of the multivariate Cox regression analysis showed that the risk score, as a continuous variable, was significantly associated with OS, suggesting that it was an independent prognostic factor (HR = 2.78, 95% CI: 2.02–3.83; $p < 0.001$; Table 2).

To validate the prognostic value of the signature in another cohort, the same formula was applied to the testing cohort for calculating the risk score. Similarly, patients in the testing cohort were divided into high- and low-risk groups based on the same cut-off value. The results revealed that a significant difference in

Table 1
Clinicopathological characteristics.

	Training cohort (n = 782)		Testing cohort (n = 339)	
Variable	high-risk group	low-risk group	high-risk group	low-risk group
N	286	496	123	216
Risk score (median)	−3.58 (−3.73, −3.37)	−4.16 (−4.39, −3.99)	−3.57 (−3.70, −3.40)	−4.19 (−4.35, −4.01)
Age (median)	61.00 (52.00, 68.00)	61.00 (53.00, 68.00)	61.00 (52.00, 68.00)	63.00 (54.00, 68.00)
Gender				
Male	87 (30.5)	154 (31.1)	89 (72.4)	148 (68.5)
Female	198 (69.2)	340 (68.5)	34 (27.6)	68 (31.5)
Unknown	1 (0.3)	2 (0.4)	0 (0)	0 (0)
Stage				
I	15 (5.2)	60 (12.1)	6 (4.9)	17 (7.9)
II	25 (8.7)	83 (16.7)	12 (9.8)	28 (13.0)
III	101 (35.3)	159 (32.1)	51 (41.5)	74 (34.3)
IV	22 (7.7)	20 (4.0)	7 (5.7)	7 (3.1)
Unknown	123 (43.1)	174 (35.1)	47 (38.2)	90 (41.7)
Lauren classification				
diffuse	50 (17.5)	96 (19.4)	28 (22.8)	30 (13.9)
intestinal	85 (29.7)	172 (34.7)	38 (30.9)	62 (28.7)
mixed	11 (3.8)	16 (3.2)	2 (1.6)	10 (4.6)
Unknown	140 (49.0)	212 (42.7)	55 (44.7)	114 (52.8)
Adjuvant chemotherapy				
No	71 (24.8)	124 (25.0)	32 (26.0)	38 (17.6)
Yes	62 (21.7)	152 (30.6)	30 (24.4)	55 (25.5)
Unknown	153 (53.5)	220 (44.4)	61 (49.6)	123 (56.9)
Tumor location				
antrum	64 (22.4)	139 (28.0)	33 (26.8)	49 (22.7)
body	62 (21.7)	112 (22.6)	26 (21.2)	33 (15.2)
cardia	25 (8.7)	42 (8.5)	11 (8.9)	20 (9.3)
Unknown	135 (47.2)	203 (40.9)	53 (43.1)	114 (52.8)
Histological grade				
G1/G2	34 (11.9)	79 (15.9)	17 (13.8)	38 (17.6)
G3	66 (23.1)	119 (24.0)	32 (26.0)	41 (19.0)
Unknown	186 (65.0)	298 (60.1)	74 (60.2)	137 (63.4)
Overall survival				
Alive	101 (35.3)	330 (66.5)	51 (41.5)	132 (61.1)
Dead	185 (64.7)	166 (33.5)	72 (58.5)	84 (38.9)
Survival time (median)	2.67 (1.04, 5.74)	5.50 (2.26, 7.64)	2.14 (1.16, 6.07)	4.05 (1.88, 7.03)

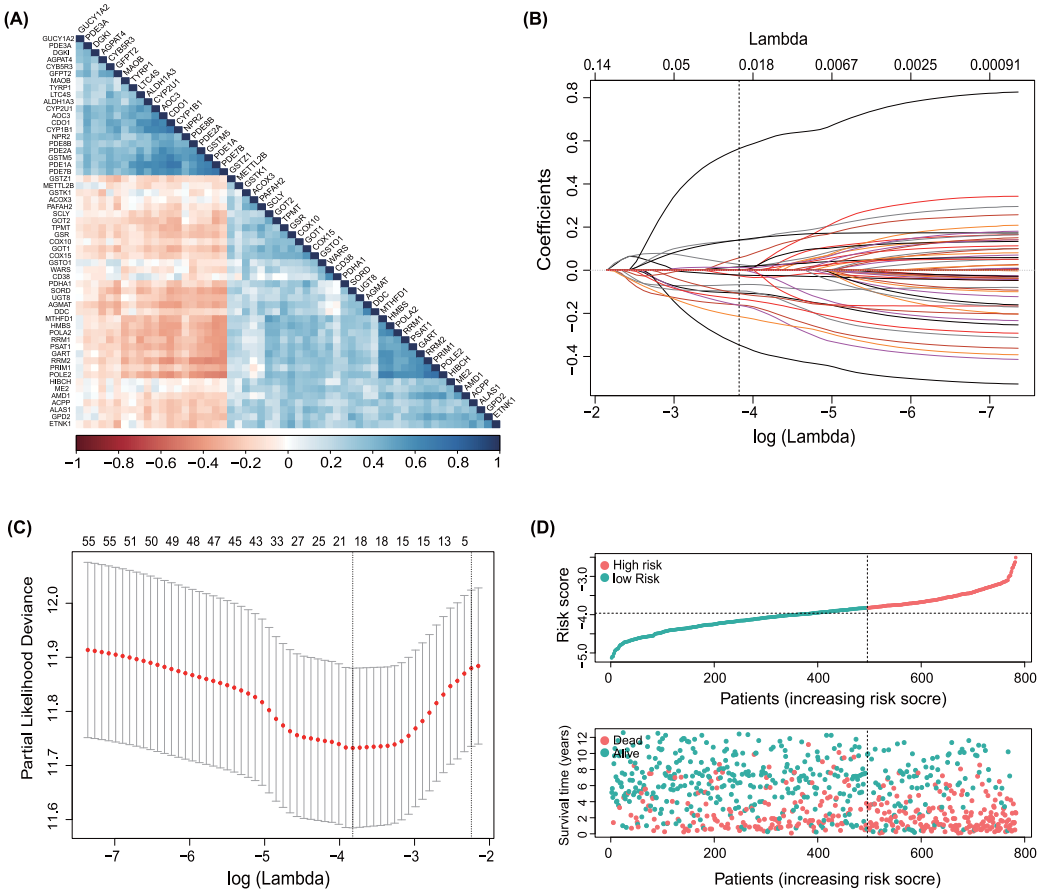

Fig. 2. Construction of the risk score in the training cohort. (A). The correlation analysis of the 56 metabolism-related genes (B). LASSO coefficient of the 20 metabolism-related genes. The dotted vertical line shows the value of lambda selected by 200-fold cross-validation via minimum criteria; (C). The 200-fold cross-validation for variable selection in the LASSO model. The two dotted vertical lines indicate the optimal values by using the minimum criteria and the 1-SE criteria; (D). Distribution of the risk score and survival state of patients in the training cohort.

survival was observed between the two groups in the testing cohort (high-risk vs low-risk patients; five years OS: 42.9% vs 62.9%; $p < 0.001$; Fig. 3B). Consistently, multivariate Cox regression analysis suggested that adjusting for covariates of age, ACT, and tumour stage, the risk score (continuous variable) was identified as an independent prognostic factor in the testing cohort (HR = 2.08, 95% CI: 1.37–3.14; $p < 0.001$; Table 3).

Subsequently, we estimated the predictive ability of the risk score for the two-, three-, and five-year OS by analyzing the time-dependent ROC curve. The ROC results in training, testing and entire cohorts were depicted in Fig. 3D, E, and F, respectively. Finally, principal component analysis (PCA) was performed to investigate the different distribution patterns between the high- and low-risk groups on the basis of the metabolic genes and whole-genome expression. The result showed that whether based on the metabolism-related genes or whole-genome expression, the two groups presented a distinct separation in the training and testing cohorts, respectively. (Fig. S2).

3.3. Correlation between the risk score and clinical features

To explore the prognostic value of the signature in a different population, the entire cohort was classified into several subgroups based on clinical features to estimate survival curves between the high- and low-risk group. However, subgroup survival analyses found that whether tumour stage I/II or III/IV, and receiving ACT or not, the patients in the high-risk group were significantly associated with a poor prognosis (all $p < 0.001$; Fig. 4A-D). Additionally, cases in the entire cohort were used to investigate the correlations between the risk score and clinicopathological characteristics of patients (Fig. 5A). The results showed no differences in the risk score between different sexes, age groups, histological grade, Lauren classifications, and tumour locations (all $p > 0.05$). However, we found that a higher risk score was significantly associated with advanced tumour stage ($p < 0.001$).

We finally evaluated the association between the risk score and immune microenvironment using CIBERSORT algorithm. As shown in Fig. 5B, B cells naïve, T cells CD4 memory resting, monocytes, macrophage M2, and mast cells resting were more enriched in the samples of the high-risk group. However, plasma cells, T cells CD8, T cells CD4 memory activated, T cells follicular helper, and macrophage M1 showed more abundant density in the samples of the low-risk group (all $p < 0.05$).

3.4. External validation in the independent dataset

After excluding 38 GC patients without survival information, we used TCGA-STAD cohort containing 337 GC patients as external validation. By using the same formula, we calculated risk score for each patient in TCGA-STAD cohort. Patients were separated into high- and low-risk groups based on the median of the risk scores as cut-off value. The survival curve showed that the low-risk group were significantly associated with lower mortality (high-risk vs low-risk patients; five years overall survival: 30.2% vs 40.4%; $p = 0.007$; Fig. 6A). This finding was consistent with observations from the training cohort.

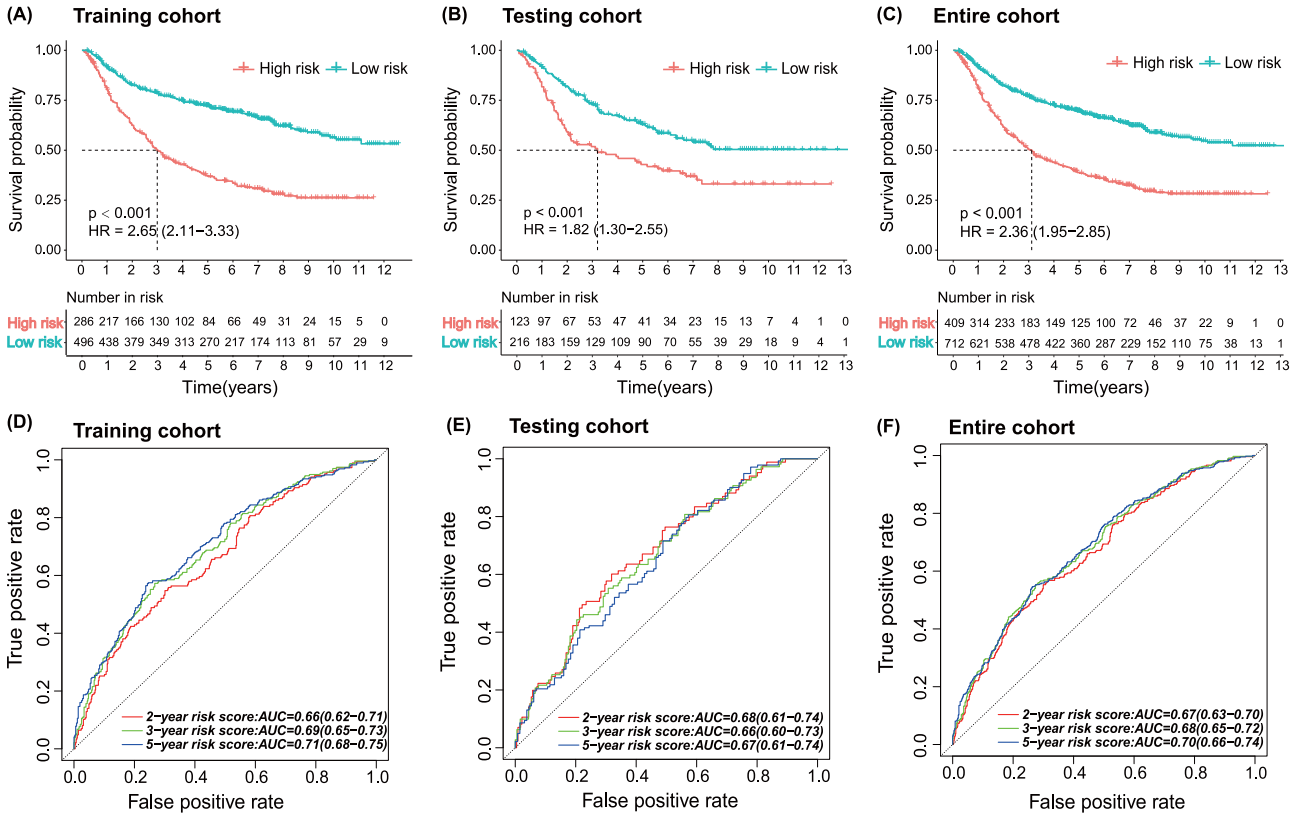

Fig. 3. The association between the overall survival and risk score. (A)–(C). Kaplan-Meier curves for overall survival between the high- and low-risk groups in the training, testing, and entire cohorts, respectively; (D)–(F). The time-dependent Receive operating characteristic (ROC) curves for the risk score in the training, testing, and entire cohorts, respectively.

Table 2
Univariate and multivariate Cox regression analyses of the risk score and clinical characteristics with the overall survival in the training cohort.

	Univariate analysis			Multivariate analysis		
Variable	HR	95%CI	*p*-value	HR	95%CI	*p*-value
Age	1.02	1.01–1.03	**<0.001**	1.02	1.01–1.03	**0.006**
Gender (male vs. female)	1.18	0.93–1.48	0.170			
Tumor stage (III/IV vs. I/II)	3.34	2.40–4.64	**<0.001**	3.49	2.45–4.98	**<0.001**
Adjuvant chemotherapy (Yes vs. No)	0.53	0.40–0.71	**<0.001**	0.48	0.35–0.65	**<0.001**
Risk score	3.61	2.83–4.60	**<0.001**	2.78	2.02–3.83	**<0.001**

Table 3
Univariate and multivariate Cox regression analyses of the risk score and clinical characteristics with the overall survival in the testing cohort.

	Univariate analysis			Multivariate analysis		
Variable	HR	95%CI	*p*-value	HR	95%CI	*p*-value
Age	1.00	0.99–1.02	0.571			
Gender (male vs. female)	0.78	0.56–1.09	0.143			
Tumor stage (III/IV vs. I/II)	3.25	1.91–5.55	**< 0.001**	3.91	2.18–7.01	**<0.001**
Adjuvant chemotherapy (Yes vs. No)	0.61	0.40–0.94	**0.026**	0.50	0.32–0.78	**0.002**
Risk score	2.75	1.95–3.89	**< 0.001**	2.08	1.37–3.14	**<0.001**

To estimate the specificity of the metabolic gene signature for GC, we applied the same formula to the mRNA sequencing data of 32 types of TCGA tumours, including 30 solid tumours and two blood system tumours. Based on the result, we found that the signature can significantly distinguish patients with different survival outcomes in seven types of tumours (cervical cancer, mesothelioma, ovarian cancer, cutaneous melanoma, kidney chromophobe, rectal cancer, and thymoma; $p < 0.05$; Fig. 6B–H). Moreover, the signature was also marginally associated with survival outcomes in four types of tumours (ocular melanoma, bladder cancer, sarcoma, and kidney papillary cell carcinoma; Fig. 6I–L). Although the results of log-rank test did not reach the level of statistical significance ($p < 0.05$), the survival curves were remarkably separated and the result of the hazard ratio was acceptable in the four tumours.

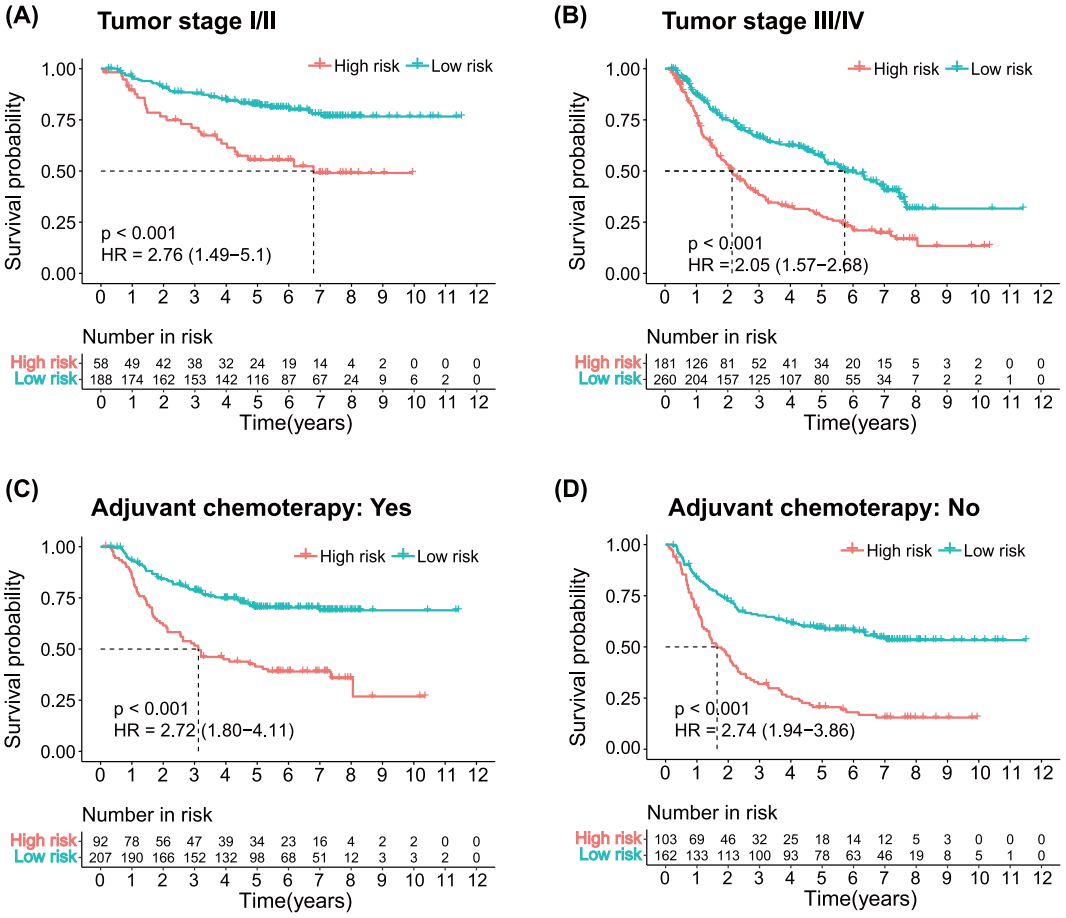

Fig. 4. Stratified survival analysis of the high- and low-risk groups in the entire cohort. (A). Kaplan-Meier curves for overall survival between high- and low-risk group in the patients with I/II stage; (B). Kaplan-Meier curves for overall survival between high- and low-risk group in the patients with III/IV stage; (C). Kaplan-Meier curves for overall survival between the high- and low-risk group in the patients who received adjuvant chemotherapy. (D). Kaplan-Meier curves for overall survival between the high- and low-risk group in the patients who did not receive adjuvant chemotherapy.

3.5. Construction and evaluation of nomogram

To reinforce the model's predictive power, we integrated the risk score and three clinical variables, including age, ACT, and tumour stage into a nomogram (Fig. 7A). The calibration curves for the nomogram showed favourable consistency between actual observation and predictive value (Fig. 7B). In Fig. 8A, we displayed the ROCs for the five-year OS of these variables. the Area Under Curves (AUCs) of the nomogram score were 0.81, 0.75, and 0.80 in the training, testing, and entire cohorts, respectively, with better prognostic efficiency compared to the other variables ($p < 0.05$). Lastly, DCA was used to compare the clinical net benefit between the nomogram and other models. As shown in Fig. 8B, the nomogram had a better net benefit across a wider scale of threshold probabilities for predicting three-year OS than conventional staging system and risk score.

3.6. Exploration of biological function

To determine the potential biological pathway, GSEA software was utilized to explore the differences in the Hallmark pathway between the two groups. According to our results, the six pathways were significantly enriched in the high-risk groups (all $p < 0.05$, FDR q < 0.25, $|NES| \geq 1$), including "myogenesis", "epithelial-mesenchymal transition", "UV response DN", "angiogenesis", "apical junction", and "hedgehog signaling" (Fig. 9A). Next, we demonstrated the top five pathways in the high- and low-risk groups in Fig. 9B. Finally, the Metascape tool was used to achieve the functional annotation for the 20 metabolism-related genes and to help us explore the potential molecular mechanisms (Fig. 9C, D). The result showed that the biological processes of these genes primarily engaged in the pathways named "tyrosine metabolism", "cofactor metabolic process", "metabolism of amino acids and derivatives", "monocarboxylic acid metabolic process", "nucleotide metabolic process", "biological oxidations", and "generation of precursor metabolites and energy".

4. Discussion

As the only predictive model generally used in clinical, the TNM staging system merely used clinicopathological features of patients to predict outcomes. Owing to the high heterogeneity of GC, patients with similar stage often have different survival outcomes, indicating that the TNM staging system has reached its limit of predicting patients' survival. Although numerous prognostic models using molecular signature have been developed by researchers, only conventional Her-2, CEA, CA19-9, and CA72-4 were applied to assisting prediction for GC patients' outcomes in clinical practice

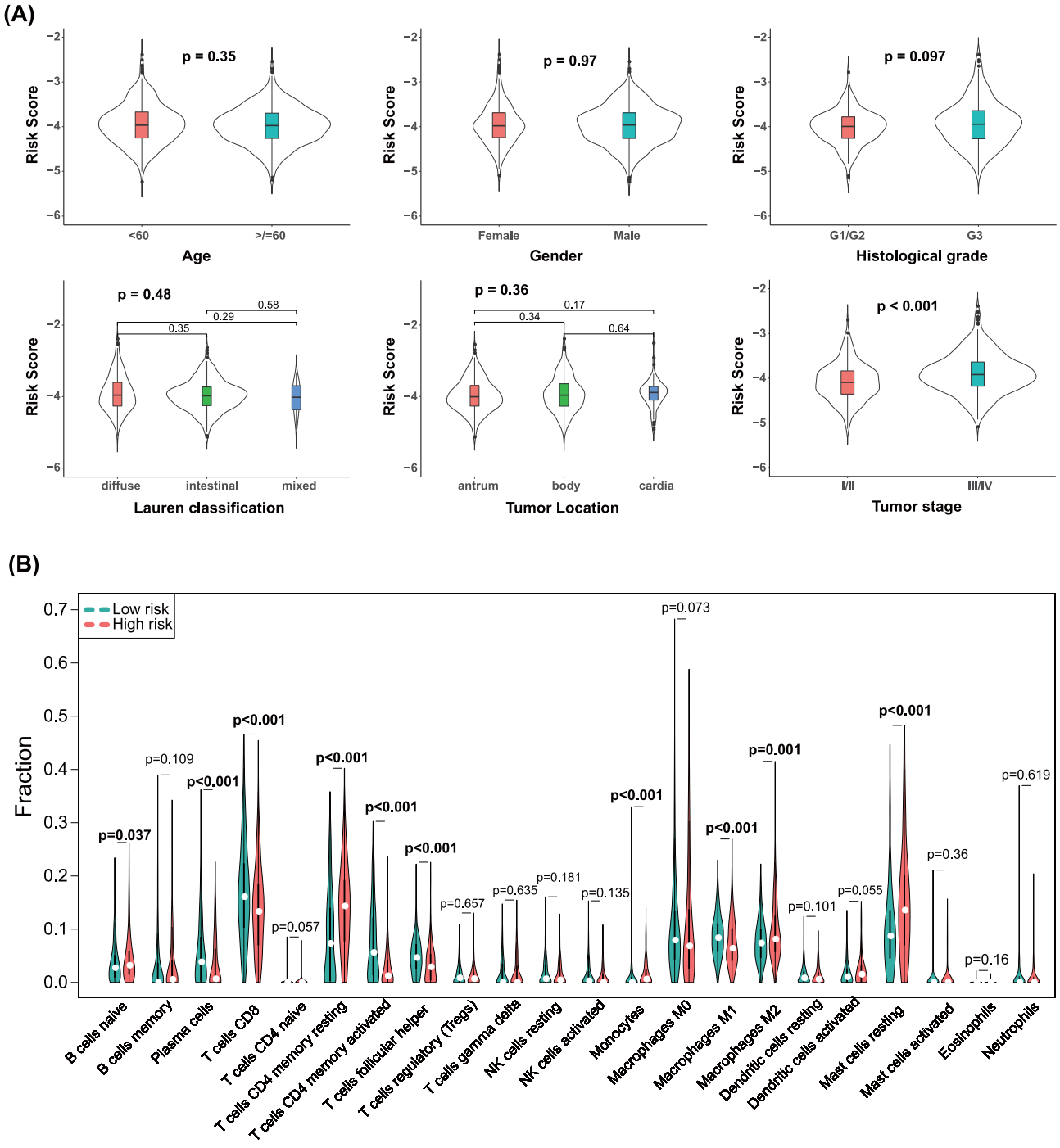

Fig. 5. (A). Distribution of the risk score in different clinicopathological features in the entire cohort. (B). Comparisons of 22 infiltrated immune cells between the high- and low-risk groups.

[27]. However, these biomarkers were mainly used in diagnosis and monitoring of recurrence in GC patients. The biomarker that can play a key role in prognostic prediction for GC were rare. With the rapid development of gene sequencing technology, a lot of solid biomarkers that have prognostic value for GC patients has been identified and validated in multiple independent datasets [28], which promotes the accuracy of prediction for prognosis.

To ensure the effectiveness and stability of the predictive model, the study involved multiple datasets (1121 cases) in constructing a novel gene signature and performed external validation in an independent cohort. The main finding showed that patients with high-risk scores were negatively associated with survival, and this observation was confirmed in the internal validation cohort. This finding was also validated in the TCGA-STAD cohort. In the analysis of specificity for the signature, we selected the 32 types of TCGA tumours to validate the 20 genes signature. According to the survival curve analysis, the signature was correlated with overall survival in the 11 types of tumours.

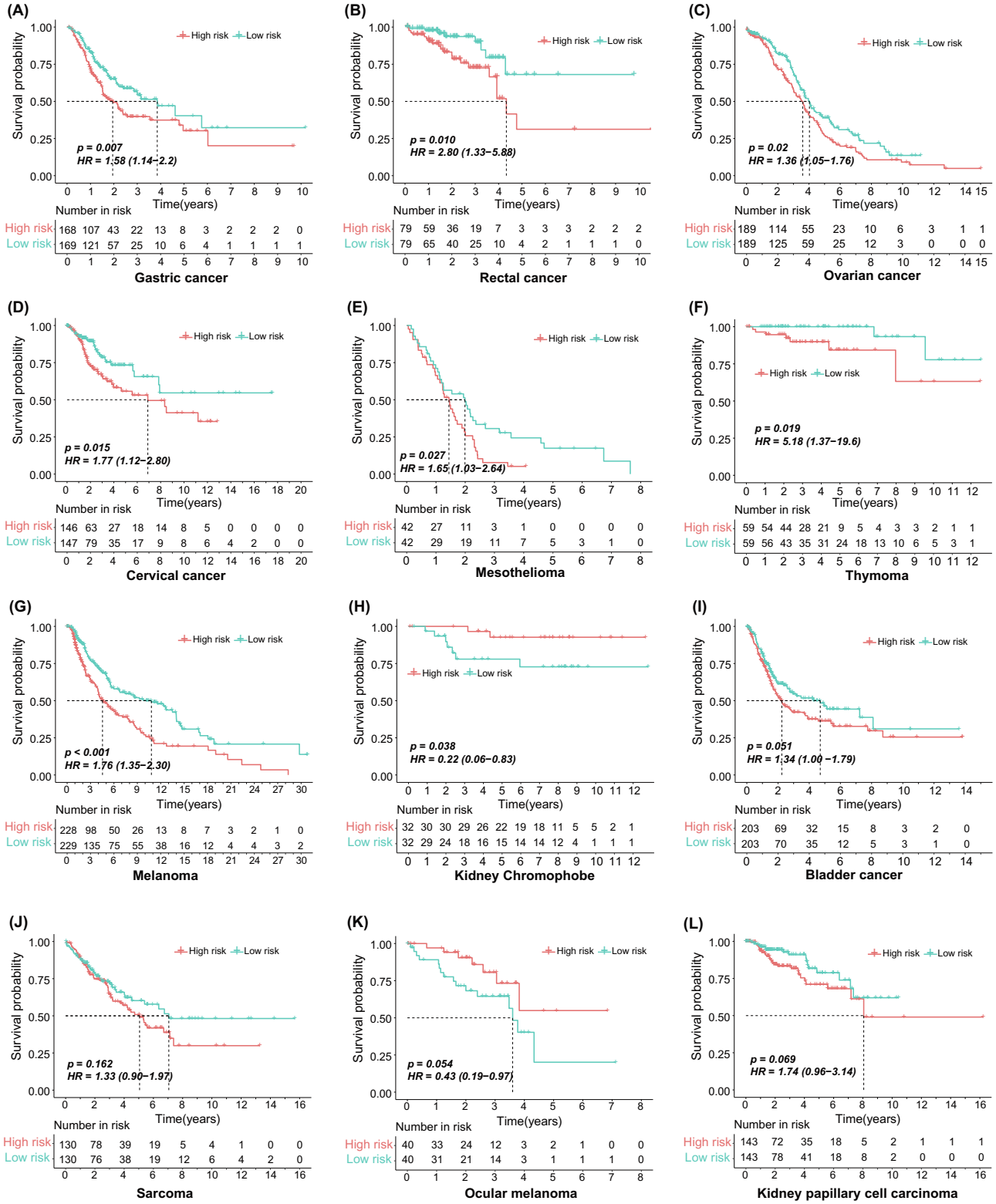

Fig. 6. External validation in the independent datasets; (A). Kaplan-Meier curves for overall survival between the high- and low-risk group in the TCGA-STAD dataset. (B)–(L). Kaplan-Meier curves for overall survival between the high- and low-risk group in the 11 types of tumour datasets.

One explanation for this result may be that the metabolic signature is specific to several tumours to some extent. In these tumours, there may be similar metabolic pathways that play an important role in tumorigenesis. Identification of these pathways will contribute to the research of the mechanism and development of the drug.

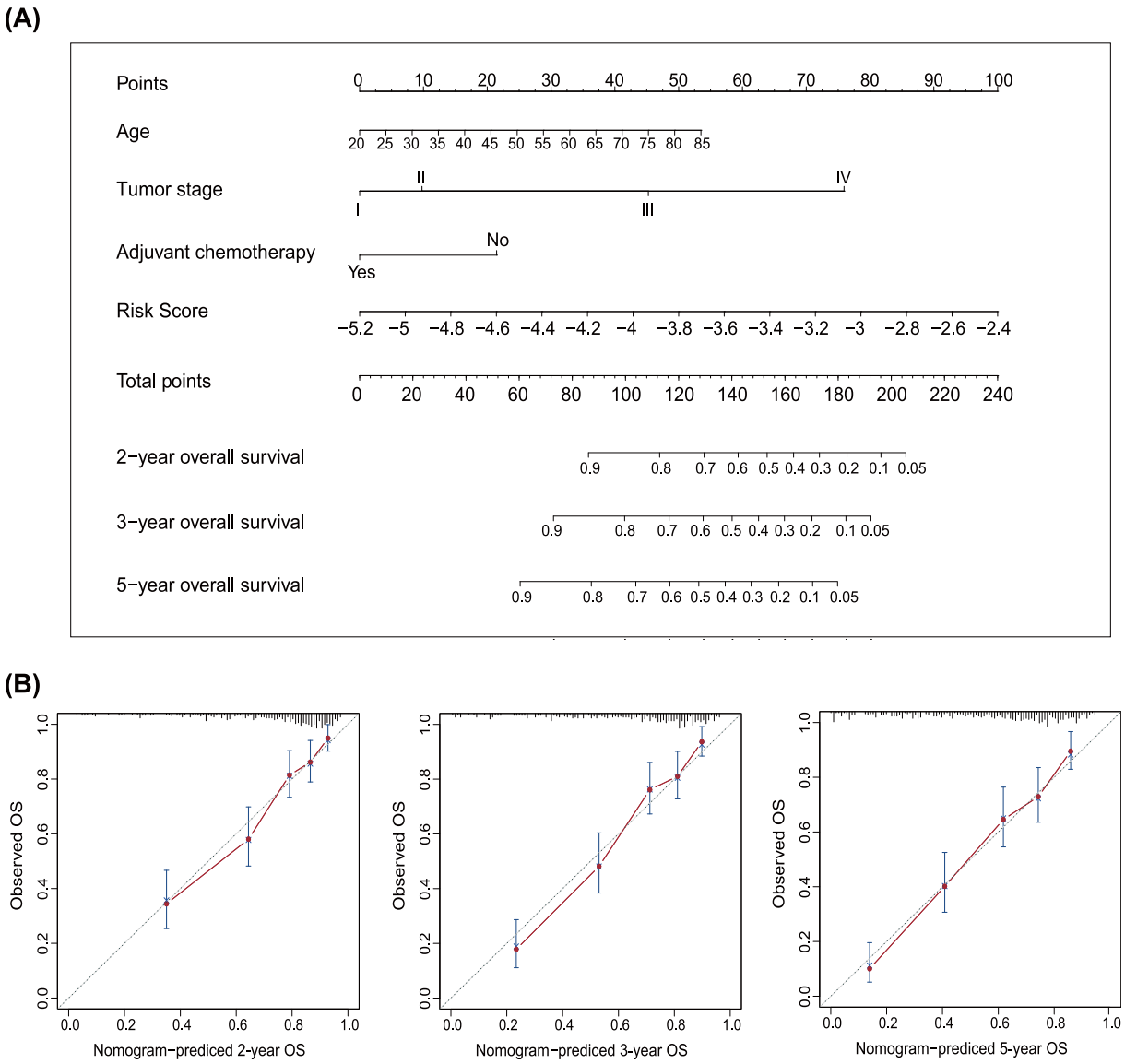

Fig. 7. Construction of nomogram; (A). A nomogram that integrated the risk score and age, adjuvant chemotherapy, and tumour stage. (B). The calibration curves for two, three, and five-year overall survival.

To determine whether tumour stage and use of ACT can affect the predictive ability of the risk score, stratification analysis was conducted to compare the two groups' survival in a different subgroup (tumour stage I/II or III/IV, and receiving ACT or not). The results showed that a survival advantage for the low-risk group could be displayed in the four subgroups, suggesting that the risk score has a broad utility in GC patients. Moreover, we integrated risk score, age, ACT, and TNM stage into a nomogram and then, calculated a nomogram score for each patient. According to the result of ROC and DCA, we found that the nomogram score had a superior predictive ability than conventional staging system ($p < 0.001$), suggesting that the risk score combined with other clinical information can develop a robust prediction for survival. In the future clinical practice, oncologists can use genetic detection to obtain information of expression of the 20 metabolic genes. Through combing with clinical variables mentioned above, our nomogram can accurately calculate the specific survival probability of each patient, which improves the individualized clinical decision making of GC patient.

Numerous researches have reported that the immune infiltrates in the tumour are of clinical importance [29–31]. Zeng et al. revealed that infiltrating immune cell was an independent prognostic biomarker in GC, and estimated its value in predicting chemotherapeutic and immunotherapeutic outcomes [32,33]. Therefore, we compared the differential abundance of tumour-infiltrating immune cells between the two groups and observed ten types of immune cells that were significantly different. Thereinto, plasma cells, T cells CD8+, T cells CD4 + memory activated, and macrophage M1 were more enriched in the low-risk group. Liu et al. found that high densities of T cells CD8 + and T cells CD4 + were associated with better clinical outcomes in GC [34]. Matsumoto et al. reported that high levels of CD8 + and CD4 + T cell infiltrate were correlated with better survival in triple-negative breast cancer [35]. Extensive literature has shown that tumour-infiltrating plasma cells have a positive prognostic effect for cancer [36]. Furthermore, we also observed a higher level of macrophage M2 and a lower level of macrophage M1 in the high-risk group than the low-risk group. Nevertheless, there is plenty of evidence

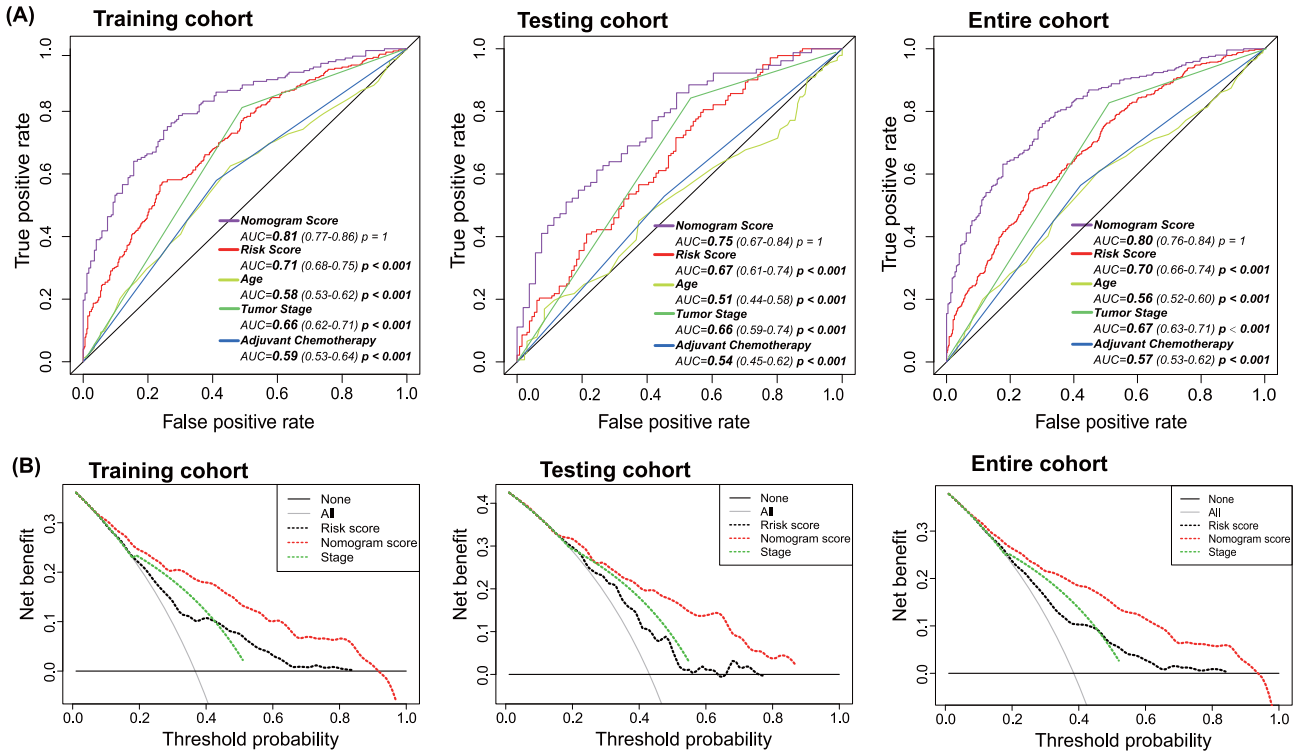

Fig. 8. Evaluation of the nomogram. (A). The ROC curves for five-year overall survival of the risk score and other clinical variables in the training, validation, and entire cohorts, respectively. (B). Decision curve analysis for the risk score, TNM staging system, and the nomogram. The solid black line represented no patients would die, and the grey line represented all patients would die.

that high infiltration of M2 macrophages in tumour and low infiltration of M1 macrophages were associated with reduced OS [37,38]. Accordingly, the above immune cell infiltration patterns may help to explain the outcome that the low-risk group had a better prognosis.

In addition, we explored the biological processes in the two groups using GSEA analysis. The results revealed that a total of six pathways were significantly enriched in the high-risk group. Thereinto, "epithelial-mesenchymal transition" (EMT) is considered integral in the development and wound healing, whereas it contributes pathologically to tumorigenesis [39]. As the canonical signalling pathway in tumorigenesis, EMT was significantly correlated with initiation, invasion, and metastasis of GC [40]. The pathway called "Angiogenesis" might be related to the metabolic alteration of glucose metabolism previously mentioned [15]. In the 2000 s, Constant et al. reported that an increased level of lactate could promote angiogenesis by inducing tumour-related stromal cells to secrete vascular endothelial growth factor (VEGF) [41]. Furthermore, overexpression of the hedgehog signalling pathway can repair the damaged gastric mucosa caused by helicobacter pylori (H pylori) infection [42]. Nevertheless, the low-risk group was primarily concentrated on MYC target and cell cycle-related pathways. Based on previous studies, MYC can drive specific metabolic pathways [43]. Activation of the cyclin-dependent kinase (CDK) can lead to the progression of the cell cycle, whereas overexpression of cyclin D1 and D2 were detected in GC [44]. In summary, these enriched pathways were mainly linked to gastric tumorigenesis and the alteration of metabolism. Exploration of underlying molecular mechanisms helps to develop novel therapeutic targets for GC.

However, our present work has several limitations. First, all of the research data are derived from a public database, which makes it difficult to collect complete clinical information for each patient. Moreover, this study involved multiple series, and the batch effect was inevitable despite applying the statistical method to reduce it. Last but not least, this was a retrospectively designed study, and the potential bias correlated with unbalanced clinicopathological features cannot be ignored. Further prospective studies and experiments are urgently needed to validate the prognostic value of metabolic genes.

5. Conclusion

In this study, for the first time, we identified a list of metabolic genes related to survival and developed a metabolism-based gene signature for GC. Through a series of bioinformatics and statistical analyses, the predictive ability of the signature was confirmed. We believe that it will lead to the discovery of a novel landscape for the therapeutic strategy of the tumour.

6. Funding statement

This research was not solicited and did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

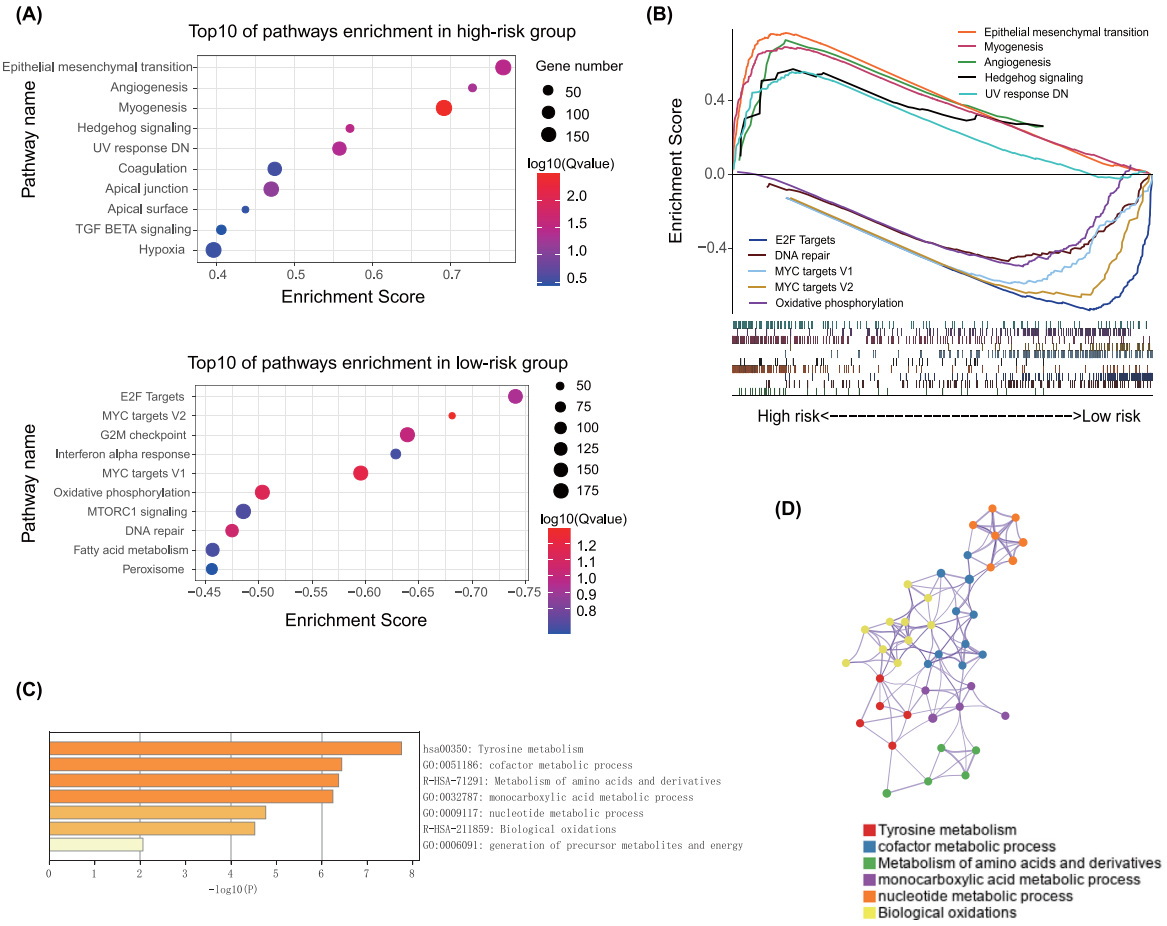

Fig. 9. Exploration of biological function. (A). Bubble plot of the top ten pathways enriched in the high- and low-risk groups; (B). Gene set enrichment analysis of the top five pathways significantly enriched in the high- and low-risk groups. (C). Bar graph of enriched pathways across 20 metabolism-related genes, coloured by p-values. (D). The network of enriched pathways, where nodes that share the same pathway are typically close to each other.

CRediT authorship contribution statement

Tianqi Luo: Investigation, Formal analysis, Writing - original draft. **Yuanfang Li:** Investigation, Data curation, Methodology. **Runcong Nie:** Investigation, Software. **Chengcai Liang:** Investigation, Validation. **Zekun Liu:** Visualization. **Zhicheng Xue:** Resources. **Guoming Chen:** Data curation. **Kaiming Jiang:** Validation, Writing - review & editing. **Ze-Xian Liu:** Project administration. **Huan Lin:** Conceptualization, Writing - review & editing. **Cong Li:** Supervision, Writing - review & editing, Writing - review & editing. **Yingbo Chen:** Conceptualization, Supervision.

Data availability Statement

The datasets generated for this study can be found in the GEO database (GSE62254, GSE84437, GSE26942, GSE13861, GSE29272; https://www.ncbi.nlm.nih.gov/geo/), and UCSC Xena website (https://gdc.xenahubs.net).

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.csbj.2020.09.037.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.
[2] Digklia A, Wagner AD. Advanced gastric cancer: Current treatment landscape and future perspectives. World J Gastroenterol 2016;22:2403–14.
[3] Kiyokawa T, Fukagawa T. Recent trends from the results of clinical trials on gastric cancer surgery. Cancer Commun (Lond) 2019;39:11.
[4] Fang C, Wang W, Deng JY, Sun Z, Seeruttun SR, et al. Proposal and validation of a modified staging system to improve the prognosis predictive performance of the 8th AJCC/UICC pTNM staging system for gastric adenocarcinoma: a multicenter study with external validation. Cancer Commun (Lond) 2018;38:67.
[5] Wu H-H, Lin W-c, Tsai K-W. Advances in molecular biomarkers for gastric cancer: miRNAs as emerging novel cancer markers. Expert Rev Mol Med 2014;16:e1.
[6] Li P, Chen S, Chen H, Mo X, Li T, et al. Using circular RNA as a novel type of biomarker in the screening of gastric cancer. Clin Chim Acta 2015;444:132–6.
[7] Jiang B, Sun Q, Tong Y, Wang Y, Ma H, et al. An immune-related gene signature predicts prognosis of gastric cancer. Medicine (Baltimore) 2019;98:e16273.
[8] Peng P-L, Zhou X-Y, Yi G-D, Chen P-F, Wang F, et al. Identification of a novel gene pairs signature in the prognosis of gastric cancer. Cancer Medicine 2018;7:344–50.
[9] Hou JY, Wang YG, Ma SJ, Yang BY, Li QP. Identification of a prognostic 5-Gene expression signature for gastric cancer. J Cancer Res Clin Oncol 2017;143:619–29.
[10] Cheng P. A prognostic 3-long noncoding RNA signature for patients with gastric cancer. J Cell Biochem 2018;119:9261–9.

[11] Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683–90.
[12] Pavlova NN, Thompson CB. The emerging hallmarks of cancer metabolism. Cell Metab 2016;23:27–47.
[13] DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv 2016;2:e1600200.
[14] Warburg O, Wind F, Negelein E. The metabolism of tumors in the body. J Gen Physiol 1927;8:519–30.
[15] O W,(1956), On respiratory impairment in cancer cells, Science, 124: 269-270.
[16] Cai Z, Zhao JS, Li JJ, Peng DN, Wang XY, et al. A combined proteomics and metabolomics profiling of gastric cardia cancer reveals characteristic dysregulations in glucose metabolism. Mol Cell Proteomics 2010;9:2617–28.
[17] Xiao S, Zhou L. Gastric cancer: metabolic and metabolomics perspectives (Review). Int J Oncol 2017;51:5–17.
[18] Adamski J, Suhre K. Metabolomics platforms for genome wide association studies—linking the genome to the metabolome. Curr Opin Biotechnol 2013;24:39–47.
[19] Liu YQ, Chai RC, Wang YZ, Wang Z, Liu X, et al. Amino acid metabolism-related gene expression-based risk signature can better predict overall survival for glioma. Cancer Sci 2019;110:321–33.
[20] Zhao S, Cai J, Li J, Bao G, Li D, et al. Bioinformatic profiling identifies a glucose-related risk signature for the malignancy of glioma and the survival of patients. Mol Neurobiol 2017;54:8203–10.
[21] Wu F, Zhao Z, Chai RC, Liu YQ, Li GZ, et al. Prognostic power of a lipid metabolism gene panel for diffuse gliomas. J Cell Mol Med 2019;23:7741–8.
[22] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118–27.
[23] Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16:385–95.
[24] Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008;8:53.
[25] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10:1523.
[26] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7.
[27] Matsuoka T, Yashiro M. Biomarkers of gastric cancer: current topics and future perspective. World J Gastroenterol 2018;24:2818–32.
[28] 2016), Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients, DOI.
[29] Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer 2012;12:298–306.
[30] Angell H, Galon J. From the immune contexture to the immunoscore: the role of prognostic and predictive immune markers in cancer. Curr Opin Immunol 2013;25:261–7.
[31] Mlecnik B, Bindea G, Kirilovsky A, Angell HK, Obenauf AC, et al. The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis. Sci Transl Med 2016;8:327ra326.
[32] Zeng D, Zhou R, Yu Y, Luo Y, Zhang J, et al. Gene expression profiles for a prognostic immunoscore in gastric cancer. Br J Surg 2018;105:1338–48.
[33] Zeng D, Li M, Zhou R, Zhang J, Sun H, et al. Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures. Cancer Immunol Res 2019. https://doi.org/10.1158/2326-6066.CIR-18-0436.
[34] Liu K, Yang K, Wu B, Chen H, Chen X, et al. Tumor-infiltrating immune cells are associated with prognosis of gastric cancer. Medicine 2015;94:e1631.
[35] Matsumoto H, Thike AA, Li H, Yeong J, Koo S-L, et al. Increased CD4 and CD8-positive T cell infiltrate signifies good prognosis in a subset of triple-negative breast cancer. Breast Cancer Res Treat 2016;156:237–47.
[36] Wouters MCA, Nelson BH. Prognostic significance of tumor-infiltrating B cells and plasma cells in human cancer. Clin Cancer Res 2018;24:6125–35.
[37] Jackute J, Zemaitis M, Pranys D, Sitkauskiene B, Miliauskas S, et al. Distribution of M1 and M2 macrophages in tumor islets and stroma in relation to prognosis of non-small cell lung cancer. BMC Immunol 2018;19:3.
[38] Xiong Y, Wang K, Zhou H, Peng L, You W, et al. Profiles of immune infiltration in colorectal cancer and their clinical significant: a gene expression-based study. Cancer Medicine 2018;7:4496–508.
[39] Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial–mesenchymal transition. Nat Rev Mol Cell Biol 2014;15:178–96.
[40] Molaei F, Forghanifard MM, Fahim Y, Abbaszadegan MR. Molecular signaling in tumorigenesis of gastric cancer. Iran Biomed J 2018;22:217–30.
[41] Constant JS, Feng JJ, Zabel DD, Yuan H, Suh DY, et al. Lactate elicits vascular endothelial growth factor from macrophages: a possible alternative to hypoxia. Wound Repair Regen 2000;8:353–60.
[42] Kim J-H, Choi YJ, Lee SH, Shin HS, Lee IO, et al. Effect of Helicobacter pylori infection on the sonic hedgehog signaling pathway in gastric cancer cells. Oncol Rep 2010;23:1523–8.
[43] Stine Z E, Walton Z E, Altman B J, Hsieh A L, Dang C V,(2015), MYC, Metabolism, and Cancer, Cancer Discov, 5: 1024-1039.
[44] Arici DS, Tuncer E, Ozer H, Simek G, Koyuncu A. Expression of retinoblastoma and cyclin D1 in gastric carcinoma. Neoplasma 2009;56:63–7.