Research Paper

Establishing a cancer driver gene signature-based risk model for predicting the prognoses of gastric cancer patients

Jun Chen[1], Chao Zhou[2], Ying Liu[3]

[1]Department of Oncology, The First Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, People's Republic of China
[2]Department of Neurology, Jiangxi Provincial People's Hospital Affiliated to Nanchang University, Nanchang 330006, Jiangxi, People's Republic of China
[3]Department of Emergency, The First Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, People's Republic of China

Correspondence to: Ying Liu; **email:** liuyingemergency@outlook.com, https://orcid.org/0000-0002-5965-2025

Keywords: gastric cancer, cancer driver gene, risk signature, nomogram, prognosis
Received: August 29, 2021 **Accepted:** February 24, 2022 **Published:** March 14, 2022

ABSTRACT

Despite the high prevalence of gastric cancer (GC), molecular biomarkers that can reliably detect GC are yet to be discovered. The present study aimed to establish a robust gene signature based on cancer driver genes (CDGs) that can predict GC prognosis. Transcriptional profiles and clinical data from GC patients were analyzed using univariate Cox regression analysis and the least absolute shrinkage and selection (LASSO)-penalized Cox regression analysis to select optimal prognosis-related genes for modeling. Time-dependent receiver operating characteristic (ROC) and Kaplan-Meier analyses were done to assess the predictive power of this gene signature. A nomogram model for prediction of survival of GC patients was established using the CDG signature and clinical information, and a seven-CDG signature was identified. Risk scores were calculated using this signature, and patients were subsequently divided into high- and low-risk groups; high-risk patients in the training and validation datasets had poorer prognoses than low-risk patients. Cox regression analysis revealed that the CDG signature is an independent prognostic factor for GC. The signature and other clinical features were used to construct a nomogram for predicting overall GC patient survival. Calibration and decision curve analysis showed that the nomogram accurately predicted survival, highlighting its clinical utility. Thus, we established a novel CDG signature and nomogram for predicting GC prognosis, which may facilitate personalized treatment of GC.

INTRODUCTION

Gastric cancer (GC) is one of the most common forms of gastrointestinal cancer and is associated with very high morbidity and mortality rates [1]. It can be histologically classified into various subtypes, including adenocarcinoma, squamous cell carcinoma, adenosquamous carcinoma, and carcinoid. Gastric adenocarcinoma accounts for 80-90% of all GC cases. In recent years, the incidence of GC has increased and GC cases have been associated with poor prognoses. Currently, surgical resection is the main option for GC treatment [2]. Therefore, identifying new therapeutic targets for GC is required [3]. Over the past few decades, several studies that focused on developing molecular targeted therapies for GC and understanding their underlying molecular mechanisms have shed light on GC pathogenesis [4]. However, despite the importance of accurate classification and risk stratification of GC patients in improving management decisions and prognosis predictions, reliable biomarkers to predict GC prognosis are lacking [5, 6].

Cancers are complicated diseases characterized by uncontrolled cellular growth, invasion, and metastasis, which are primarily caused by genetic mutations [7, 8]. These mutations are termed “drivers” due to their ability to drive tumorigenesis and confer certain selective advantages to somatic tissue cells over their neighboring cells. Mutations in cancer driver genes (CDGs) affect cellular homeostasis and numerous cellular processes. Recently, Francisco et al. reported molecular-level changes that occur during malignant tumor progression [7]. Their study represents the most comprehensive analysis performed to date, as they analyzed over 28,000 samples of 66 cancer types and revealed 568 CDGs; these results suggest the involvement of a variety of molecular mechanisms. CDGs are important factors that affect the occurrence and development of GC, and they play important roles in GC prognosis. This necessitates the development of a robust and reliable CDG signature to improve individualized survival predictions for GC.

This study aimed to build a scoring model by classifying GC patients on the basis of a CDG signature, in combination with other clinicopathological factors, to improve the ability to predict GC patient prognoses, thereby helping to guide individualized treatment. This study represents the first investigation into the clinical value of CDGs in predicting GC prognoses. CDGs are expected to become novel key GC biomarkers and open new avenues for the development of novel GC treatment methods.

RESULTS

Construction of the CDG signature for GC

Overlapping prognostic CDGs from The Cancer Genome Atlas Stomach Adenocarcinoma (TCGA-STAD) and GSE62254 databases were selected to determine the candidate CDGs. Twenty-four CDGs were identified for final analysis (Supplementary Figure 1A). Next, least absolute shrinkage and selection operator (LASSO)-penalized Cox analysis was performed to narrow down the list of CDGs, and 12 genes were identified for downstream analyses (Supplementary Figure 1B). Multivariate Cox analysis was performed based on the CDGs selected by LASSO analysis (Supplementary Figure 1C). The prognostic risk scores of the CDG signature were determined as follows:

Risk score = (-0.06570) * (expression level of Damage Specific Deoxyribose Nucleic Acid [DNA] Binding Protein 2 [DDB2]) + 0.04589 * (expression level of Aminopeptidase [ENPEP]) + 0.00243* (expression level of Guanine Nucleotide binding protein, Alpha Stimulating activity polypeptide [GNAS]) - 0.10790 * (expression level of Musashi Ribose Nucleic Acid [RNA] Binding Protein 2 [MSI2]) + 0.14947 * (expression level of myosin Va [MYO5A]) + 0.22932 * (expression level of Pleomorphic Adenoma Gene 1 [PLAG1]) - 0.18526 * (expression level of RNA Binding Motif 15 [RBM15]; Supplementary Figure 1D).

CDG expression and its mutations in GC

To study the differences in CDG expression between tumor and normal tissues, we examined CDG messenger RNA (mRNA) levels in samples from TCGA-STAD. The results revealed that tumor tissues had significantly higher expression of *DDB2*, *MSI2*, and *RBM15* than the normal tissues (Figure 1A). However, no differences were observed in the expression of *ENPEP*, *GNAS*, *MYO5A*, or *PLAG1*. We also examined the levels of proteins encoded by these CDGs using clinical samples from the Human Protein Atlas (HPA) database and found that the differences in protein levels were consistent with the observed differences in mRNA levels (Figure 1B).

The prognostic value of the CDG signature

To evaluate the prognostic value of the CDG signature in the training set, the GC patients were divided into high-risk (n = 167) and low-risk (n= 167) groups according to the median risk score. The corresponding signature risk score survival statuses were ranked and displayed on a dot-plot (Figure 2A, 2B). Individuals exhibited a greater risk of mortality with increasing risk score (Figure 2C). Heatmaps of the seven prognostic CDGs are displayed in Figure 2D.

Kaplan-Meier analysis revealed that in the training set, patients in the high-risk group had shorter overall survival (OS) than those in the low-risk group (P < 0.001) (Figure 2E). Time-dependent receiver operating characteristic (ROC) analysis demonstrated that the area under the curve (AUC) values for one year, three years, and five years were 0.712, 0.613, and 0.611, respectively (Figure 2F). We also compared the prognostic value to traditional clinicopathological predictors. Time-dependent ROC analyses for grade-based and stage-based prediction of OS (TCGA) are shown in Figure 2G, 2H. Univariate and multivariate Cox regression analyses showed that age and risk score were both independent prognostic factors for GC (Figure 3A, 3B).

Validation of the CDG signature

To validate the CDG signature, patients in the validation set were divided into high- and low-risk groups, according to the median risk score (calculated using the CDG signature). The results were compatible with those obtained in the training set derived from TCGA. Figure 4A shows the heatmap of the seven

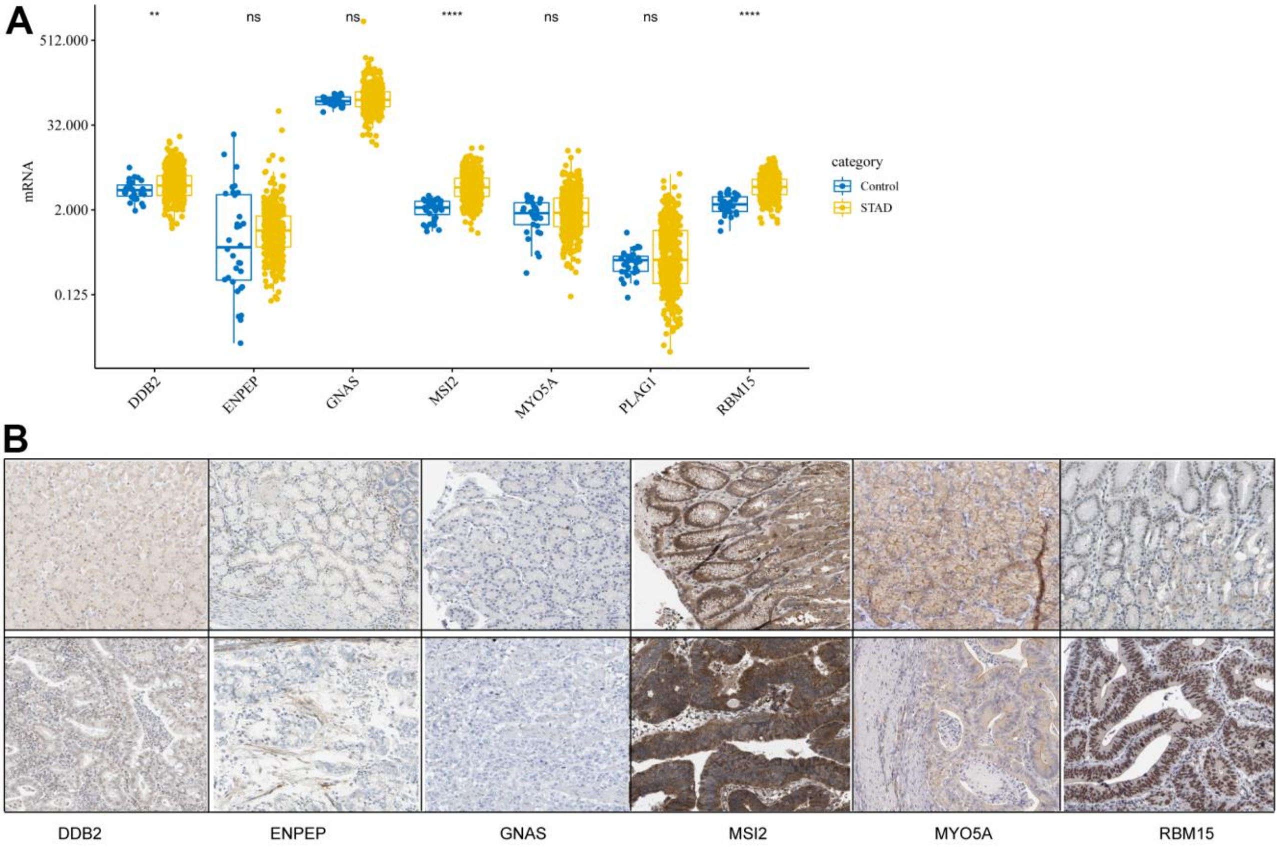


Figure 1. Expression levels of cancer driver genes (CDGs) and their alterations in gastric cancer (GC). (**A**) CDG mRNA expression levels in GC obtained from The Cancer Genome Atlas Stomach Adenocarcinoma (TCGA-STAD). (**B**) Expression levels of proteins encoded by CDGs in normal tissues as obtained from the Human Protein Atlas (HPA) database (Data for *GLAP1* was not available at HPA database).

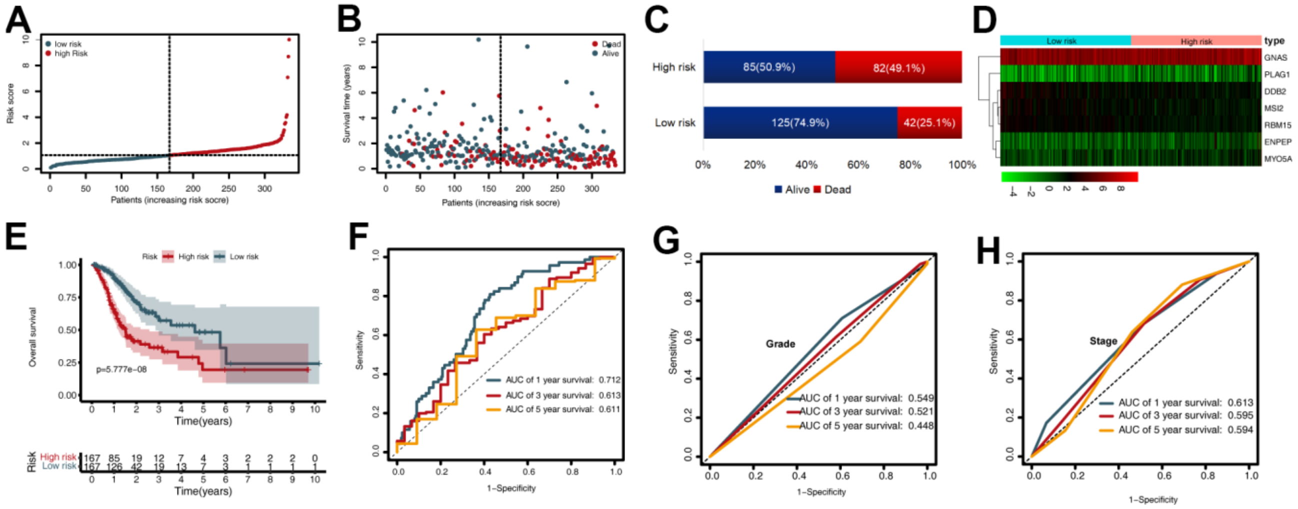


Figure 2. Prognostic value of the cancer driver gene (CDG) signature in The Cancer Genome Atlas (TCGA) training set. (**A**) Distribution of risk scores per patient. (**B**) Relationships between overall survival (OS) status and survival time of gastric cancer (GC) patients ranked on the basis of risk score. (**C**) Comparison of mortality risk between the two groups in TCGA cohort. (**D**) Heatmap representing the expression profiles of the seven CDGs. (**E**) Kaplan-Meier analysis of OS between high- and low-risk groups in TCGA set. (**F**) Time-dependent receiver operating characteristic (ROC) analysis for OS prediction in TCGA set. (**G**) Time-dependent ROC analysis for grade prediction in TCGA set of OS. (**H**) Time-dependent ROC analysis for stage prediction in TCGA set of OS.

prognostic CDGs. The corresponding signature risk score survival statuses were ranked and are displayed as a dot-plot (Figure 4A–4C). Individuals were at greater risk of mortality and recurrence with increasing risk scores (Figure 4E, 4F). Kaplan-Meier analysis also showed that the high-risk group had shorter OS and disease-free survival (DFS) than the low-risk group (*P* < 0.001; Figure 4G, 4H). Time-dependent ROC

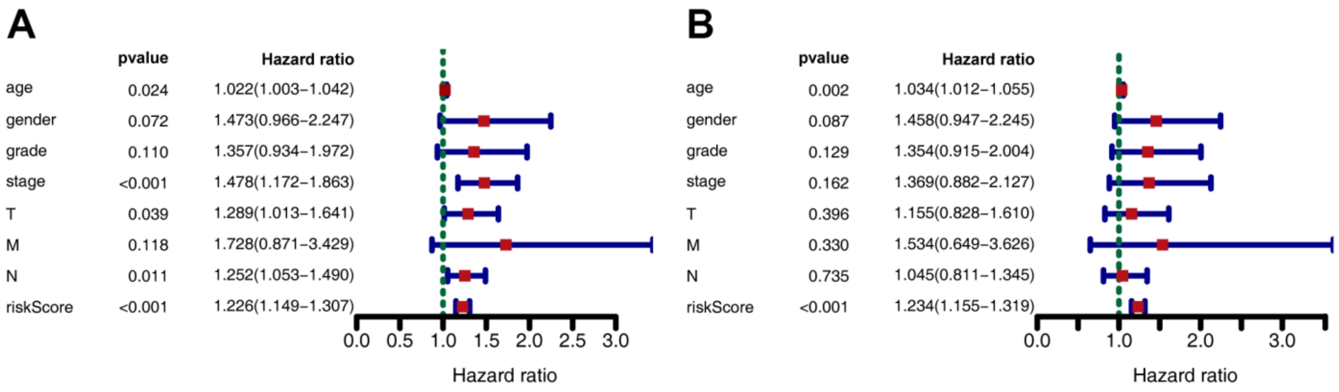


Figure 3. Forest plot depicting associations between risk factors and other clinical features and the prognosis of gastric cancer. (**A**) Univariate Cox regression analysis. (**B**) Multiple Cox regression analysis.

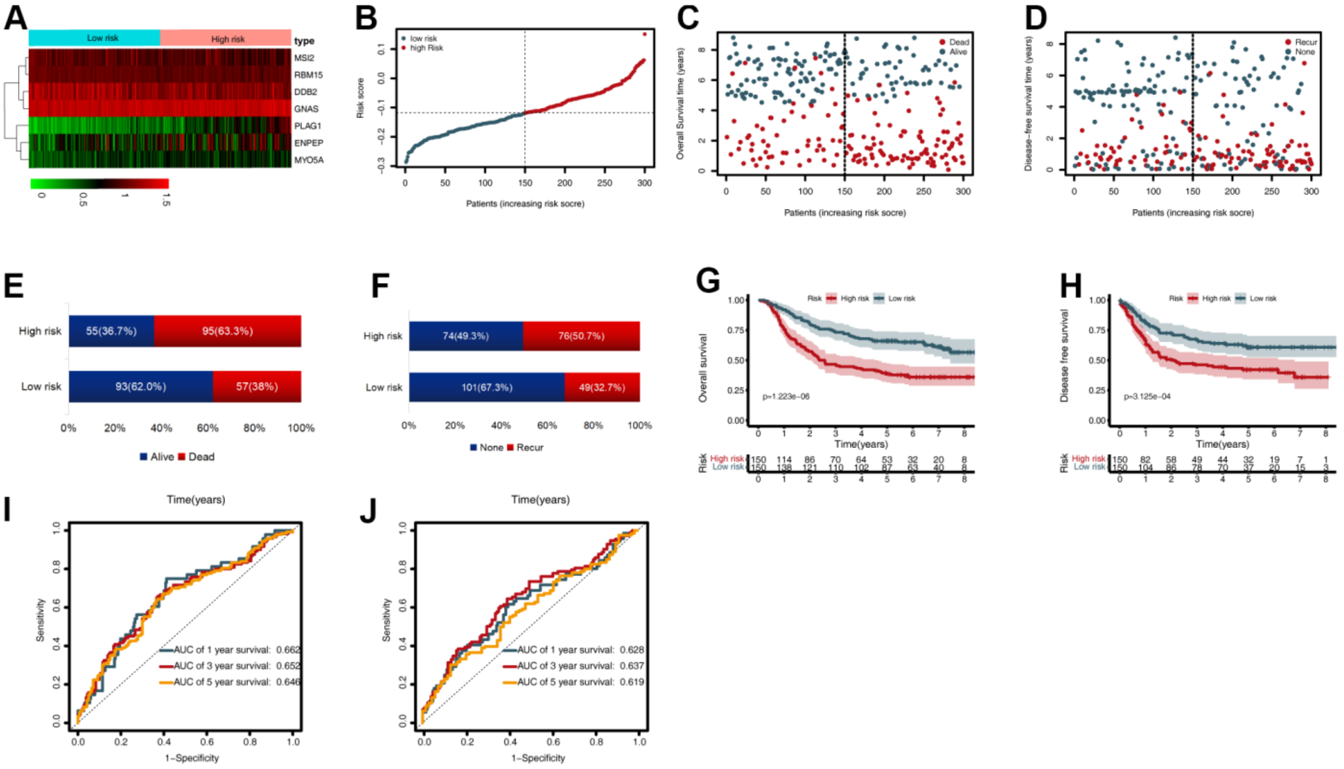


Figure 4. External validation of the cancer driver gene (CDG) signature using the gastric cancer (GC) data from the Gene Expression Omnibus (GEO) validation set. (**A**) Heatmap representing expression profiles of the seven CDGs. (**B**) Distribution of risk scores per patient. (**C**) Relationships between overall survival (OS) status and survival time in GC patients ranked by risk score. (**D**) Relationships between disease-free survival (DFS) status and survival time in GC patients ranked by risk score. (**E**) Comparison of OS risk between the two groups. (**F**) Comparison of DFS risk between the two groups. (**G**) Kaplan-Meier analysis of OS between high- and low-risk groups in GSE62254. (**H**) Kaplan-Meier analysis of DFS between high- and low-risk groups in GSE62254. (**I**) Time-dependent receiver operating characteristic (ROC) analysis for OS prediction in the GSE62254 cohort. (**J**) Time-dependent ROC analysis for DFS prediction in the GSE62254 cohort.

analysis of OS showed that the 1-, 3-, and 5-year AUC values were 0.662, 0.652, and 0.646 respectively (Figure 4I). The AUC values of one-, three-, and five-year DFS were 0.628, 0.637, and 0.619, respectively (Figure 4J).

Subgroup analysis of the CDG signature

To further estimate the utility of the CDG signature in predicting survival outcomes, stratification analysis was conducted based on specific clinicopathological characteristics. These subgroups included age (<65 or ≥65 years), gender, grade, stage, and T, N, and M stages (Figure 5A–5P). Stratification of the training and validation datasets revealed that the CDG signature could categorize patients into different survival groups and provide statistically significant prognostic values (Tables 1, 2).

Gene set enrichment analysis (GSEA)

To further analyze the functions of the seven CDGs identified, GSEA was conducted for the high- and low-risk patients in the four datasets. GSEA results for

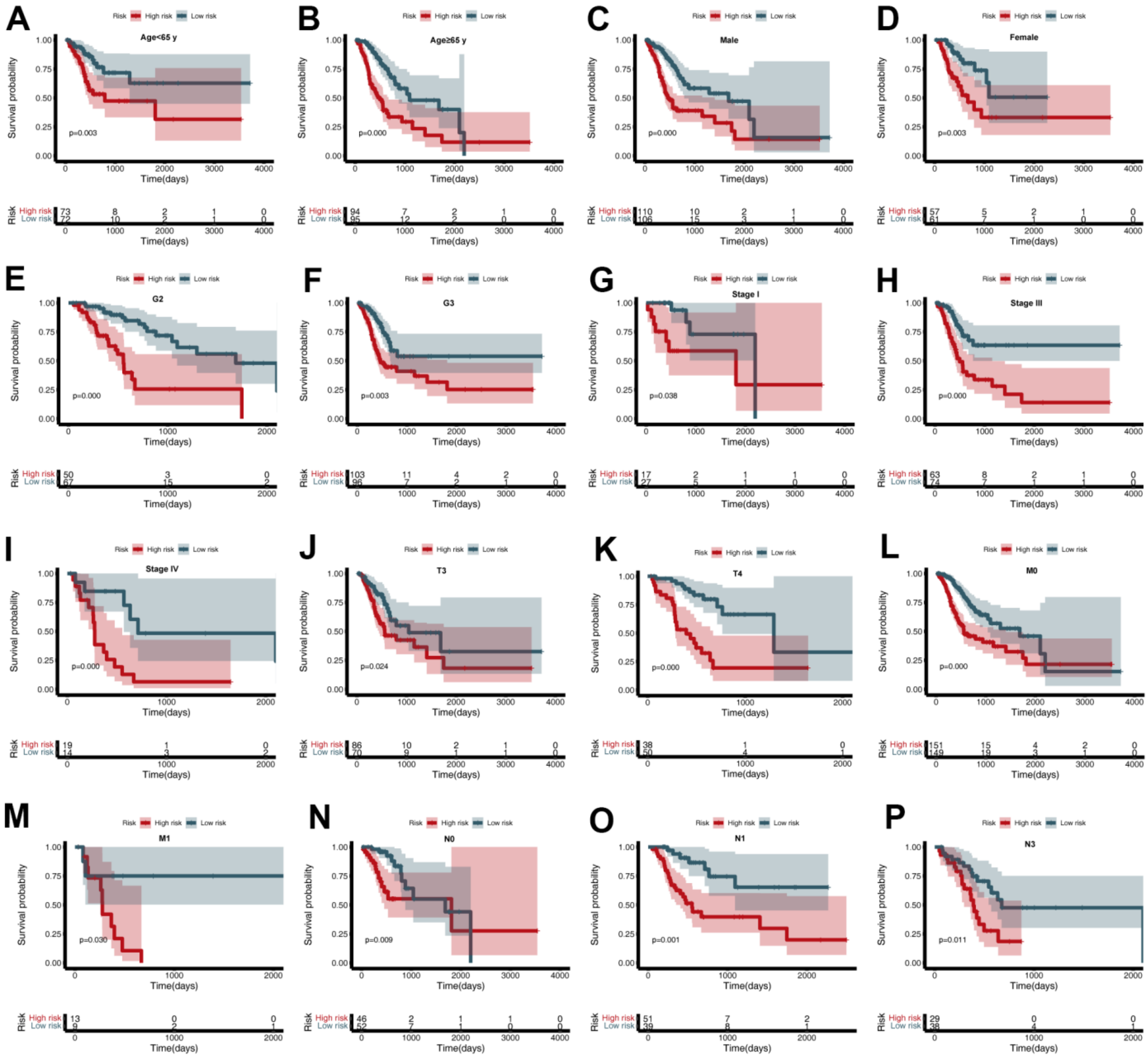


Figure 5. Subgroup analysis of the cancer driver gene (CDG) signature. (**A**) Age < 65 years, (**B**) Age ≥ 65 years, (**C**) Male, (**D**) Female, (**E**) G2, (**F**) G3, (**G**) Stage I, (**H**) Stage III, (**I**) Stage IV, (**J**) T3, (**K**) T4, (**L**) M0, (**M**) M1, (**N**) N0, (**O**) N1, and (**P**) N3.

Table 1. Stratified survival analyses based on clinical characteristics and CDG signature in TCGA-STAD cohort.

Characteristics	Number		%	Overall survival	
	High-risk	Low-risk		HR (95% CI)	P value
Age (years)					
< 65	73	72	43.4%	2.627(1.378-5.007)	0.003
≥ 65	94	95	56.6%	2.791(1.761-4.424)	0.000
Sex					
Male	110	106	64.7%	2.623(1.682-4.090)	0.000
Female	57	61	35.3%	1.705(1.204-2.415)	0.003
Grade					
G1	8	1	2.7%	-	-
G2	50	67	35.0%	2.115(1.518-2.947)	0.000
G3	103	96	59.6%	1.554(1.219-1.981)	0.000
Unknown	6	3	2.7%	-	-
Stage					
I	17	27	13.2%	1.940(1.037-3.629)	0.038
II	58	48	31.7%	1.225(0.844-1.777)	0.286
III	63	74	41.0%	1.717(1.290-2.284)	0.000
IV	19	14	9.9%	2.111(1.251-3.563)	0.005
Unknown	10	4	4.2%	-	-
T stage					
T1	2	12	4.2%	-	-
T2	37	35	21.6%	1.425(0.944-2.150)	0.092
T3	86	70	46.7%	1.359(1.041-1.774)	0.024
T4	38	50	26.3%	2.228(1.539-3.225)	0.000
Unknown	4	0	1.2%	-	-
M stage					
M0	151	149	89.8%	1.576(1.290-1.925)	0.000
M1	13	9	6.6%	2.360(1.087-5.122)	0.030
Unknown	3	9	3.6%	-	-
N stage					
N0	46	52	29.3%	1.708(1.146-2.544)	0.009
N1	51	39	26.9%	2.044(1.346-3.105)	0.001
N2	32	36	20.4%	1.467(0.974-2.210)	0.067
N3	29	38	20.1%	1.587(1.113-2.263)	0.011
Unknown	9	2	3.3%	-	-

HR, hazard ratio; CI, confidence interval.

the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways indicated that the "calcium signaling pathway," "cell adhesion molecules," (CAMs) "extracellular matrix receptor interaction," "focal adhesion," and "gap junction" categories were highly enriched in the high-risk group (Figure 6A). GSEA results for Gene Ontology (GO) terms indicated that the "collagen-containing extracellular matrix," "contractile fiber," "glycosaminoglycan binding," "hormone binding," and "muscle system processes" were highly enriched in the high-risk group (Figure 6B).

Analysis of tumor immunity

TCGA-STAD gene expression matrix was uploaded to the Cell type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) platform to estimate the proportions of the 22 immune cell types. The high-risk group had high proportions of activated

Table 2. Stratified survival analyses based on clinical characteristics and CDG signature in the GSE62254 cohort.

Characteristics	Number		%	Overall survival		Disease-free survival	
	High-risk	Low-risk		HR (95% CI)	*P* value	HR (95% CI)	*P* value
Age (years)							
< 65	81	80	53.7%	2.431(1.491-3.963)	0.000	2.201(1.318-3.675)	0.003
≥ 65	69	70	46.3%	1.443(1.154-1.805)	0.001	1.299(1.007-1.675)	0.044
Sex							
Male	103	96	66.3%	1.260(1.032-1.538)	0.023	1.155(0.928-1.436)	0.197
Female	47	54	33.7%	2.074(1.538-2.796)	0.000	2.019(1.454-2.805)	0.000
Lauren pathological classification							
Intestinal	59	87	48.7%	1.450(1.124-1.869)	0.004	1.327(1.003-1.757)	0.048
Diffuse	82	53	45.0%	1.476(1.157-1.884)	0.002	1.388(1.068-1.805)	0.014
Mixed	9	10	6.3%	1.251(0.722-2.170)	0.425	1.062(0.548-2.058)	0.859
Stage							
I	7	23	10.0%	0.937(0.313-2.805)	0.908	1.272(0.383-4.227)	0.694
II	41	56	32.3%	1.638(1.140-2.352)	0.008	1.156(0.760-1.760)	0.498
III	56	40	32.0%	1.341(1.008-1.784)	0.044	1.278(0.932-1.754)	0.128
IV	46	31	25.7%	1.263(0.975-1.637)	0.077	1.228(0.932-1.619)	0.145
T stage							
T1+T2	75	113	62.7%	1.510(1.202-1.896)	0.000	1.370(1.056-1.777)	0.018
T3+T4	75	37	37.3%	1.217(0.953-1.556)	0.116	1.091(0.843-1.411)	0.509
M Stage							
M0	132	141	91.0%	1.441(1.205-1.722)	0.000	1.348(1.109-1.638)	0.003
M1	18	9	9.0%	4.708(1.538-14.411)	0.007	2.186(0.702-6.803)	0.177
N Stage							
N0	12	26	12.7%	1.123(0.561-2.246)	0.744	1.301(0.615-2.752)	0.491
N1+N2+N3	138	124	87.3%	1.473(1.240-1.749)	0.000	1.345(1.116-1.620)	0.002

HR, hazard ratio; CI, confidence interval.

natural killer (NK) cells, monocytes, M2 macrophages, resting dendritic cells, and resting mast cells (Figure 7A). The low-risk group had high proportions of CD8 T-cells, CD4 memory-activated T-cells, follicular helper T-cells, resting NK cells, and M1 macrophages (Figure 7B).

Establishment and evaluation of the nomogram

To predict the survival of GC patients, we constructed a nomogram based on the training set, which included the CDG signature risk score, age, gender, and pathological stage (Figure 8A). Time-dependent ROC analysis was performed to assess the predictive accuracy of the nomogram. Plotting the one-, three-, and five-year ROC values of OS in the training set of the nomogram revealed AUC values of 0.696, 0.639, and 0.632, respectively (Figure 8B). Time-dependent ROC analyses from the nomogram for one-, three- and five-year OS probabilities in the validation set returned AUC values of 0.825, 0.784, and 0.767, respectively (Figure 8C). In addition, ROC analysis of the DFS predictions in the validation set revealed that the nomogram was highly discriminatory, with AUC values of 0.825, 0.784, and 0.767 for one-, three-, and five-year DFS levels, respectively (Figure 8D).

Calibration and decision curve analysis (DCA) revealed the reliability of the nomogram for predicting prognoses in the training and validation sets. The calibration plot revealed that the predictions made using the nomogram were consistent with the true observations (Supplementary Figure 2A–2C). The DCA curves for the predictive nomogram revealed that it had high net benefit (Supplementary Figure 2D–2F). The web-based calculator (https://prognosis.shinyapps.io/STAD/) could predict the OS of GC patients based on the established nomogram (Supplementary Figure 3A, 3B) and is convenient in terms of its usage and visualization of the prognostic nomogram.

DISCUSSION

Prognostic outcomes of GC patients are highly variable. Thus, there is an urgent need to find new GC biomarkers and to construct new prognostic models for predicting GC survival in order to develop personalized treatment plans [9]. In this study, we developed a prognostic CDG signature and a corresponding nomogram for predicting GC patient survival. We have developed a promising tool for predicting GC outcomes and guiding personalized GC therapy.

Altering the expression of CDGs can increase cell proliferation and survival, leading to clonal expansion and tumor growth [10]. Different cancer types may be associated with both common and specific driver genes, and different genes may play different roles in various cancer types. Here, we developed a seven-CDG prognostic signature based on *DDB2*, *ENPEP*, *GNAS*, *MSI2*, *MYO5A*, *PLAG1*, and *RBM15*. DDB2 was originally identified as a novel tumor suppressor via nucleotide excision repair [11], and it is abnormally expressed in several tumor tissues [12–15]. However, increasing evidence suggests that DDB2 exhibits dual functions in cancer cell proliferation. Qiao et al. reported that DDB2-silencing inhibits proliferation and migration of GC cells [16]. ENPEP is an essential and highly specific proangiogenic enzyme. ENPEP functions in tumor proliferation, migration, and drug resistance in breast and colorectal cancers [17, 18]. *GNAS* is a complex gene locus that gives rise to multiple translated and non-translated gene products [19–22]. At present, few studies have investigated the role of *GNAS* in GC, thus future studies should systematically elucidate its functions. MSI2 is a member of the Musashi family of RNA-binding proteins, which are overexpressed in various tumors, including ovarian, pancreatic, bladder, and lung cancers [23–26]. MSI2 overexpression is correlated with poor prognoses of liver and pancreatic

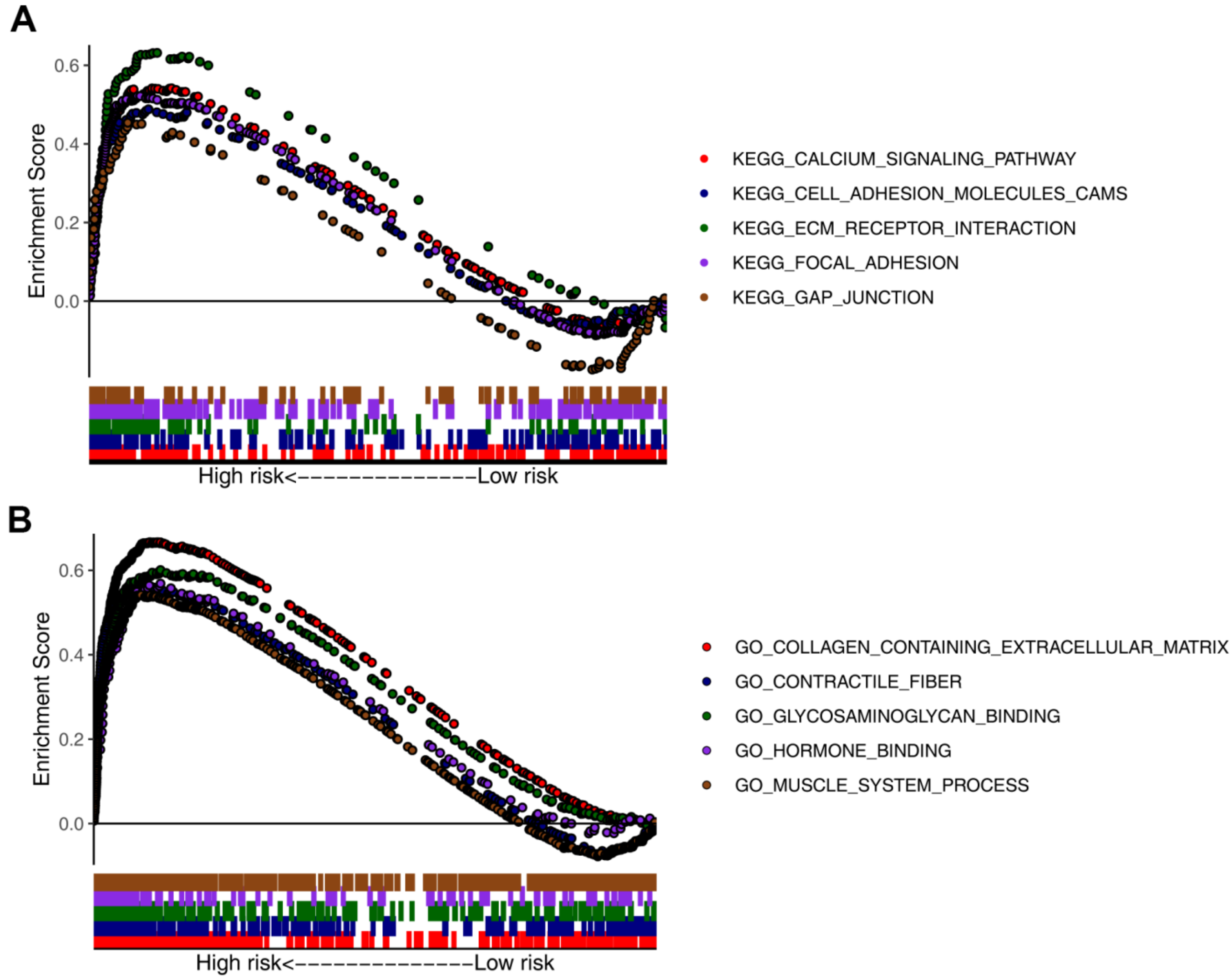


Figure 6. Gene set enrichment analysis (GSEA) of high- and low-risk groups. Top five representatives from (**A**) Gene Ontology (GO) term enrichment analysis and (**B**) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis.

cancer patients [27, 28]. Early studies on MYO5A focused on its roles in neuron formation and function, and in neurological disease. However, the functions and clinical significance of MYO5A in GC remain unclear. Recent studies have reported that MYO5A plays a role in tumorigenesis. Zhao et al. reported that serum MYO5A levels are a valuable predictor of cervical nodal occult metastasis and can be used to assess prognosis [29]. PLAG1 is a transcription factor involved in various cancers, such as lipoblastoma, hepatoblastoma, acute myeloid leukemia, uterine leiomyoma, and leiomyosarcoma [30]. RBM15, which is a member of the split ends family of proteins, determines cell-fate in many tissues including blood and is overexpressed in hepatocellular carcinoma [31]. While the studies highlighted above have revealed the functions of these CDGs in other cancers, few studies have investigated the roles of these CDGs in GC tumorigenesis.

While several studies have focused on the functions of CDGs, systematic analysis of their prognostic potential for GC is still required. The CDG signature identified in

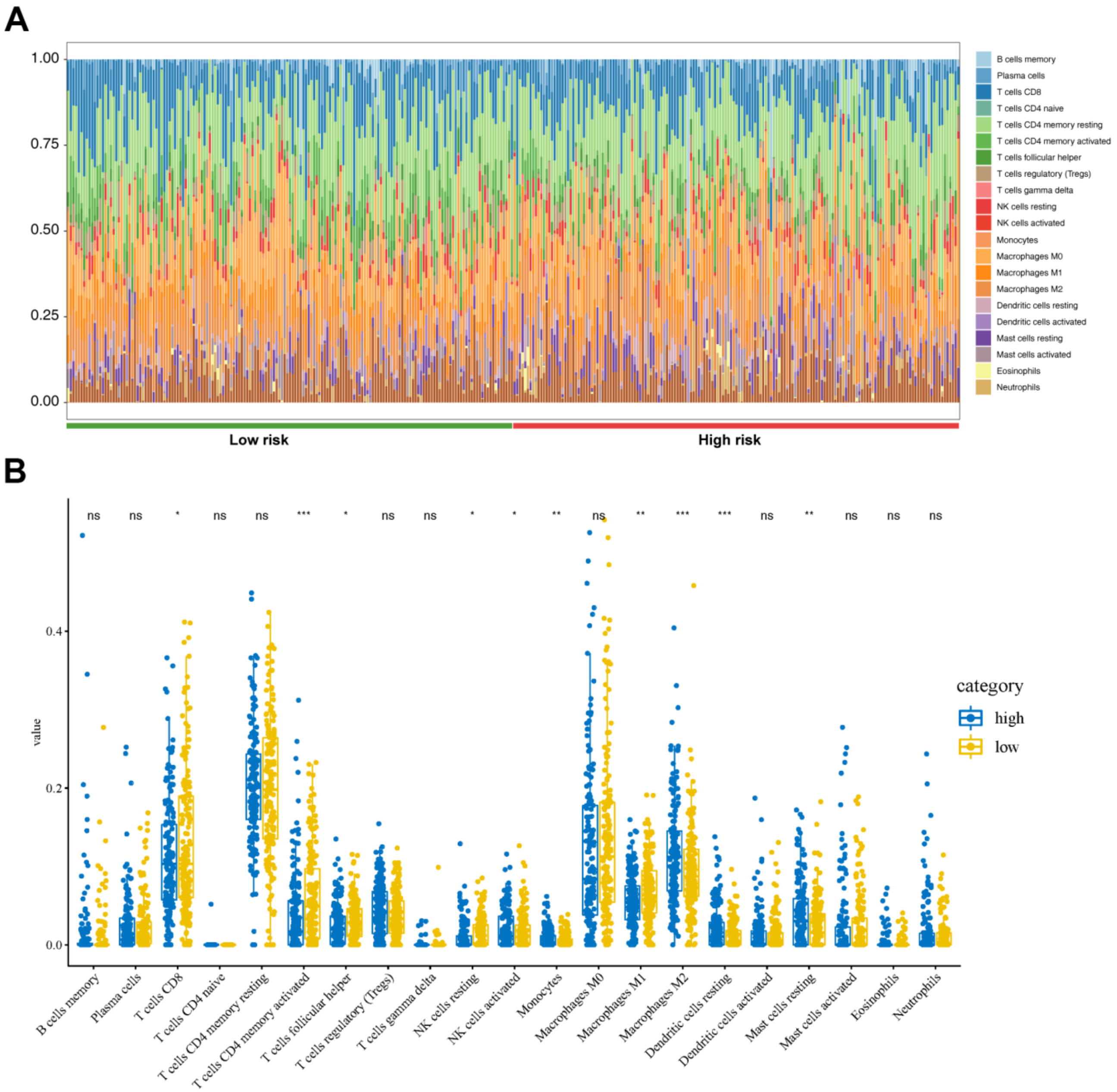


Figure 7. Tumor immunity analysis based on the CDG signature. (**A**) Relative proportion of immune cells between high- and low-risk groups. (**B**) Violin plot depicting differences in the abundances of 22 types of immune cells between the high- and low-risk groups.

our study was significantly associated with the survival of GC patients, and this association remained significant after controlling for clinical and pathological features. We constructed a nomogram for predicting one-, three- and five-year OS values for GC based on this CDG signature, age, gender, and stage. ROC analysis, calibration plots, and DCA were used to verify the prognostic accuracy of the model, and the results showed that this model had strong predictive ability. We also created a simple online tool to perform this analysis in clinical settings.

To further our understanding of the mechanisms associated with this CDG signature, GSEA was conducted to compare the low- and high-risk groups. Terms and categories such as “calcium signaling pathway,” “CAMs,” “extracellular matrix receptor interaction,” “focal adhesion,” “gap junction,” “collagen-containing extracellular matrix,” “contractile fiber,” “glycosaminoglycan binding,” “hormone binding,” and “muscle system processes” were highly enriched in the high-risk group, indicating that the seven CDGs are involved in these signaling pathways in GC.

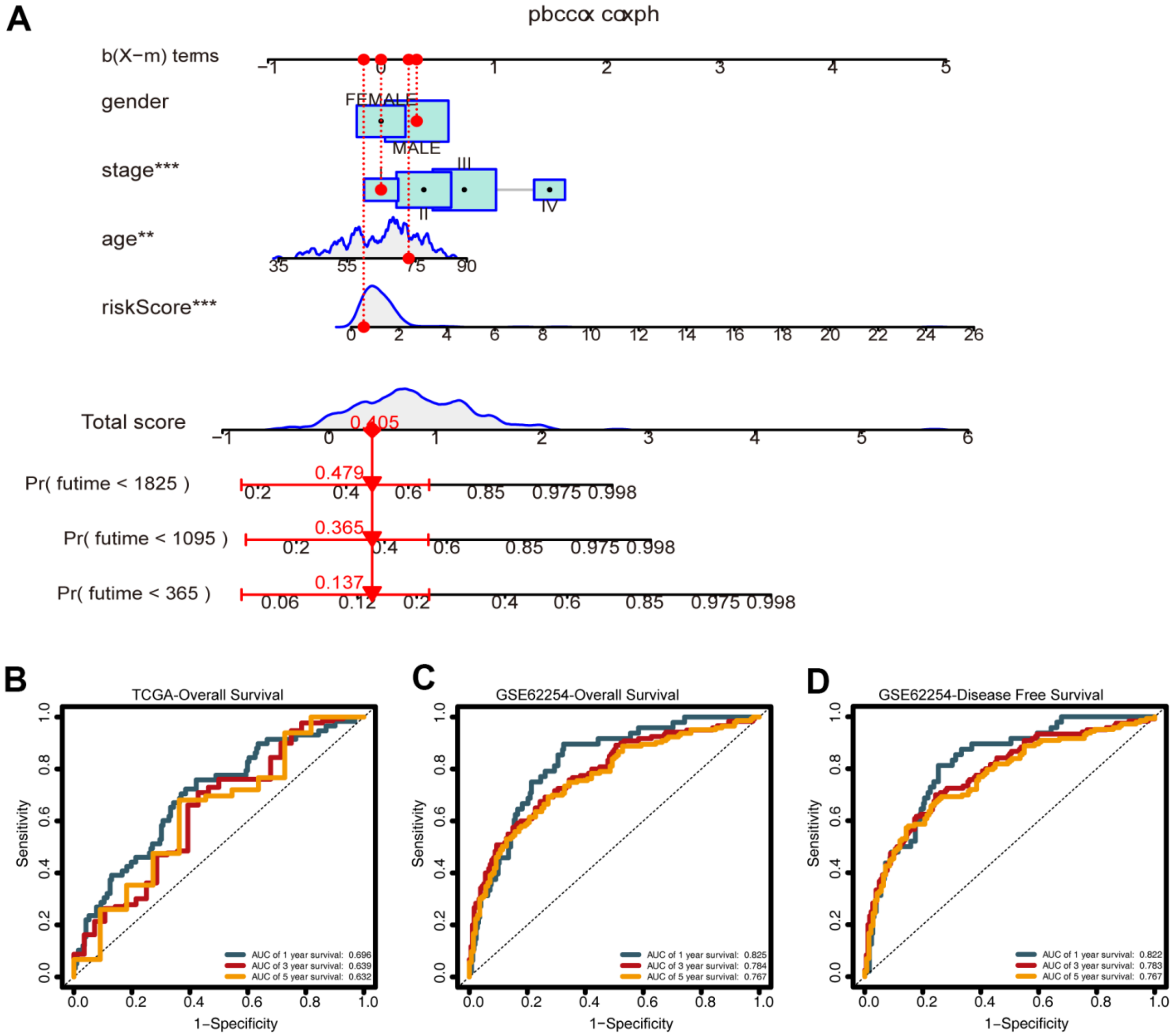


Figure 8. Construction of the nomogram. (**A**) A nomogram for predicting one-, three-, and five-year overall survival (OS) generated by integrating the risk score, age, gender, and stage. (**B**) Time-dependent receiver operating characteristic ROC curves of the nomogram for OS prediction from the training set. (**C**) Time-dependent ROC curves of the nomogram for OS prediction from the validation set. (**D**) Time-dependent ROC curves of the nomogram for DFS predictions from the validation set.

There is evidence that CDGs are closely related to tumor cell immune infiltration. In non-small cell lung cancer, GNAS promotes migration and invasion of cancer cells by altering macrophage polarization [32]. However, studies focusing on the role of the seven CDGs, identified in this study, in immune infiltration remain limited. In the present study, CIBERSORT was used to calculate the proportions of 22 immune cell subsets in GC, revealing that the high-risk group had high proportions of activated NK cells, monocytes, M2 macrophages, resting dendritic cells, and resting mast cells. These findings provide insight into the mechanisms associated with these CDGs in GC.

Many previous studies have constructed prognostic gene signatures. Chen et al. constructed a stemness index-related signature for GC with an AUC value of 0.688 [33]. Ren et al. reported angiogenesis-related gene expression signatures for predicting DFS in GC patients with an AUC value of 0.673 [34]. ROC analysis in the validation group for the nomogram in our study (0.825, 0.784, and 0.767 for OS and 0.822, 0.783, and 0.767 for DFS) indicated that the prognostic index is a stable predictor for the prognosis of GC patients. We also performed ROC analysis with traditional clinicopathological predictors (stag and grade), which demonstrated that our CDG signature has high prognostic value compared to that of these predictors.

However, our study has the following limitations. First, the data used in this study were obtained from two different databases and a non-database case was not used for external verification. Second, we did not investigate the mechanisms underlying these CDGs in GC. Third, the levels of immune cell infiltration were calculated based on algorithmic evaluations and, thus, require experimental validation. Therefore, further genetic and experimental studies with larger sample sizes and experimental validation are needed.

In summary, this is the first study to identify and validate a CDG signature that could independently predict the OS and DFS of GC patients. A prognostic nomogram was constructed by integrating age, sex, and TNM stage, which performed well in predicting the survival of GC patients. Our study, thus, generated a clinically useful tool for improving prognostic management of GC.

MATERIALS AND METHODS

Data collection and processing

Publicly available transcriptomic and clinical data associated with GC samples were obtained from TCGA (https://tcga-data.nci.nih.gov/tcga/) and the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/gds/) and analyzed retrospectively. We used the RNA-Seq fragments per kilobase of transcript per million mapped reads (FPKM) data from TCGA. After excluding cases with follow-up times of <30 days, 634 patients were enrolled in the study, including 334 patients from TCGA-STAD project and 300 from the GSE62254 cohort [35]. Data from the GSE62254 cohort were obtained from the Asian Cancer Research Group (ACRG) Gastric cohort, which included 199 male and 91 female GC patients. The median age was 64 years and the range was 24-86 years.

Data for 568 CDGs (Supplementary Table 1) were downloaded from the Integrative OncoGenomics (IntOGen) pipeline (https://www.intogen.org/search) [7]. The immunohistochemical data associated with proteins encoded by each CDG in GC and normal tissues were obtained from the HPA (https://www.proteinatlas.org/).

Construction of the CDG signature

To narrow down the screening range, overlapping prognostic CDGs were selected from TCGA-STAD and GSE62254 cohorts via Cox univariate analysis. TCGA-STAD and GSE62254 were then used for model training (n = 334) and validation (n = 300), respectively. Previously selected CDGs were further screened and confirmed by LASSO Cox regression analysis (with the penalty parameter estimated by 10-fold cross-validation) using the “glmnet” package. A formula was developed using the CDG signature constructed above, where β corresponds to the correlation coefficient:

Risk score = $\beta 1 \times$ (expression of RNA1) + $\beta 2 \times$ (expression of RNA2) + $\cdots$ + $\beta n \times$ (expression of RNAn)

The patients in each dataset were assigned to a high- or a low-risk group using the median risk score as a cutoff. The ROC curves were created using the “survivalROC” package, and the AUC values were calculated to evaluate the predictive potential of the CDG signature.

Validation of the CDG signature

To validate the CDG signature, the patients in the validation set were separated into high- or low-risk groups according to the median risk score, which was calculated according to the CDG signature. Kaplan-Meier curve and time-dependent ROC analyses were conducted to assess CDG signature categorization.

Subgroup analysis of the CDG signature

To validate the effectiveness of the prognostic CDG signature, stratification analysis was performed on the training and validation sets using different demographic and clinical characteristics. The GC cases were divided into two risk groups according to their characteristics and risk scores, and Cox regression analysis was performed to analyze differences between the subgroups.

Estimation of immune cell infiltration

To analyze the relationship between the CDG signature and immune cell characteristics, CIBERSORT was used to estimate the fractions of immune cell types between the high- and low-risk groups [36]. Statistical analysis of the proportions of 22 immune cell types in each of the 334 GC samples was performed using the Wilcoxon rank-sum test.

GSEA

GSEA was performed to explore the GO terms and KEGG pathways that were significantly enriched in high-risk GC samples (http://www.broadinstitute.org/gsea). Gene sets were considered significantly enriched when FDR < 0.05 and |NES| > 1.

Construction and evaluation of the nomogram

We designed a novel nomogram model containing the CDG signature and clinicopathological predictors to establish a quantitative clinical tool to monitor and predict outcomes of GC patients. Subsequently, we developed a web-based calculator based on this model for clinical applications. Time-ROC curves and calibration plots were generated, and DCA was performed to evaluate the clinical utility of the novel nomogram.

Statistical analysis

All statistical analyses were performed using R software (version 4.0.5, R Development Core Team, 2021) and GraphPad Prism (version 8.3.0, GraphPad software, Inc., San Diego, CA, USA). All statistical tests (two-tailed) with $P < 0.05$ were considered statistically significant.

AUTHOR CONTRIBUTIONS

Liu Ying conceived and designed the study. Chen Jun performed literature searches. Liu Ying and Zhou Chao generated the figures and tables. Chen Jun analyzed the data. Chen Jun wrote the manuscript, and Liu Ying critically reviewed the manuscript.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

This study was supported by The Project of Jiangxi Education Department (No. GJJ200223 and No. GJJ170122).

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. 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. https://doi.org/10.3322/caac.21492 PMID:30207593

2. Akce M, Jiang R, Alese OB, Shaib WL, Wu C, Behera M, El-Rayes BF. Gastric squamous cell carcinoma and gastric adenosquamous carcinoma, clinical features and outcomes of rare clinical entities: a National Cancer Database (NCDB) analysis. J Gastrointest Oncol. 2019; 10:85–94. https://doi.org/10.21037/jgo.2018.10.06 PMID:30788163

3. Tazawa H, Suzuki T, Saito A, Ishikawa A, Komo T, Sada H, Shimada N, Hadano N, Onoe T, Sudo T, Shimizu Y, Kuraoka K, Tashiro H. Utility of TMPRSS4 as a Prognostic Biomarker and Potential Therapeutic Target in Patients with Gastric Cancer. J Gastrointest Surg. 2022; 26:305–13. https://doi.org/10.1007/s11605-021-05101-2 PMID:34379296

4. Douchi D, Yamamura A, Matsuo J, Melissa Lim YH, Nuttonmanit N, Shimura M, Suda K, Chen S, Pang S, Kohu K, Abe T, Shioi G, Kim G, et al. Induction of gastric cancer by successive oncogenic activation in the corpus. Gastroenterology. 2021; 161:1907–23.e26. https://doi.org/10.1053/j.gastro.2021.08.013 PMID:34391772

5. Qin Y, Deng Y, Jiang H, Hu N, Song B. Artificial Intelligence in the Imaging of Gastric Cancer: Current Applications and Future Direction. Front Oncol. 2021; 11:631686. https://doi.org/10.3389/fonc.2021.631686 PMID:34367946

6. Tsuburaya A, Guan J, Yoshida K, Kobayashi M, Yoshino S, Tanabe K, Yoshikawa T, Oshima T, Miyashita Y, Sakamoto J, Tanaka S. Clinical biomarkers in adjuvant chemotherapy for gastric cancer after D2 dissection by a pooled analysis of individual patient data from large

randomized controlled trials. Gastric Cancer. 2021; 24:1184–93.
https://doi.org/10.1007/s10120-021-01228-y
PMID:34365541

7. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, Gonzalez-Perez A, Lopez-Bigas N. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020; 20:555–72.
https://doi.org/10.1038/s41568-020-0290-x
PMID:32778778

8. Huang D, Sun W, Zhou Y, Li P, Chen F, Chen H, Xia D, Xu E, Lai M, Wu Y, Zhang H. Mutations of key driver genes in colorectal cancer progression and metastasis. Cancer Metastasis Rev. 2018; 37:173–87.
https://doi.org/10.1007/s10555-017-9726-5
PMID:29322354

9. Kim W, Kim SJ. Heat Shock Factor 1 as a Prognostic and Diagnostic Biomarker of Gastric Cancer. Biomedicines. 2021; 9:586.
https://doi.org/10.3390/biomedicines9060586
PMID:34064083

10. Takeda H. A Platform for Validating Colorectal Cancer Driver Genes Using Mouse Organoids. Front Genet. 2021; 12:698771.
https://doi.org/10.3389/fgene.2021.698771
PMID:34262603

11. Scrima A, Konícková R, Czyzewski BK, Kawasaki Y, Jeffrey PD, Groisman R, Nakatani Y, Iwai S, Pavletich NP, Thomä NH. Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex. Cell. 2008; 135:1213–23.
https://doi.org/10.1016/j.cell.2008.10.045
PMID:19109893

12. Bagchi S, Raychaudhuri P. Damaged-DNA Binding Protein-2 Drives Apoptosis Following DNA Damage. Cell Div. 2010; 5:3.
https://doi.org/10.1186/1747-1028-5-3
PMID:20205757

13. Chen HH, Fan P, Chang SW, Tsao YP, Huang HP, Chen SL. NRIP/DCAF6 stabilizes the androgen receptor protein by displacing DDB2 from the CUL4A-DDB1 E3 ligase complex in prostate cancer. Oncotarget. 2017; 8:21501–15.
https://doi.org/10.18632/oncotarget.15308
PMID:28212551

14. Yang H, Liu J, Jing J, Wang Z, Li Y, Gou K, Feng X, Yuan Y, Xing C. Expression of DDB2 Protein in the Initiation, Progression, and Prognosis of Colorectal Cancer. Dig Dis Sci. 2018; 63:2959–68.
https://doi.org/10.1007/s10620-018-5224-z
PMID:30054844

15. Liu J, Li H, Sun L, Feng X, Wang Z, Yuan Y, Xing C. The Differential Expression of Core Genes in Nucleotide Excision Repair Pathway Indicates Colorectal Carcinogenesis and Prognosis. BioMed Res Int. 2018; 2018:9651320.
https://doi.org/10.1155/2018/9651320
PMID:29568775

16. Qiao S, Guo W, Liao L, Wang L, Wang Z, Zhang R, Xu D, Zhang Y, Pan Y, Wang Z, Chen Y. DDB2 is involved in ubiquitination and degradation of PAQR3 and regulates tumorigenesis of gastric cancer cells. Biochem J. 2015; 469:469–80.
https://doi.org/10.1042/BJ20150253
PMID:26205499

17. Feliciano A, Castellvi J, Artero-Castro A, Leal JA, Romagosa C, Hernández-Losa J, Peg V, Fabra A, Vidal F, Kondoh H, Ramón Y Cajal S, Lleonart ME. miR-125b acts as a tumor suppressor in breast tumorigenesis via its novel direct targets ENPEP, CK2-α, CCNJ, and MEGF9. PLoS One. 2013; 8:e76247.
https://doi.org/10.1371/journal.pone.0076247
PMID:24098452

18. Chuang HY, Jiang JK, Yang MH, Wang HW, Li MC, Tsai CY, Jhang YY, Huang JC. Aminopeptidase A initiates tumorigenesis and enhances tumor cell stemness via TWIST1 upregulation in colorectal cancer. Oncotarget. 2017; 8:21266–80.
https://doi.org/10.18632/oncotarget.15072
PMID:28177885

19. Kozasa T, Itoh H, Tsukamoto T, Kaziro Y. Isolation and characterization of the human Gs alpha gene. Proc Natl Acad Sci USA. 1988; 85:2081–5.
https://doi.org/10.1073/pnas.85.7.2081
PMID:3127824

20. Weinstein LS, Liu J, Sakamoto A, Xie T, Chen M. Minireview: GNAS: normal and abnormal functions. Endocrinology. 2004; 145:5459–64.
https://doi.org/10.1210/en.2004-0865
PMID:15331575

21. Peters J, Williamson CM. Control of imprinting at the Gnas cluster. Adv Exp Med Biol. 2008; 626:16–26.
https://doi.org/10.1007/978-0-387-77576-0_2
PMID:18372788

22. Plagge A, Kelsey G, Germain-Lee EL. Physiological functions of the imprinted Gnas locus and its protein variants Galpha(s) and XLalpha(s) in human and mouse. J Endocrinol. 2008; 196:193–214.
https://doi.org/10.1677/JOE-07-0544
PMID:18252944

23. Fox RG, Lytle NK, Jaquish DV, Park FD, Ito T, Bajaj J, Koechlein CS, Zimdahl B, Yano M, Kopp J, Kritzik M, Sicklick J, Sander M, et al. Image-based detection

and targeting of therapy resistance in pancreatic adenocarcinoma. Nature. 2016; 534:407–11. https://doi.org/10.1038/nature17988 PMID:27281208

24. Kudinov AE, Deneka A, Nikonova AS, Beck TN, Ahn YH, Liu X, Martinez CF, Schultz FA, Reynolds S, Yang DH, Cai KQ, Yaghmour KM, Baker KA, et al. Musashi-2 (MSI2) supports TGF-β signaling and inhibits claudins to promote non-small cell lung cancer (NSCLC) metastasis. Proc Natl Acad Sci USA. 2016; 113:6955–60. https://doi.org/10.1073/pnas.1513616113 PMID:27274057

25. Lee J, An S, Choi YM, Lee J, Ahn KJ, Lee JH, Kim TJ, An IS, Bae S. Musashi-2 is a novel regulator of paclitaxel sensitivity in ovarian cancer cells. Int J Oncol. 2016; 49:1945–52. https://doi.org/10.3892/ijo.2016.3683 PMID:27600258

26. Tsujino T, Sugito N, Taniguchi K, Honda R, Komura K, Yoshikawa Y, Takai T, Minami K, Kuranaga Y, Shinohara H, Tokumaru Y, Heishima K, Inamoto T, et al. MicroRNA-143/Musashi-2/KRAS cascade contributes positively to carcinogenesis in human bladder cancer. Cancer Sci. 2019; 110:2189–99. https://doi.org/10.1111/cas.14035 PMID:31066120

27. Zhou L, Sheng W, Jia C, Shi X, Cao R, Wang G, Lin Y, Zhu F, Dong Q, Dong M. Musashi2 promotes the progression of pancreatic cancer through a novel ISYNA1-p21/ZEB-1 pathway. J Cell Mol Med. 2020; 24:10560–72. https://doi.org/10.1111/jcmm.15676 PMID:32779876

28. Wang X, Wang R, Bai S, Xiong S, Li Y, Liu M, Zhao Z, Wang Y, Zhao Y, Chen W, Billiar TR, Cheng B. Musashi2 contributes to the maintenance of CD44v6+ liver cancer stem cells via notch1 signaling pathway. J Exp Clin Cancer Res. 2019; 38:505. https://doi.org/10.1186/s13046-019-1508-1 PMID:31888685

29. Zhao X, Zhang W, Ji W. MYO5A inhibition by miR-145 acts as a predictive marker of occult neck lymph node metastasis in human laryngeal squamous cell carcinoma. OncoTargets Ther. 2018; 11:3619–35. https://doi.org/10.2147/OTT.S164597 PMID:29950866

30. Voz ML, Mathys J, Hensen K, Pendeville H, Van Valckenborgh I, Van Huffel C, Chavez M, Van Damme B, De Moor B, Moreau Y, Van de Ven WJ. Microarray screening for target genes of the proto-oncogene PLAG1. Oncogene. 2004; 23:179–91. https://doi.org/10.1038/sj.onc.1207013 PMID:14712223

31. Liu J, Sun G, Pan S, Qin M, Ouyang R, Li Z, Huang J. The Cancer Genome Atlas (TCGA) based m^{6}A methylation-related genes predict prognosis in hepatocellular carcinoma. Bioengineered. 2020; 11:759–68. https://doi.org/10.1080/21655979.2020.1787764 PMID:32631107

32. Li Z, Feng C, Guo J, Hu X, Xie D. GNAS-AS1/miR-4319/NECAB3 axis promotes migration and invasion of non-small cell lung cancer cells by altering macrophage polarization. Funct Integr Genomics. 2020; 20:17–28. https://doi.org/10.1007/s10142-019-00696-x PMID:31267263

33. Chen X, Zhang D, Jiang F, Shen Y, Li X, Hu X, Wei P, Shen X. Prognostic Prediction Using a Stemness Index-Related Signature in a Cohort of Gastric Cancer. Front Mol Biosci. 2020; 7:570702. https://doi.org/10.3389/fmolb.2020.570702 PMID:33134315

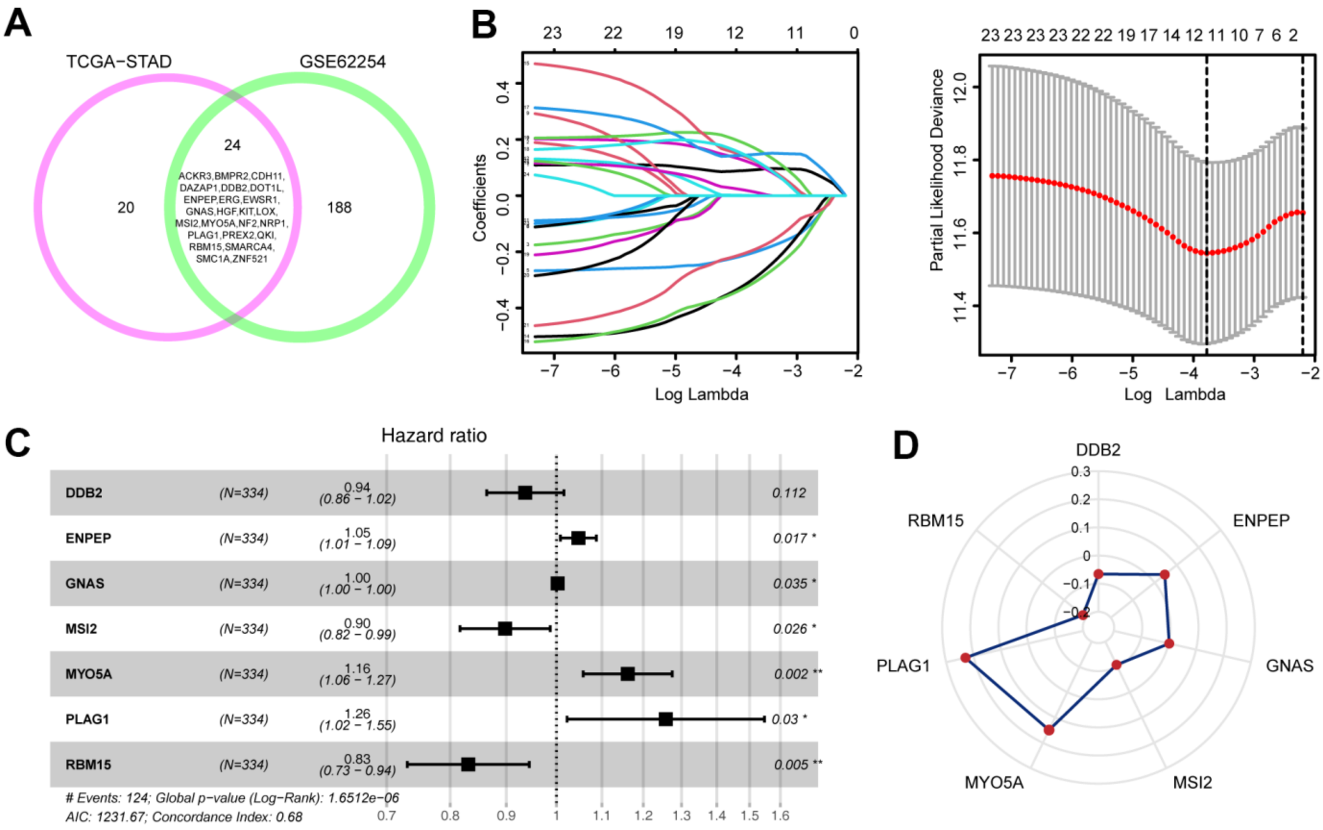
34. Ren H, Zhu J, Yu H, Bazhin AV, Westphalen CB, Renz BW, Jacob SN, Lampert C, Werner J, Angele MK, Bösch F. Angiogenesis-Related Gene Expression Signatures Predicting Prognosis in Gastric Cancer Patients. Cancers (Basel). 2020; 12:12. https://doi.org/10.3390/cancers12123685 PMID:33302481

35. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, Liu J, Yue YG, Wang J, Yu K, Ye XS, Do IG, Liu S, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med. 2015; 21:449–56. https://doi.org/10.1038/nm.3850 PMID:25894828

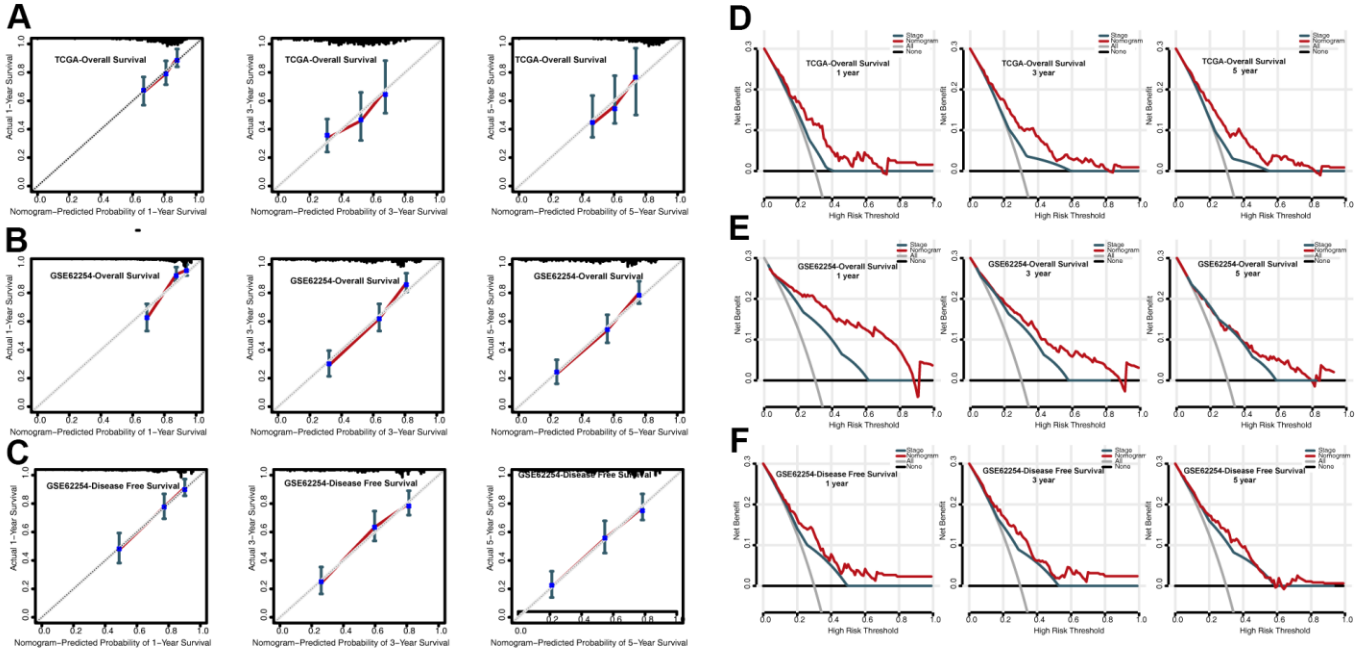
36. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12:453–7. https://doi.org/10.1038/nmeth.3337 PMID:25822800

SUPPLEMENTARY MATERIALS

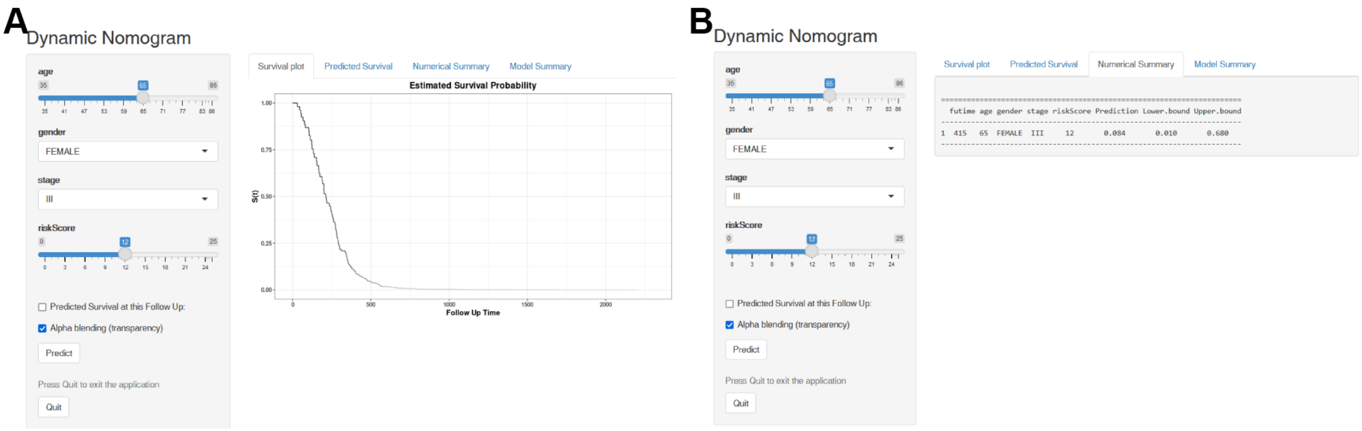
Supplementary Figures



Supplementary Figure 1. Establishment of a prognostic CDG (cancer driver gene) signature. (**A**) Screening of prognosis-related CDGs. (**B**) Least absolute shrinkage and selection operator (LASSO) regression analysis to screen prognosis-related genes from the survival-related CDGs. (**C**) Forest plot of multivariate Cox regression analyses used to construct the prognostic CDG signature. (**D**) Seven-CDG signature for GC based on Cox regression coefficient.



Supplementary Figure 2. Evaluation of the cancer driver gene (CDG)-based nomogram. (**A**) Calibration plot for overall survival (OS) prediction from the training set of the nomogram. (**B**) Calibration plot for OS prediction from the validation set of the nomogram. (**C**) Calibration plot for disease-free survival (DFS) prediction from the validation set of the nomogram. (**D**) Decision curve analysis (DCA) for OS prediction from the training set of the nomogram. (**E**) DCA for OS prediction from the validation set of the nomogram. (**F**) DCA for DFS prediction from the validation set of the nomogram.



age: 65

gender: FEMALE

stage: III

riskScore: 12

□ Predicted Survival at this Follow Up:

☑ Alpha blending (transparency)

Press Quit to exit the application

```
==========================================================================
  futime age gender stage riskScore Prediction Lower.bound Upper.bound
--------------------------------------------------------------------------
1  415   65  FEMALE  III     12       0.084       0.010       0.680
--------------------------------------------------------------------------
```

Supplementary Figure 3. Establishment of an easy-to-operate web-based calculator for predicting gastric cancer (GC) prognoses (https://prognosis.shinyapps.io/STAD/). (**A**) Overall survival rate calculator. (**B**) 95% confidence interval of the overall survival rate determined using the web-based calculator.

Supplementary Table

Please browse Full Text version to see the data of Supplementary Table 1.

Supplementary Table 1. List of cancer driver genes.