A 14 transcription factors-associated nomogram predicts the recurrence-free survival of gastric cancer

Xianxiong Ma
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technolog

Hengyu Chen (chenhy9012@163.com)
Department of Pancreatic Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology https://orcid.org/0000-0001-9637-3992

Ming Yang
Department of Pancreatic Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology

Zunxiang Ke
Department of Vascular Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology

Peng Zhao
Department of Hepatobiliary Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology

Lei Li
Department of Breast and Thyroid Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology

Kaixiong Tao
Department of Gastrointestinal Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology

Research

Keywords: Signature; Gastric cancer; Transcription factor; GEO; TCGA; nomogram.

Posted Date: May 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-27170/v1

Abstract

Background: We aimed to construct and validate a novel transcription factors (TFs) signature for the prediction of gastric cancer (GC) patient's recurrence-free survival (RFS) from TCGA and Gene Expression Omnibus (GEO) database and improve the predictive ability of RFS in GC patients.

Methods: We searched TCGA database and GEO database to obtain gene expression data and related clinical information for GC. In total, 722 TFs and 384 GC patients with intact clinical information were identified to develop a novel TF signature. All TFs were included in a univariate Cox regression model. We then used the least absolute shrinkage and selection operator (LASSO) Cox regression model which included only TFs with $P < 0.05$ in the univariate model to identify candidate TFs related to RFS. After further adjustment, multivariate Cox regression was performed based on the candidate TFs for the identification of TF signatures in the RFS evaluation of GC patients.

Results: We successfully confirmed the high ability of the 14-TF panel for predicting GC patients' RFS by receiver operating characteristic (ROC). AUCs at 1, 3, 5 years in internal validation dataset were 0.827, 0.817, 0.811, respectively. Some similar results were calculated in external validation dataset (0.808, 0.907, 0.813, respectively) and entire dataset (0.815, 0.849, 0.801, respectively). Besides, our model makes a good distinction between the high-risk group and the low-risk group. Furthermore, a nomogram was developed via risk score, sex, cancer status and tumor grade, and C-index, ROC, the calibration plots as well as decision curve analysis (DCA) demonstrated good ability and clinical application of the nomogram.

Conclusions: We successfully established and validated a novel 14-TF-associated nomogram for predicting the RFS of GC.

Background

Gastric cancer (GC) is the most common cause of cancer-related death worldwide [1]. Despite efforts to improve the survival of patients with GC, a favorable prognosis has not been obtained [2]. Presently, the prognostic models for GC are primarily based on the Union for International Cancer Control (UICC) Tumor-Node-Metastasis staging system. However, the results for patients with a similar TNM stage yield great differences due to the inherent heterogeneity [3-6]. Thus, the determination of effective signatures for predicting the prognosis of GC could improve individualized clinical management.

Numerous researches reported that TFs played a significant role in the progression and prognosis of cancer. For instance, Edwards et al. revealed that ZEB1 served as a TF which was prognostic and predictive in diffuse gliomas [7]. Oktay et al. indicated that thyroid TF-1 played a key role in the prognosis of lung adenocarcinoma (LUAD) [8]. Su et al. reported that TF 7 served as a poor prognostic marker of glioblastoma multiforme by enhancing proliferation by upregulating c-myc [9]. Lee et al. suggested that combination immunohistochemistry for SMAD4 and runt-related TF 3 may determine a favorable prognostic subgroup of pancreatic ductal adenocarcinomas [10]. Fan et al. reported that

microRNA-301a-3p overexpression may contribute to cell invasion and proliferation by targeting runt-related TF 3 in prostate cancer [11]. Therefore, the studies on TF are promising in identifying predictive biomarkers to help doctors offer personalized treatments for cancer and may improve patients' survival time. However, few researches have revealed the key role of TFs as independent biomarkers for GC prognosis. The identification of TFs as independent and valuable predictors for GC prognosis by a comprehensive and systematic method is essential.

In our study, we obtained genes expression data and clinical information for GC from TCGA and GEO databases and corresponding TFs and eligible patients were determined to investigate transcription factor marker for GC prognosis. We identified a 14-TF signature for predicting RFS of GC patients by bioinformatic integrated analysis. According to Kaplan–Meier method□ROC analysis, we confirmed the high ability of the 14-TF signature in prognostic assessment for GC. Besides, we developed a predictive nomogram that integrated the 14-TF signature with the conventional clinicopathological factors and the result suggested a good predictive value of our nomogram.

Materials And Methods

Data source and processing

We searched TCGA database and GEO database using TCGAbiolinks package [12] and GEOquery package, respectively [13] to obtain genes expression data and related clinical information for GC. In total, 24991 genes and 407 GC patients with intact clinical information in TCGA database were included. Samples without prognostic data or non-TF genes were excluded from subsequent analysis. TFs were determined based on TRRUST database (https://www.grnpedia.org/trrust/downloadnetwork.php) [14]. Raw counts of expression matrix were converted to transcripts per million (TPM). Genes with 0 expression more than 20% of the samples were excluded. Finally, 722 TFs and 384 patients with GC were identified for Univariate Cox regression analyses. Similarly, raw data of GSE26253 was preprocessed and normalized by the robust multichip averaging (RMA) [15] method using affy packages [16] of R (v3.6.1). In the end, 432 patients in GSE26253 were included as an external validation set. LASSO method was used for identifying the candidate TFs to predict RFS of GC patients. LASSO COX regression model was conducted via a publicly available R package for 1000 iterations.

Gene sets enrichment analysis and protein-protein interaction (PPI) analyses

The TFs screened by the univariate Cox regression analyses ($p<0.05$) were used to perform Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using R clusterProfiler package [17]. GO analysis was significantly involved in cellular components, biologic processes and molecular functions of genes [18]. KEGG pathway analysis was mainly associated with the molecular interaction, reaction, and relation networks [19].

A PPI network was built on the basis of the Search Tool for the Retrieval of Interacting Genes (STRING, http://string-db.org) and visualized by Cytoscape (ver. 3.5.1) [20]. Only experimentally confirmed

interplays with a combined score >0.4 were selected as significant. The plug-in Molecular Complex Detection (MCODE) was employed to screen the prime module from the PPI network. The criteria were defined as follows: MCODE scores >3 and the number of nodes ≥4.

Generation of TF-associated signature

The association between the TFs expression and patient's RFS was evaluated by univariable Cox regression analysis to select TFs relevant to patients' RFS. Then, the determined TFs were used to perform LASSO analysis for selecting the candidate TFs reliably associated with RFS. After that, multivariate Cox regression was executed based on the candidate TFs for the predictive TF signatures in the survival RFS evaluation of GC patients.

384 patients were randomly assigned to training set (n = 269) and internal validation set (n = 115). The training cohort was employed for the identification of prognostic TF signature. Internal validation set and external validation set as well as entire TCGA dataset were applied to validate our results. The TF risk score formula was then established to determine RFS risk for every patient with the coefficients from the multivariate Cox regression analysis. Patients with GC in each set were stratified into high-risk or low-risk group with the corresponding median risk score as the cutoff point. Survival differences between the high-risk and low-risk groups in each set were weighed using the Kaplan-Meier method, and compared by the log-rank test. We conducted ROC analysis to assess the sensitivity and specificity of the survival prediction based on the TF risk score. The greater the AUC value was, the more superior the model was for the hazard prediction. Then stratified analysis was acted based on clinical parameters in the whole TCGA set. All ROC and Kaplan-Meier curves were drew with R (version 3.6.1).

Gene set variation analysis (GSVA)

To unearth the 14-TF signature-related signaling pathways, single sample gene sets enrichment analysis (ssGSEA) was carried out through the TCGA GC mRNA dataset with GSVA package [21]. The top 20 important pathways positively related to risk score were investigated. Patients were assigned into high- or low-risk cohorts with the cutoff of the median risk score. Adjusted $P < 0.05$ was considered to be of significance.

Construction of the nomogram

The univariate Cox proportional hazard analysis and multivariate Cox proportional hazard analysis were acted on the basis of risk score and other clinicopathological factors. The factors with $P<0.05$ from multivariate Cox proportional hazard analysis were employed to establish a nomogram via the 'rms' R package. Hazard ratios (HR) and corresponding 95% confidence interval (CI) were measured by Cox proportional hazard models. The prognostic value of the nomogram was assessed by C-index, ROC, calibration plots and DCA. The outcome of the nomogram was listed in the calibrate curve and the 45-degree line implied an ideal performance.

Results

Clinical characteristics of the study populations

The study was performed on 384 patients who were clinically and pathologically diagnosed with GC. Of these patients, 247(64.32%) were male and 137(35.68%) were female. The median age at diagnosis was 68 years (range, 35–90) and the median RFS was 383 days respectively. The 3-year RFS rate of all patients was 10.4%. The pathologic stage was defined according to the American Joint Committee on Cancer (AJCC) Cancer staging manual. The stage of GC patients ranged from I to IV, and 56 (14.58%) patients in state I, 119 (30.99%) patients in stage Ⅱ, 144 (37.5%) patients in stage Ⅲ, and 42 (10.94%) in stage IV, 23 (5.99%) patients in stage X (X: the stage can not be identified). The GC dataset from TCGA consists of three histological types: stomach adenocarcinoma (213, 55.49%), stomach-intestinal adenocarcinoma (170, 44.27%), not available (1, 0.26%). Patients were separated into three groups according to the cancer status of samples, which contains tumor free (262, 68.23%), with tumor (69, 17.97%), indeterminate (53, 13.80%). In addition, race list included Asian, Black or African American, native Hawaiian or other Pacific Islander, White, indeterminate, and the White group was the most common (239, 62.24%). The complete list of clinicopathological characteristics of all the included patients in TCGA and GEO databases was detailed in **Table 1**. The study flowchart was presented in **Figure 1**.

Gene sets enrichment analysis and PPI analysis

Figure 2A and Figure 2B showed the top 10 enriched GO terms and genes correlated with the top 5 GO terms respectively. **Figure 2C and Figure 2D** listed the top 10 KEGG pathways and genes correlated with the top 3 enriched KEGG pathways respectively. Genes with interaction greater than 8 were set as hub genes. Finally, 6 hub genes were selected based on PPI analysis: (WDR5, TBP, PAX5, MYB, POU5F1, SMAD3) (**Figure 2E**). The top 2 core sub-modules from the PPI network were employed for annotating gene function. Enrichment analysis indicated that the genes in these 2 sub-modules were primarily associated with DNA replication and wnt signaling (**Figure 2F**).

Identification of TFs significantly associated with RFS and establishment of prognostic signatures

Univariate Cox regression analysis and LASSO Cox regression analysis were conducted to identify the relationship between the 722 TFs and RFS in patients with GC. As a result, 28 TFs (**Figure 3A & 3B**) were revealed to be significantly correlated with GC patients' RFS after LASSO Cox regression analysis (**Figure 3A & 3B**). Finally, 14 TFs (NOTCH3, NR5A1, WDR5, RARB, SRCAP, SMAD3, ONECUT1, PITX3, TRAF6, MTA2, JDP2, FOSL1, GLI1, MTFl) were revealed to be significantly related to GC patients' RFS by multivariate Cox analysis. Risk score = 6e-05*NOTCH3 + 0.00878*NR5A1 - 0.00124*WDR5 + 0.00233*RARB + 1e-04*SRCAP + 0.00025*SMAD3 + 0.00217*ONECUT1 + 0.08996*PITX3 - 0.00186*TRAF6 - 0.00039*MTA2 - 0.00233*JDP2 + 0.00012*FOSL1 + 0.00196*GLI1 - 0.00257*MTF1. The 14-TF signature was employed for predicting RFS of GC patients. Obviously, the high TF expression of NOTCH3, NR5A1, RARB, SRCAP, SMAD3, ONECUT1, PITX3, FOSL1 and GLI1 was corresponding to a

higher risk. Nevertheless, the low TF expression of WDR5, TRAF6, MTA2, MTF1 and JDP2 was corresponding to higher risk (**Figure 4**) (**Figure S1**).

Relationship between the 14-TF signature and GC patients' RFS in internal validation dataset and external validation dataset as well as the whole dataset

Kaplan–Meier analysis was applied to measure the difference in RFS between the two groups. RFS for the high-score GC patients was shorter than that for the low-score GC patients in internal validation set (P= 3e-10) (**Figure 5A**). A similar outcome was observed in the external validation dataset (p =6e-05) (**Figure 5C**) and entire dataset (p = 1e-13) (**Figure 5E**).

Evaluation of the predictive performance of the 14-TF signature by using ROC analysis

Time-dependent ROC curves were drew to assess the predictive power of the 14-TF signature. The AUC of the 14-TF signature at 1, 3, 5 years in internal validation dataset were 0.827, 0.817, 0.811, respectively (**Figure 5B**). A high predictive power was also presented in external validation dataset (0.808, 0.907, 0.813) (**Figure 5D**) and entire dataset (0.815, 0.849, 0.801) (**Figure 5F**). The result suggested that the 14-TF signature was a stable predictor for RFS of GC patients.

Furthermore, patients were ranked with the risk scores (**Figure 6A**), and the dot plot was drew via their survival status (**Figure 6B**). The outcome implied that the high-risk cohort generated a greater mortality rate than that in the low-risk cohort. Heatmap of 14 TFs grouped according to risk score was presented in **Figure 6C**, which confirmed our previous boxplot. A similar result was obtained in GSE26253 (**Figure S2**). Besides, subgroup analysis was acted by a few clinicopathological factors consisting of age, gender, stage, histologic type, anatomic site and metastasis status. The result demonstrated a good predictive power of the 14-TF in the majority of sub-groups (**Figure S3-S7**).

Determination of the 14-TF signature-related biological pathways

Patients were assigned into high- or low-risk cohort in accordance to the cutoff of the median risk score. Top 20 pathways that were more activated in the high-risk patients than that in low-risk patients were exhibited in (**Figure 7A**) (Table S4). The same trend was evident in the enriched pathways and risk score (**Figure 7B**), suggesting a good correlation between the pathways and the risk score.

Nomogram development

We performed univariate and multivariate Cox model via TF related risk score and a few other clinicopathological factors to weigh independence of the 14-TF signature as a prognostic predictor of GC patients. Hazard ratios (HRs) demonstrated that the 14-TF signature was importantly correlated with the RFS of GC patients (P= 2.60E-13, HR 2.11, 95% CI 1.72-2.57) by the outcome of Cox regression analysis (**Table 2**) (**Figure 8**), implying that the 14-TF signature functioned as an independent prognostic predictor. A nomogram (**Figure 9**) combining TFs risk score with other clinical factors (p<0.2 in multivariate Cox analysis) was developed. The importance between the 14-TF signature and the clinicopathological

factors was observed in (**Figure 10A**). The result showed that C-index (0.788, 95%CI: 0.741-0.835), AUC (0.865, 0.921, 0.907) (**Figure 10B**) and calibration plot presented a good performance (**Figure 10C & 10D & 10E**). In addition, the DCA implied that the established nomogram had more crucial clinical value for the prediction of RFS in GC patients than that in treat all or treat none cohort. The particular benefit was obtained for GC patients' 3-year recurrent risks (**Figure 10F**), suggesting strong robustness of our model.

Discussion

GC remains a severe challenge for public health worldwide. Currently, the prognostic models for GC are mainly based on the UICC Tumor-Node-Metastasis staging system. Whereas, the results for patients with a similar TNM stage yield great difference due to the inherent heterogeneity [3-6]. The identification of novel prognostic predictors and the establishment of more valuable prognostic models are urgently required.

In this study, we identified a combination of 14 TFs (NOTCH3, NR5A1, WDR5, RARB, SRCAP, SMAD3, ONECUT1, PITX3, TRAF6, MTA2, JDP2, FOSL1, GLI1, MTFl) and effectively predicted RFS in GC patients using the univariate Cox proportional hazard analysis, the LASSO Cox regression analysis and multivariate Cox proportional hazard analysis. Various experiments have indicated that the above 14 TFs were significant in cancer development. For instance, Ganguly et al. reported that Notch3 promoted prostate cancer-induced bone lesion development by MMP-3 [22]. Impaired steroidogenic factor 1 (NR5A1) activity played a significant role in mutant Y1 mouse adrenocortical tumor cells [23]. Expression of WD repeat domain 5 (WDR5) was relevant to progression and reduced prognosis in papillary thyroid carcinoma [24]. Methylation of L1RE1, RARB, and RASSF1 served as potential biomarkers for the differential diagnosis of lung cancer [25]. The chromatin remodeling factor SRCAP regulated expression of prostate-specific antigen as well as cellular proliferation in prostate cancer cells [26]. MicroRNA-17 served as an oncogene via downregulating smad3 expression in hepatocellular carcinoma [27]. Loss of ONECUT1 expression revealed a key role in human pancreatic cancer cells [28]. PITX3 DNA methylation functioned as an independent biomarker for predicting the overall survival of patients with head and neck squamous cell carcinoma [29]. TRAF6 enhanced the progression and growth of colorectal cancer by nuclear shuttle regulation NF-kB/c-jun signaling pathway [30]. miR-1236-3p depressed invasion and metastasis in gastric cancer through targeting MTA2 [31]. ATF3 and JDP2 deficiency in cancer-related fibroblasts enhanced tumor growth by SDF-1 transcription [32]. FOSL1 promoted growth and metastasis of human prostate cancer cells via epithelial mesenchymal transition pathway [33]. DZIP1 enhanced proliferation, migration, and invasion of oral squamous carcinoma by the GLI1/3 pathway [34]. Tumors expressing the mtfl-mrfp1-wttk fusion reporter showed a key role in the study of construction and validation of improved triple fusion reporter gene vectors for molecular imaging of living subjects [35].

The result of GO terms and KEGG pathways on the basis of the TFs screened by the Univariate Cox regression analyses suggested that cell fate commitment term, human T−cell leukemia virus 1 infection and transcriptional misregulation in cancer pathways were associated with cancer development. For example, Gastaldi et al. reported that met signaling modulated growth, repopulating potential and basal

cell-fate commitment of mammary luminal progenitors: indication for basal-like breast cancer [36], which revealed a key role of cell fate commitment term in cancer development. A previous research revealed that the human T cell leukemia/lymphotropic virus-1 (HTLV-I) served as the etiologic agent of adult T cell leukemia (ATL), an aggressive and deadly leukemia of CD4+ T lymphocytes [37]. Zhao et al. demonstrated that transcriptional misregulation in cancer pathway was involved in colorectal cancer [38]. The result suggested that cell fate commitment term, Human T−cell leukemia virus 1 infection and transcriptional misregulation in cancer pathways may be useful therapeutic strategies for GC. Whereas, further exploration was needed to confirm the hypothesis.

Top 6 hub genes (SMAD3, POU5F1, TBP, MYB, WDR5, PAX5) were screened according to PPI analysis. Researchers have revealed that the aforementioned 6 hub genes may be important in cancer development. For instance, long noncoding RNA OPA-interacting protein 5 antisense transcript 1 upregulated SMAD3 expression to promote metastasis of cervical cancer via sponging miR-143-3p [39]. The POU5F1 gene expression served as a prognostic marker in colorectal cancer [40]. TXNIP (VDUP-1, TBP-2) functioned as a primary redox regulator commonly suppressed in cancer through epigenetic mechanisms [41]. MYB modulated the DNA damage response and components of the homology-directed repair pathway in human estrogen receptor-positive breast cancer cells [42]. LncRNA GCAWKR contributed to gastric cancer development through scaffolding the chromatin modification factors WDR5 and KAT2A [43]. PAX5 was expressed in small-cell lung cancer and positively modulated c-Met transcription [44].

Limitations in this study were the following: First, more independent external validation sets were needed to evaluate the power of the 14-TF signature. Second, we constructed the nomogram based on the TCGA dataset solely due to the lack of complete clinical information of GSE262563 dataset. Whereas, there were still a few valuable virtues in our study. LASSO method was used to filter variables between univariate and multivariate Cox analysis, eliminating the interference of multicollinearity. In addition, no studies have combined TF signature with clinical factors to predict RFS for GC yet. We combined TF bioinformatics analysis with clinical factors, which may help direct translational study and the application of molecularly targeted therapy. We built a nomogram that combined both the 14-TF signature and the conventional clinicopathological factors to predict 1, 3-and 5-year RFS. The outcome implied good ability of 14-TF signature for predicting RFS of GC patients in the clinical routine, which made our results more significant. Furthermore, the DCA was also employed to measure the value of our nomogram. Various studies suggested that DCA was implemented for the assessment of predictive models in clinical studies. For instance, Ishioka et al. developed a nomogram incorporating serum c-reactive protein level to predict overall survival of patients with advanced urothelial carcinoma and its evaluation by Decision Curve Analysis [45]. Mo et al. revealed predictive factors of synchronous colorectal peritoneal metastases: development of a nomogram and study of its utilities using decision curve analysis [46].

Traditional indicators for diagnosis such as sensitivity, specificity only assess whether the diagnostic accuracy of a predictive model precedes another, whereas the clinical utility of a specific strategy is not

included [47]. DCA which functioned as an indicator for clinical utility may address this issue [48]. The result revealed an important clinical utility of DCA in present study.

Conclusion

We identified a 14-TF signature for predicting the RFS of GC patients by bioinformatic integrated analysis. According to Kaplan–Meier method□ROC analysis, we confirmed that the 14-TF signature was an effective prognostic predictor for GC patients' RFS. In addition, we built a predictive nomogram that integrated the 14-TF signature and the conventional clinicopathological factors and the result proved a good predictive capacity of our nomogram.

Abbreviations

TF: transcription factor; GC: gastric cancer; STAD: stomach adenocarcinoma; TGGA: The Cancer Genome Atlas; UICC: Union for International Cancer Control; PPI: protein-protein interaction; DEGs: differentially expressed genes; GO: gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; STRING: search Tool for the Retrieval of Interacting Genes; LASSO: the least absolute shrinkage and selection operator; RFS: recurrence free survival. GEO: Gene Expression Omnibus; ROC: receiver operating characteristic; AUC: area under curve; DCA: decision curve analysis;

Declarations

Authors' Contributions

HC, XM and MY designed, extracted, analyzed, and interpreted the data from TCGA and GEO databases. ZK and PZ wrote the manuscript. LL made substantial contributions to the conception of the work and substantively revised it. All authors have read and approved the final manuscript.

Acknowledgments

Not available.

Funding

This study was supported by the National Natural Science Foundation of China (Grant number: No. 81702397 to Lei Li, NO. 81874184 to Kaixiong Tao)

Availability of data and materials

All data generated or analyzed during the present study are included in this published article or are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Torre, L.A., et al., *Global cancer statistics, 2012.* CA Cancer J Clin, 2015. **65**(2): p. 87-108.
2. Ajani, J.A., et al., *Gastric Cancer, Version 3.2016, NCCN Clinical Practice Guidelines in Oncology.* J Natl Compr Canc Netw, 2016. **14**(10): p. 1286-1312.
3. Razzak, M., *Genetics: new molecular classification of gastric adenocarcinoma proposed by The Cancer Genome Atlas.* Nat Rev Clin Oncol, 2014. **11**(9): p. 499.
4. McLean, M.H. and E.M. El-Omar, *Genetics of gastric cancer.* Nat Rev Gastroenterol Hepatol, 2014. **11**(11): p. 664-74.
5. Jiang, Y., et al., *Immunomarker Support Vector Machine Classifier for Prediction of Gastric Cancer Survival and Adjuvant Chemotherapeutic Benefit.* Clin Cancer Res, 2018. **24**(22): p. 5574-5584.
6. Jiang, Y., et al., *Association of Adjuvant Chemotherapy With Survival in Patients With Stage II or III Gastric Cancer.* JAMA Surg, 2017. **152**(7): p. e171087.
7. Edwards, L.A., et al., *ZEB1 Is a Transcription Factor That Is Prognostic and Predictive in Diffuse Gliomas.* Front Neurol, 2018. **9**: p. 1199.
8. Oktay, E., et al., *The prognostic role of thyroid transcription factor-1 in lung adenocarcinoma.* J Cancer Res Ther, 2018. **14**(Supplement): p. S1201-s1208.
9. Su, Y., et al., *Transcription factor 7 functions as an unfavorable prognostic marker of glioblastoma multiforme by promoting proliferation by upregulating c-Myc.* Neuroreport, 2018. **29**(9): p. 745-752.
10. Lee, Y., et al., *Combination immunohistochemistry for SMAD4 and Runt-related transcription factor 3 may identify a favorable prognostic subgroup of pancreatic ductal adenocarcinomas.* Oncotarget, 2017. **8**(44): p. 76699-76711.
11. Fan, L., et al., *MicroRNA301a3p overexpression promotes cell invasion and proliferation by targeting runtrelated transcription factor 3 in prostate cancer.* Mol Med Rep, 2019. **20**(4): p. 3755-3763.
12. Colaprico, A., et al., *TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.* Nucleic Acids Res, 2016. **44**(8): p. e71.
13. Davis, S. and P.S. Meltzer, *GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.* Bioinformatics, 2007. **23**(14): p. 1846-7.

14. Han, H., et al., *TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions.* Nucleic Acids Res, 2018. **46**(D1): p. D380-d386.
15. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
16. Gautier, L., et al., *affy--analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-15.
17. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* Omics, 2012. **16**(5): p. 284-7.
18. Qian, Z., Y.D. Cai, and Y. Li, *A novel computational method to predict transcription factor DNA binding preference.* Biochem Biophys Res Commun, 2006. **348**(3): p. 1034-7.
19. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
20. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
21. Hänzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-seq data.* BMC Bioinformatics, 2013. **14**: p. 7.
22. Ganguly, S.S., et al., *Notch3 promotes prostate cancer-induced bone lesion development via MMP-3.* Oncogene, 2019.
23. Frigeri, C., et al., *Impaired steroidogenic factor 1 (NR5A1) activity in mutant Y1 mouse adrenocortical tumor cells.* Mol Endocrinol, 2000. **14**(4): p. 535-44.
24. Xu, W., et al., *Expression of WD Repeat Domain 5 (WDR5) is Associated with Progression and Reduced Prognosis in Papillary Thyroid Carcinoma.* Med Sci Monit, 2019. **25**: p. 3762-3770.
25. Walter, R.F.H., et al., *Methylation of L1RE1, RARB, and RASSF1 function as possible biomarkers for the differential diagnosis of lung cancer.* PLoS One, 2018. **13**(5): p. e0195716.
26. Slupianek, A., et al., *The chromatin remodeling factor SRCAP modulates expression of prostate specific antigen and cellular proliferation in prostate cancer cells.* J Cell Physiol, 2010. **224**(2): p. 369-75.
27. Lu, Z., et al., *microRNA-17 functions as an oncogene by downregulating Smad3 expression in hepatocellular carcinoma.* Cell Death Dis, 2019. **10**(10): p. 723.
28. Jiang, X., et al., *Loss of ONECUT1 expression in human pancreatic cancer cells.* Oncol Rep, 2008. **19**(1): p. 157-63.
29. Sailer, V., et al., *PITX3 DNA methylation is an independent predictor of overall survival in patients with head and neck squamous cell carcinoma.* Clin Epigenetics, 2017. **9**: p. 12.
30. Zhu, G., et al., *TRAF6 promotes the progression and growth of colorectal cancer through nuclear shuttle regulation NF-kB/c-jun signaling pathway.* Life Sci, 2019. **235**: p. 116831.
31. An, J.X., et al., *miR-1236-3p inhibits invasion and metastasis in gastric cancer by targeting MTA2.* Cancer Cell Int, 2018. **18**: p. 66.

32. Avraham, S., et al., *ATF3 and JDP2 deficiency in cancer associated fibroblasts promotes tumor growth via SDF-1 transcription.* Oncogene, 2019. **38**(20): p. 3812-3823.
33. Luo, Y.Z., P. He, and M.X. Qiu, *FOSL1 enhances growth and metastasis of human prostate cancer cells through epithelial mesenchymal transition pathway.* Eur Rev Med Pharmacol Sci, 2018. **22**(24): p. 8609-8615.
34. Yan, W., et al., *DZIP1 Promotes Proliferation, Migration, and Invasion of Oral Squamous Carcinoma Through the GLI1/3 Pathway.* Transl Oncol, 2019. **12**(11): p. 1504-1515.
35. Ray, P., R. Tsien, and S.S. Gambhir, *Construction and validation of improved triple fusion reporter gene vectors for molecular imaging of living subjects.* Cancer Res, 2007. **67**(7): p. 3085-93.
36. Gastaldi, S., et al., *Met signaling regulates growth, repopulating potential and basal cell-fate commitment of mammary luminal progenitors: implications for basal-like breast cancer.* Oncogene, 2013. **32**(11): p. 1428-40.
37. Mamane, Y., et al., *Repression of IRF-4 target genes in human T cell leukemia virus-1 infection.* Oncogene, 2002. **21**(44): p. 6751-65.
38. Zhao, Z.W., et al., *The identification of a common different gene expression signature in patients with colorectal cancer.* Math Biosci Eng, 2019. **16**(4): p. 2942-2958.
39. Chen, X., et al., *Long noncoding RNA OPA-interacting protein 5 antisense transcript 1 upregulated SMAD3 expression to contribute to metastasis of cervical cancer by sponging miR-143-3p.* J Cell Physiol, 2019. **234**(4): p. 5264-5275.
40. Miyoshi, N., et al., *The POU5F1 gene expression in colorectal cancer: a novel prognostic marker.* Surg Today, 2018. **48**(7): p. 709-715.
41. Zhou, J., Q. Yu, and W.J. Chng, *TXNIP (VDUP-1, TBP-2): a major redox regulator commonly suppressed in cancer by epigenetic mechanisms.* Int J Biochem Cell Biol, 2011. **43**(12): p. 1668-73.
42. Yang, R.M., et al., *MYB regulates the DNA damage response and components of the homology-directed repair pathway in human estrogen receptor-positive breast cancer cells.* Oncogene, 2019. **38**(26): p. 5239-5249.
43. Ma, M., et al., *lncRNA GCAWKR Promotes Gastric Cancer Development by Scaffolding the Chromatin Modification Factors WDR5 and KAT2A.* Mol Ther, 2018. **26**(11): p. 2658-2668.
44. Kanteti, R., et al., *PAX5 is expressed in small-cell lung cancer and positively regulates c-Met transcription.* Lab Invest, 2009. **89**(3): p. 301-14.
45. Ishioka, J., et al., *Development of a nomogram incorporating serum C-reactive protein level to predict overall survival of patients with advanced urothelial carcinoma and its evaluation by decision curve analysis.* Br J Cancer, 2012. **107**(7): p. 1031-6.
46. Mo, S., et al., *Predictive factors of synchronous colorectal peritoneal metastases: Development of a nomogram and study of its utilities using decision curve analysis.* Int J Surg, 2018. **54**(Pt A): p. 149-155.
47. Fitzgerald, M., B.R. Saville, and R.J. Lewis, *Decision curve analysis.* Jama, 2015. **313**(4): p. 409-10.

48. Zhang, Z., et al., *Decision curve analysis: a technical note.* Ann Transl Med, 2018. **6**(15): p. 308.

Tables

Table 1. Clinical characteristics of included patients.

Characteristics	Total	Training dataset (n=269)	Testing dataset (n=115)	GSE26253 (n=432)
Age				
<65	155(40.36)	106(39.41)	49(42.61)	
>=65	229(59.64)	163(60.59)	66(57.39)	
Sex				
Female	137(35.68)	99(36.8)	38(33.04)	
Male	247(64.32)	170(63.2)	77(66.96)	
Histological type				
Stomach Adenocarcinoma	213(55.49)	149(55.39)	64(55.65)	
Stomach- Intestinal Adenocarcinoma	170(44.27)	120(44.61)	50(43.48)	
Not available	1(0.26)		1(0.87)	
Stage				
Stage I	56(14.58)	43(15.99)	13(11.3)	68(15.7)
Stage II	119(30.99)	80(29.74)	39(33.91)	167(38.7)
Stage III	144(37.5)	105(39.03)	39(33.91)	130(30.1)
Stage IV	42(10.94)	28(10.41)	14(12.17)	67(15.5)
Stage X	23(5.99)	13(4.83)	10(8.7)	
Tumor				
T1	21(5.47)	14(5.2)	7(6.09)	
T2	88(22.92)	65(24.16)	23(20)	
T3	167(43.49)	118(43.87)	49(42.61)	
T4	100(26.04)	68(25.28)	32(27.83)	
TX	8(2.08)	4(1.49)	4(3.48)	
Node				
N1	105(39.03)	69(37.1)	36(43.37)	
N2	79(29.37)	55(29.57)	24(28.92)	
N3	66(24.54)	49(26.34)	17(20.48)	
NX	19(7.06)	13(6.99)	6(7.23)	
Metastasis status				
M0	339(88.28)	239(88.85)	100(86.96)	
M1	26(6.77)	18(6.69)	8(6.96)	
MX	19(4.95)	12(4.46)	7(6.09)	
Cancer status				
Tumor free	262(68.23)	184(68.4)	78(67.83)	
With tumor	69(17.97)	49(18.22)	20(17.39)	
Not available	53(13.80)	36(13.38)	17(14.78)	
Ethnicity				
Hispanic or Latino	5(1.3)	2(0.74)	3(2.61)	
Not Hispanic or Latino	274(71.35)	198(73.61)	76(66.09)	
Not available	105(27.35)	69(25.65)	36(31.3)	
Residual tumor				

R0	311(80.99)	218(81.04)	93(80.87)
R1	13(3.39)	10(3.72)	3(2.61)
R2	18(4.69)	13(4.83)	5(4.35)
RX	19(4.95)	12(4.46)	7(6.09)
Not available	23(5.99)	16(5.95)	7(6.09)
Race			
Asian	77(20.05)	52(19.33)	25(21.74)
Black or African American	12(3.12)	9(3.35)	3(2.61)
Native Hawaiian or other Pacific Islander	1(0.26)	1(0.37)	
White	239(62.24)	171(63.57)	68(59.13)
Not available	55(14.33)	36(13.38)	19(16.52)
Grade			
G1	10(2.6)	7(2.6)	3(2.61)
G2	144(37.5)	102(37.92)	42(36.52)
G3	222(57.81)	153(56.88)	69(60)
GX	8(2.08)	7(2.6)	1(0.87)
Pylori infection			
No	141(36.72)	96(35.69)	45(39.13)
Yes	19(4.95)	12(4.46)	7(6.09)
Not available	224(58.33)	161(59.85)	63(54.78)
Number of positive lymph nodes			
0	99(25.78)	72(26.77)	27(23.48)
1-2	77(20.05)	53(19.7)	24(20.87)
>2	165(42.97)	115(42.75)	50(43.48)
Not available	43(11.2)	29(10.78)	14(12.17)
Barrett's esophagus			
No	198(51.56)	129(47.96)	69(60)
Yes	20(5.21)	18(6.69)	2(1.74)
Not available	166(43.23)	122(45.35)	44(38.26)
Antireflux treatment			
No	134(34.9)	93(34.57)	41(35.65)
Yes	42(10.94)	25(9.29)	17(14.78)
Not available	208(54.17)	151(56.14)	57(49.56)

Table 2. Univariate Cox regression analysis and multivariate Cox regression analysis outcome based on TFs risk score and other clinical factors.

	Univariate Cox analysis				Multivariate Cox analysis			
Characteristics	HR	HR.95L	HR.95H	P value	HR	HR.95L	HR.95H	P value
Score	2.718282	2.25954	3.270159	2.87E-26	2.105799	1.724836	2.570906	2.60E-13
Sex	1.874275	1.183928	2.967164	0.007355	1.416686	0.883074	2.272744	0.148641
Histological type	0.898199	0.736756	1.095018	0.288211				
T	1.075018	0.917442	1.259658	0.371066				
N	1.296917	1.125086	1.49499	0.000337	1.067364	0.815654	1.396751	0.634735
M	1.112185	0.727047	1.701341	0.623965				
Stage	1.229799	1.055993	1.432213	0.007796	0.919478	0.704143	1.200665	0.537464
Grade	1.280151	1.042481	1.572005	0.018423	1.185943	0.948725	1.482473	0.13421
HP infection	0.670859	0.475503	0.946476	0.023013	1.240031	0.818149	1.879458	0.310594
Reflux history	0.609943	0.439642	0.846213	0.00308	0.810763	0.567863	1.157561	0.248239
Residual tumor	1.482089	1.143333	1.921214	0.002962	0.992409	0.732518	1.344507	0.960774
Ethnicity	0.973056	0.672818	1.407272	0.884638				
Race	1.094327	0.885531	1.352353	0.403998				
Number of positive lymph nodes	1.058645	1.03801	1.07969	1.39E-08	1.019857	0.985855	1.055031	0.25574
Cancer status	2.480676	2.008099	3.064469	3.59E-17	1.796603	1.384077	2.332083	1.07E-05
Age	0.999936	0.985464	1.01462	0.993097				
Anatomic site	0.956627	0.828779	1.104196	0.544645				
Barrett’s esophagus	0.960004	0.676377	1.362567	0.819296				

Figures

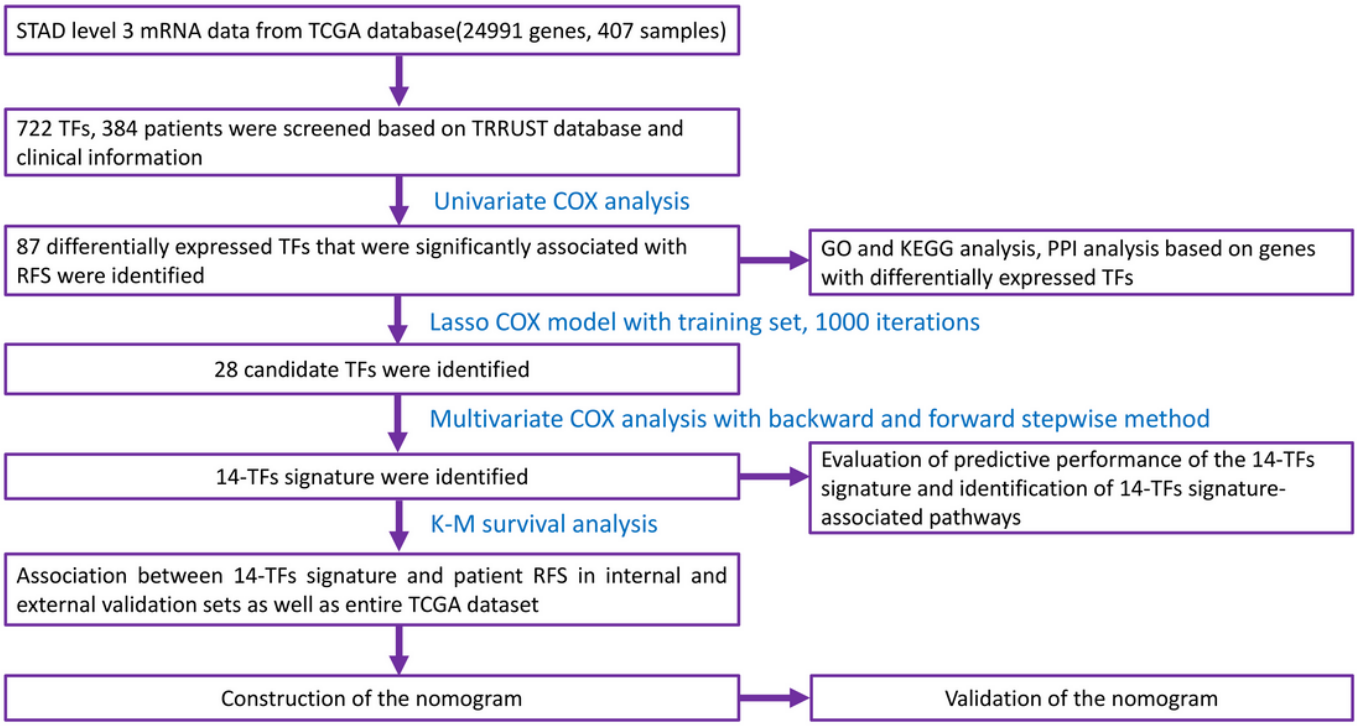

Figure 1

Flow chart of the present study.

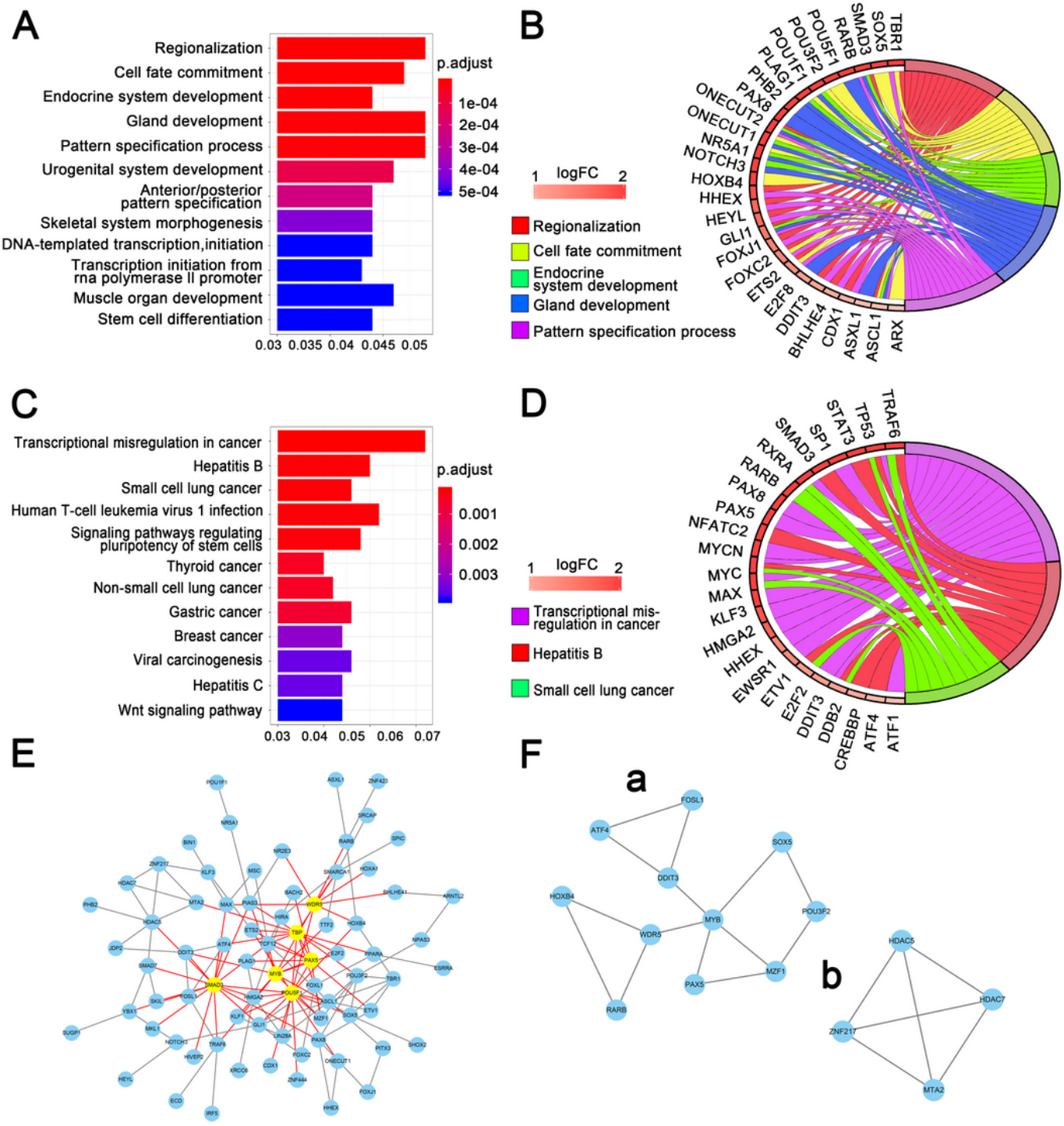

Figure 2

Gene sets enrichment analysis and protein-protein interaction analysis. (A) The top 10 enriched GO terms. The original P values were transformed by '-log (P.adj value)' to plot the bar chart. (B) Genes associated with the top 5 GO terms. (C) The top 10 KEGG pathways (D) Genes associated with the top 3 enriched KEGG pathways (E) The protein-protein interaction network was built based on the 87 transcription factors. Yellow nodes stood for hub genes. (F) The top 2 sub-module from the PPI network.

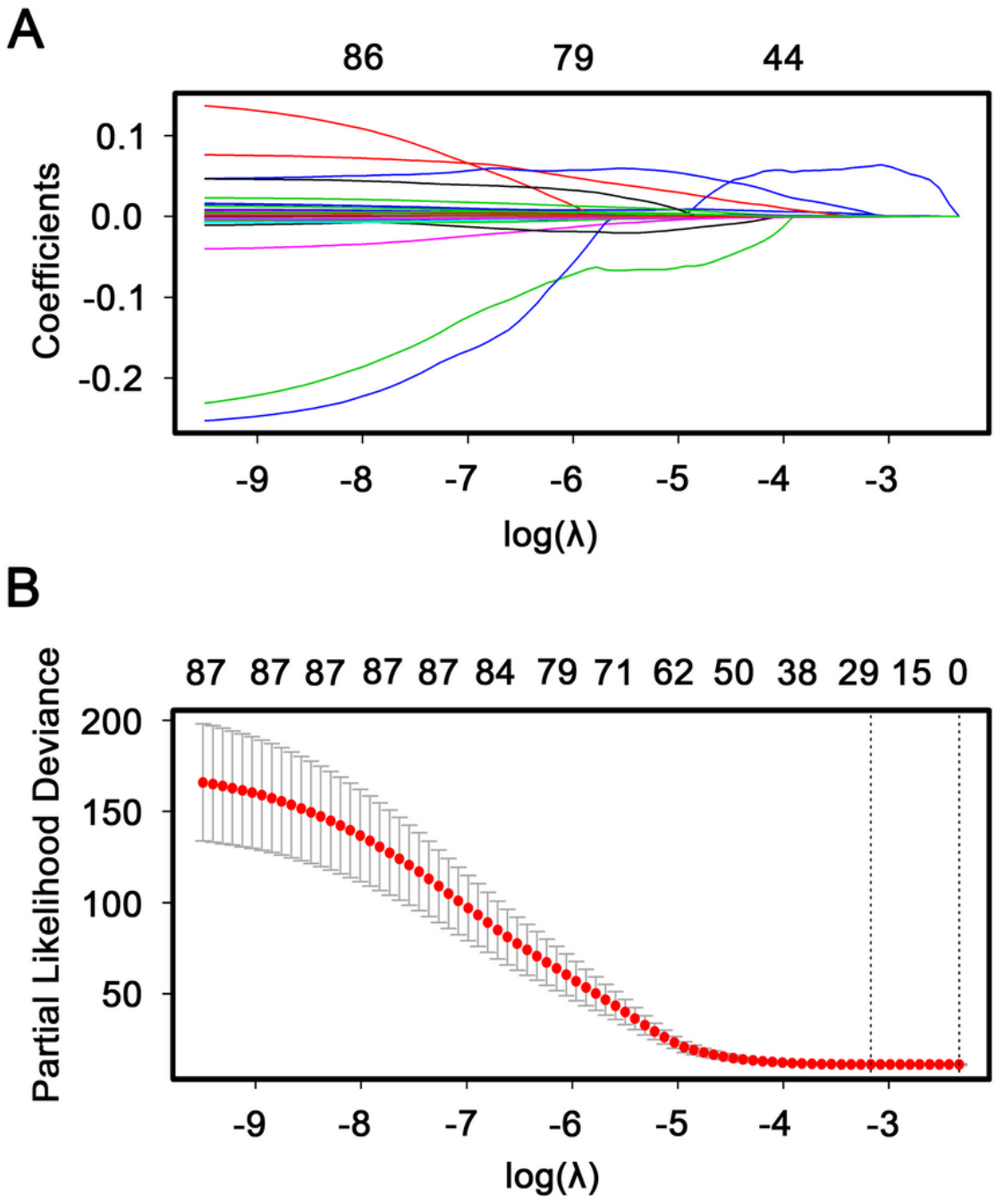

Figure 3

Candidate transcription factors selection using the LASSO Cox regression model. (A) 10-fold cross-validation for tuning parameter selection in the LASSO model via minimum criteria (the 1-SE criteria). (B) LASSO coefficient profiles of the 87 transcription factors. A coefficient profile plot was created against log (lambda) sequence. Vertical line was drawn at the value selected based on 10-fold cross-validation, where optimal lambda resulted in 28 non-zero coefficients.

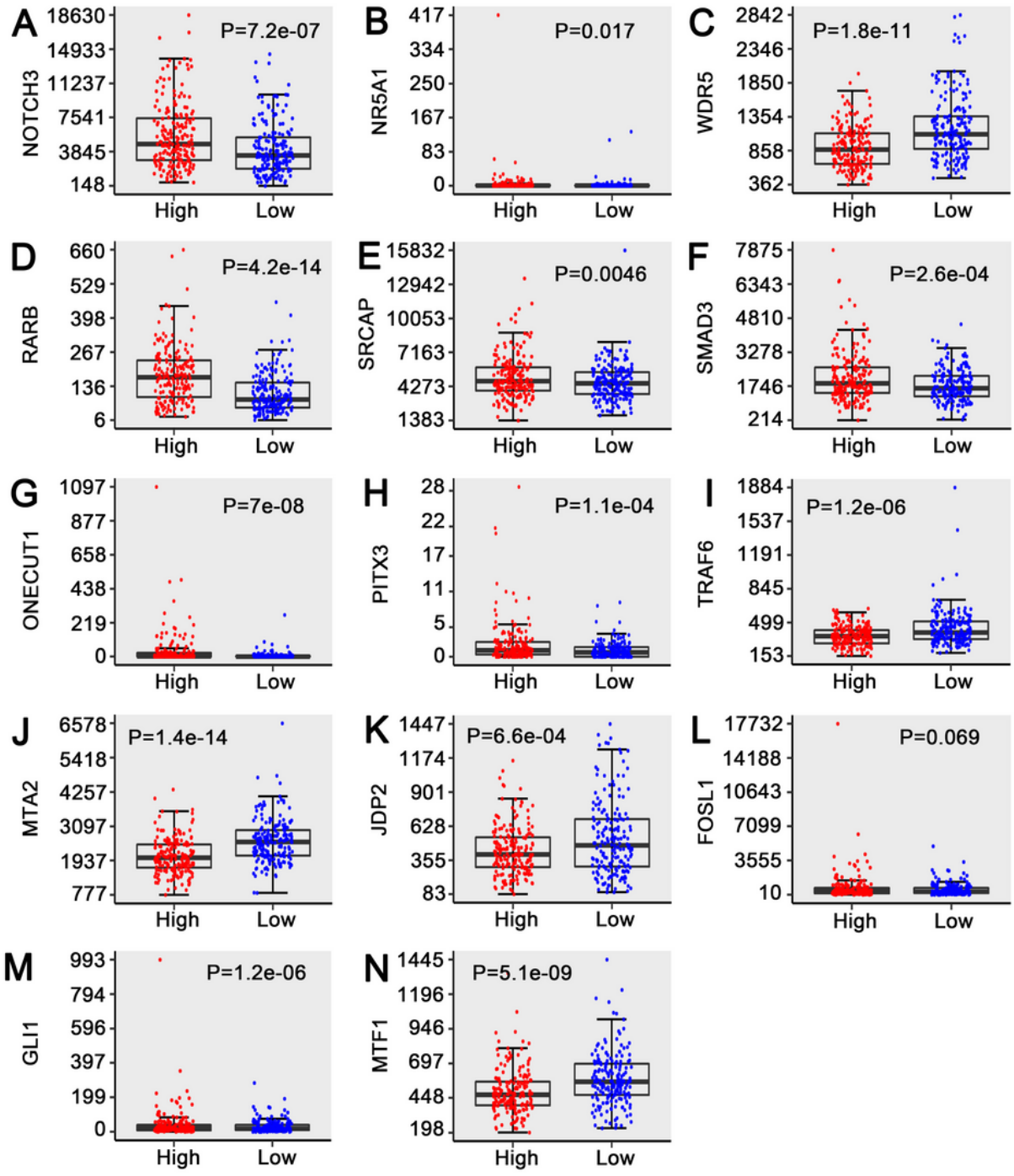

Figure 4

Boxplots of 14 transcription factors expression values against risk group in the TCGA dataset. “High” and “Low” represent the high-risk and low-risk group, respectively. The differences between the 2 groups were measured by Mann-Whitney U test, and P values were provided in the graphs.

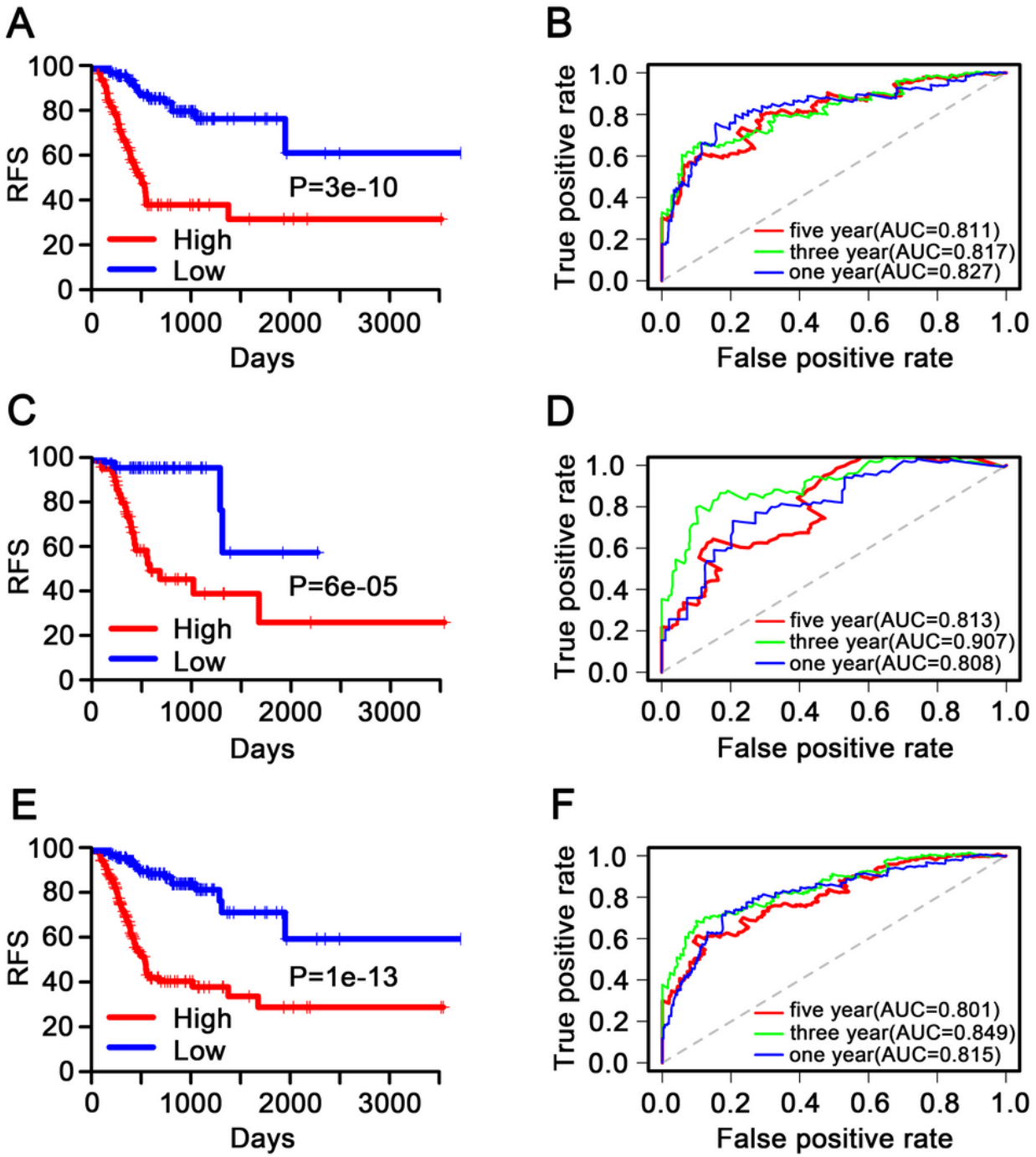

Figure 5

Kaplan-Meier and ROC analysis of patients with GC in internal validation set, external validation set and entire TCGA dataset, respectively. (A, C, E) Kaplan-Meier analysis with two-sided log-rank test was carried out to assess the differences in RFS between the low-risk and high-risk patients. (B, D, F) 1-, 3-, 5-year ROC curves of the 14-TF signature were used to weigh the sensitivity and specificity in predicting the RFS of GC patients.

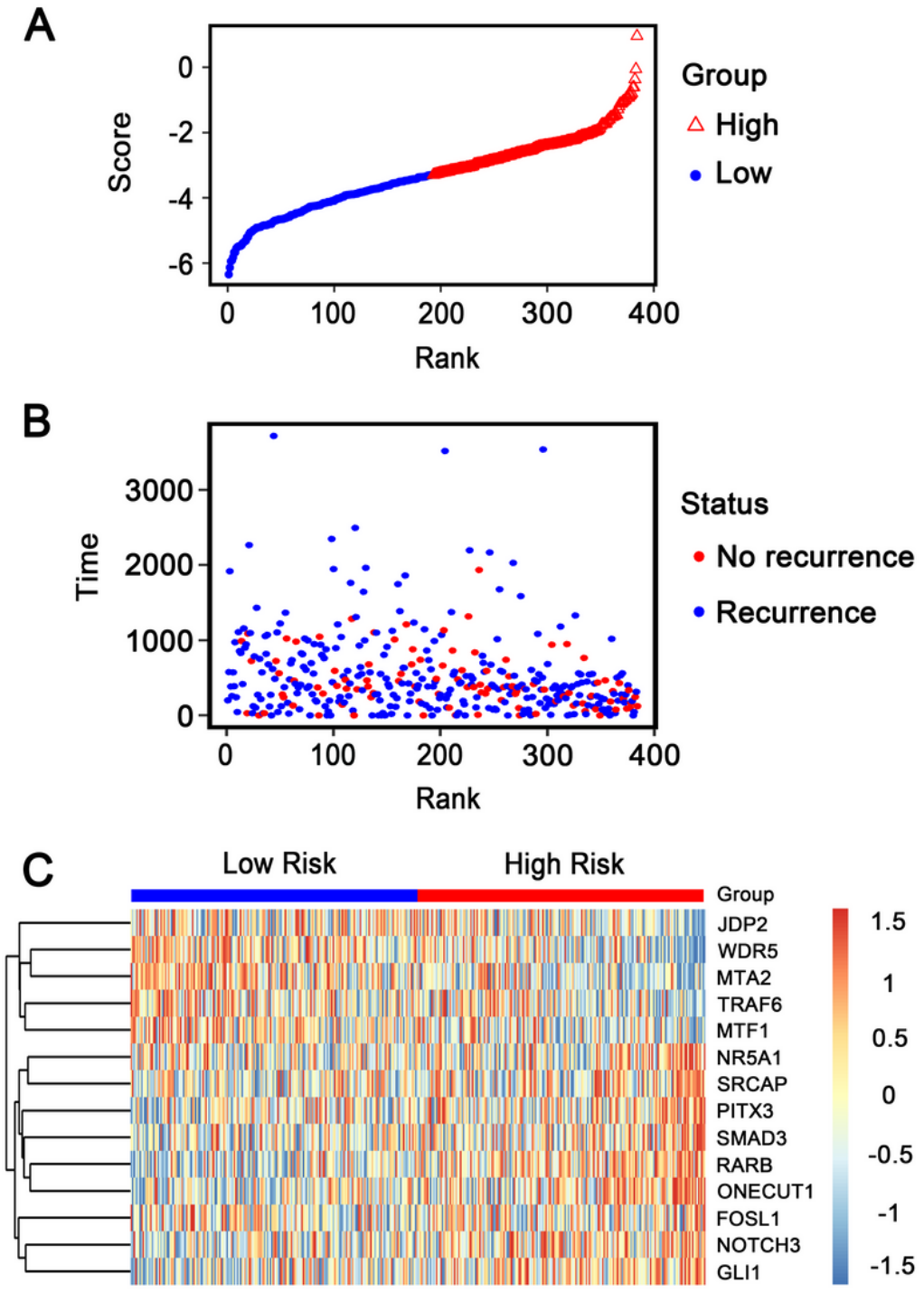

Figure 6

TFs risk score analysis of 384 GC patients in the TCGA dataset (A) TFs risk score distribution against the rank of risk score. Median risk score was the cut-off point. (B) Recurrence free survival status of GC patients. (C) Heatmap of 14 TFs expression profiles of GC patients.

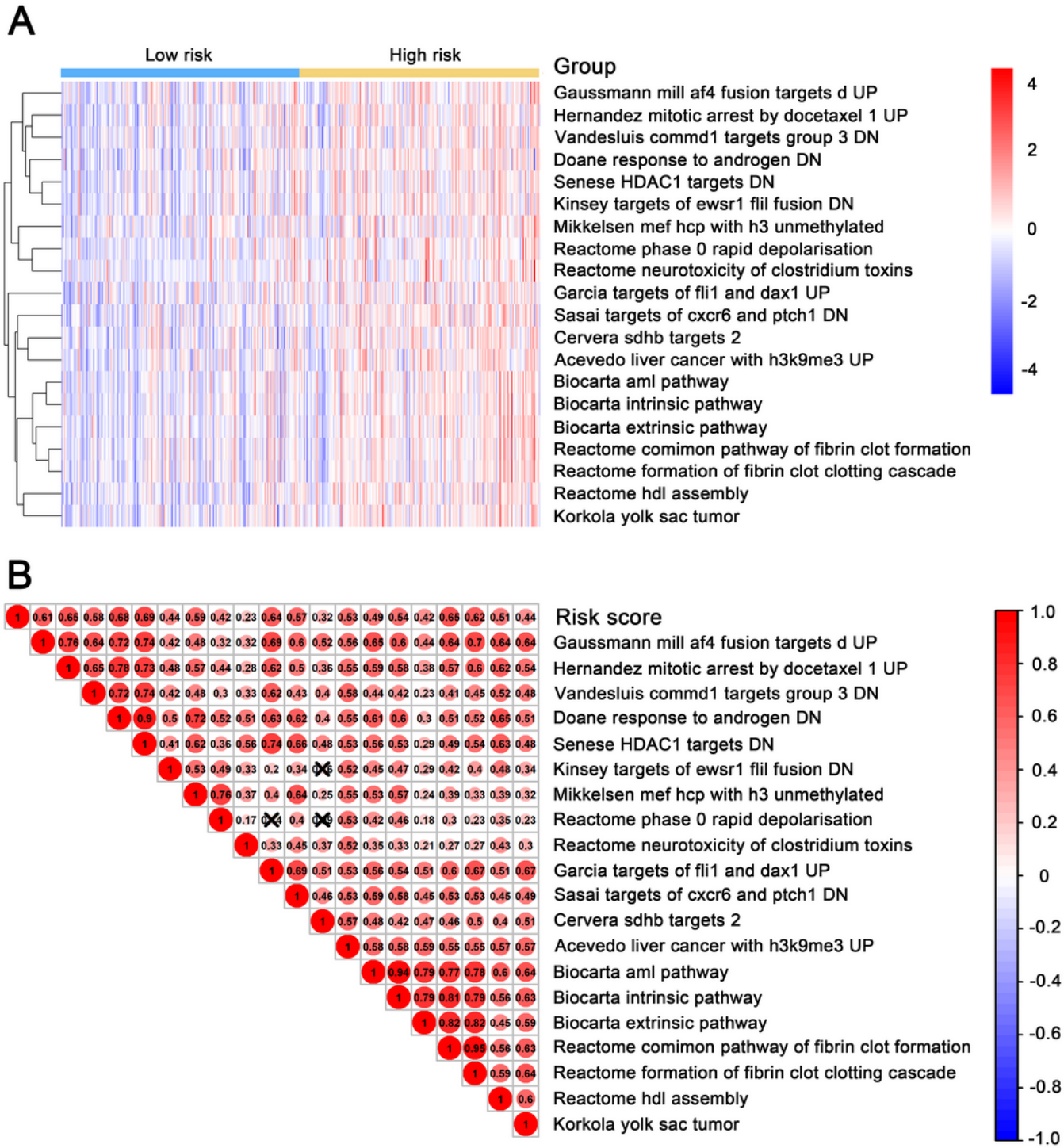

Figure 7

Identification of the 14 TF signature-relevant biological pathways. (A) Heatmap of top 20 enriched pathways associated with high risk group. (B) Association graph between risk scores and top 20 pathways. NO. of P-LN represents the number of positive lymph node.

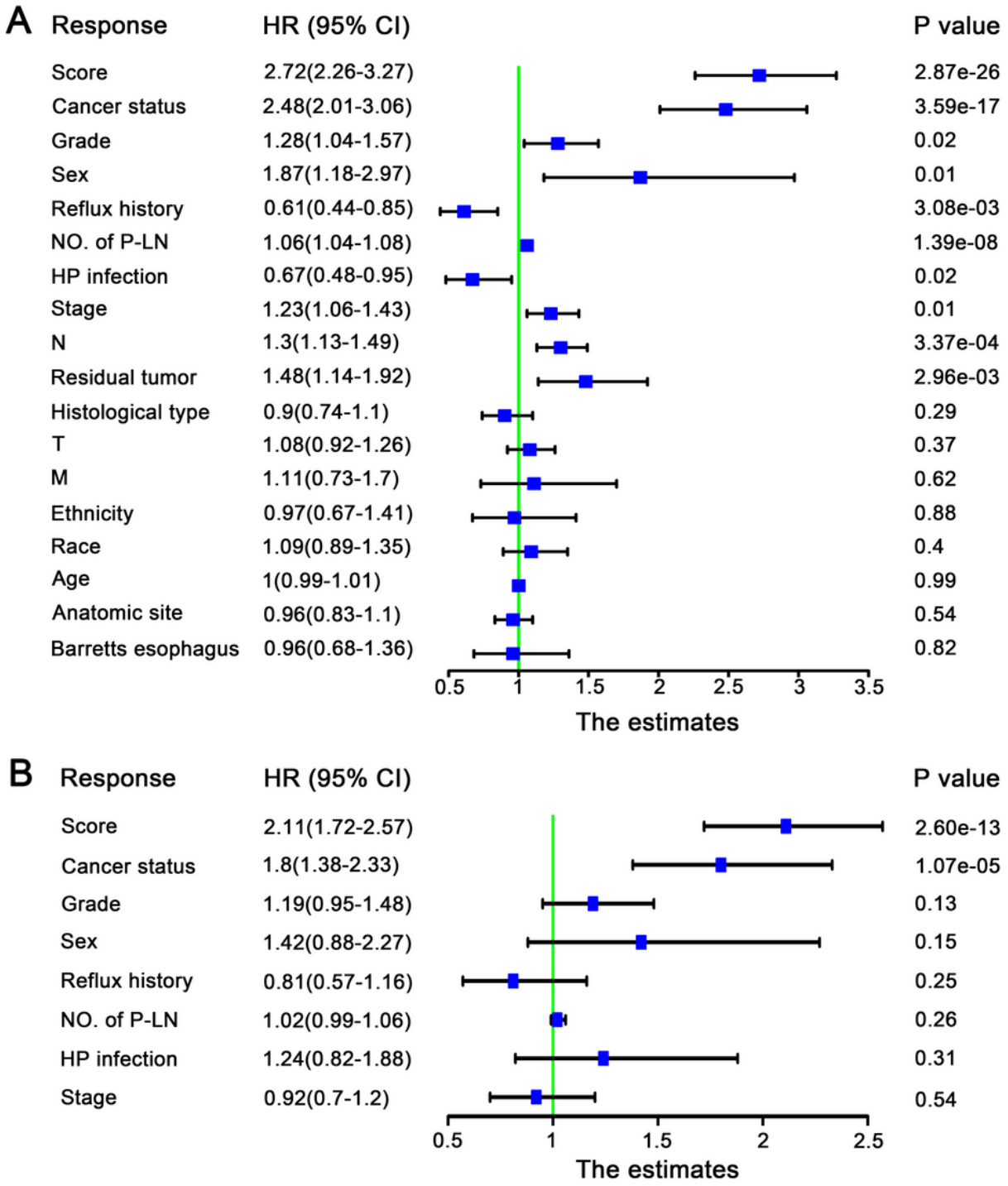

Figure 8

Forest plot summary of analysis of RFS. (A) Univariable Cox analysis of the risk score and other clinical factors. (B) Multivariable Cox analysis of the risk score and other clinical factors. The blue squares on the transverse lines represent the hazard ratio (HR), and the black transverse lines represent 95% CI.

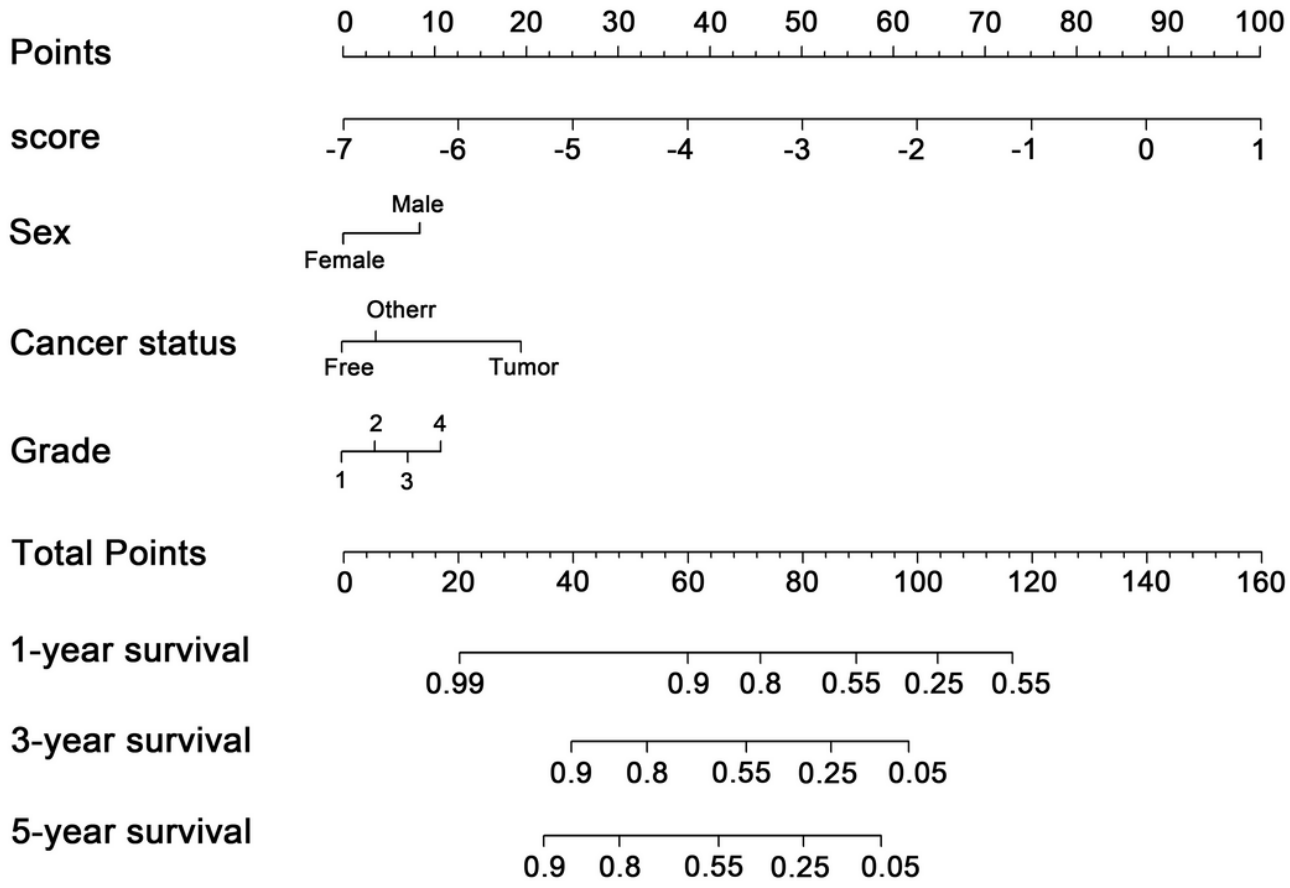

Figure 9

TFs-associated nomogram for the prediction of GC's RFS. The nomogram was developed in the entire TCGA cohort, with the TFs risk score, sex, cancer status and tumor grade.

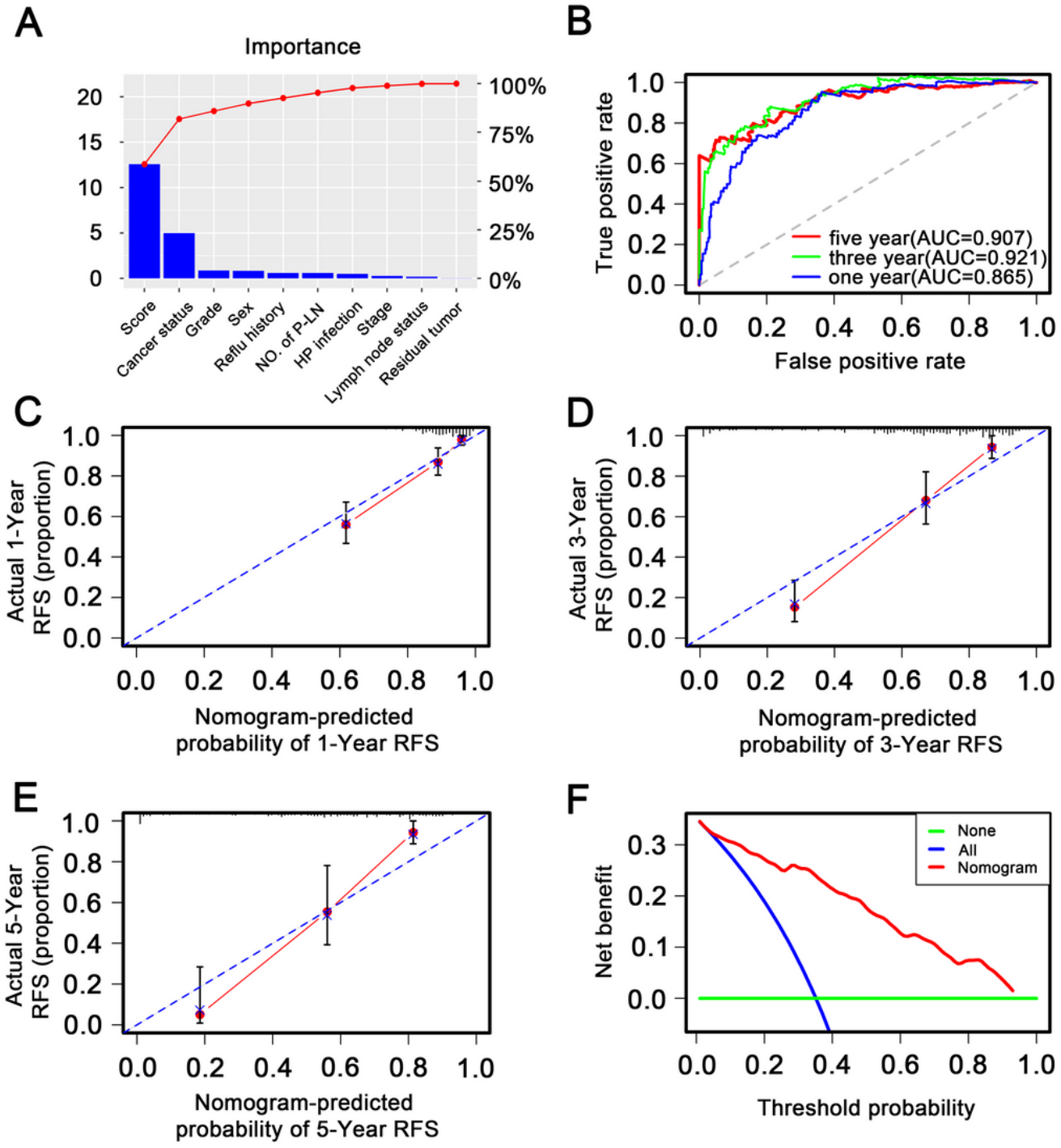

Figure 10

Validation of TFs-associated nomogram in entire TCGA dataset. (A) The higher the bar chart, the greater the percentage. NO. of P-LN represents the number of positive lymph node. (B) 1-, 3-, 5-year ROC curves for the TFs-related nomogram. (C, D, E) represented the 1-, 3-, 5-year nomogram calibration curves, respectively. The closer the dotted line fit to the ideal line, the better the predictive accuracy of the nomogram is. (F) The decision curve analysis (DCA) for the nomogram. The net benefit was plotted

versus the threshold probability. The red line referred to the nomogram. The blue line referred to the treat-all and the green line referred to the treat-none.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2.xlsx
- FigureS7.tif
- FigureS1.tif
- FigureS6.tif
- FigureS2.tif
- FigureS8.tif
- FigureS5.tif
- FigureS4.tif
- TableS4.xlsx
- TableS1.xlsx
- TableS3.xlsx
- FigureS3.tif