# A risk score model with five long non-coding RNAs for predicting prognosis in gastric cancer: a integrated analysis combing TCGA and GEO dataset

**Yiguo Wu** [Equal first author, 1], **Junping Deng** [Equal first author, 2], **Shuhui Lai** [1], **Jing Wu** [Corresp., 3], **Yujuan You** [Corresp. 4]

[1] Department of Medicine, Nanchang University, Nan Chang, China

[2] Department of General Surgery, The First Affiliated Hospital of Nanchang University, Nan Chang, China

[3] Shenzhen Prevention and Treatment Center for Occupational Diseases, Shen Zhen, China

[4] Department of Anesthesiology, The Second Affiliated Hospital of Nanchang University, Nan Chang, China

Corresponding Authors: Jing Wu, Yujuan You
Email address: 446346807@qq.com, 506737972@qq.com

**Background.** Gastric cancer(GC) is an one of the most common digestive carcinoma, and the prognosis for these patients may be poor. There are evidence that some long non-coding RNAs(lncRNAs) could predict the prognosis of gastric cancer. However, few lncRNA signatures have been used to predict the prognosis of the cancer. We herein aimed at constructing a risk score model combining with lncRNAs to predict the prognosis of gastric cancer and providing some new potential therapeutic targets in the future. **Methods.** We performed bayes analysis and survival analysis to screen out differential expressed genes that had significantly different survival times by using gastric cancer patient expression profile data from The Cancer Genome Atlas(TCGA). We then established a formula including five lncRNAs to predict prognosis in GC patients. In addition, to verified the prognostic effect of this risk score model ,two independent the Gene Expression Omnibus(GEO) datasets(GSE62254( N=300 ) and GSE 15459 (N=200)) were employed to act as validation groups. **Results.** Based on the character of five-lncRNA, high or low risk subgroups can be divided among GC patients. The prognostic value of the five-lncRNA signature was confirmed in both TCGA dataset and the other two independent GEO datasets. Furthermore, stratification analysis found that the prognostic value of this risk model was independent in GC patients with Ⅱ-Ⅳ stage. Moreover, we constructed a nomogram model combing the clinic factors and five lncRNAs to heighten the accuracy of prognostic prediction. Enrichment analysis based on Kyoto Encyclopedia of Genes and Genomes (KEGG) suggests five lncRNAs may be touched upon multiple cancer occurrence and progress-related pathways. **Conclusion.** Our results showed that the risk score model combining five-lncRNA signature predicts prognosis of GC patients well especially in stage Ⅱ-Ⅳ and may provide potential therapeutic targets in future.

# A risk score model with five long non-coding RNAs for predicting prognosis in gastric cancer: a integrated analysis combing TCGA and GEO dataset

Yiguo Wu[1], Junping Deng[2], Shuhui Lai[1], Jing Wu [3], Yujuan You [4]
[1] Department of Medicine, Nanchang University, Nanchang City, Jiangxi Province, China
[2] Department of General Surgery, The First Affiliated Hospital of Nanchang University, Nanchang City, Jiangxi Province, China
[3] Department of Health Surveillance, Shenzhen Prevention and Treatment Center for Occupational Diseases, Shenzhen City, Guangdong Province, China
[4]Department of Anesthesiology, The Second Affiliated Hospital of Nanchang University, Nanchang City, Jiangxi Province, China

Corresponding Author:
Yujuan You
No. 1 Minde Road, Nanchang City, Jiangxi Province，330006, China
Email address: 506737972@qq.com
Jing Wu
No.2019 Buxin Road, Luohu District, Shenzhen City, Guangdong Province, 518020, China
Email address:446346807@qq.com

## Abstract

**Background.** Gastric cancer(GC) is an one of the most common digestive carcinoma, and the prognosis for these patients may be poor. There are evidence that some long non-coding RNAs(lncRNAs) could predict the prognosis of gastric cancer. However, few lncRNA signatures have been used to predict the prognosis of the cancer. We herein aimed at constructing a risk score model combining with lncRNAs to predict the prognosis of gastric cancer and providing some new potential therapeutic targets in the future.

**Methods.** We performed bayes analysis and survival analysis to screen out differential expressed genes that had significantly different survival times by using gastric cancer patient expression profile data from The Cancer Genome Atlas(TCGA). We then established a formula including five lncRNAs to predict prognosis in GC patients. In addition, to verified the prognostic effect of this risk score model ,two independent the Gene Expression Omnibus(GEO) datasets(GSE62254(N=300) and GSE15459(N=200)) were employed to act as validation groups.

**Results.** Based on the character of five-lncRNA, high or low risk subgroups can be divided among GC patients. The prognostic value of the five-lncRNA signature was confirmed in both TCGA dataset and the other two independent GEO datasets. Furthermore, stratification analysis

found that the prognostic value of this risk model was independent in GC patients with Ⅱ-Ⅳ stage. Moreover, we constructed a nomogram model combing the clinic factors and five lncRNAs to heighten the accuracy of prognostic prediction. Enrichment analysis based on Kyoto Encyclopedia of Genes and Genomes (KEGG) suggests five lncRNAs may be touched upon multiple cancer occurrence and progress-related pathways.

**Conclusion.** Our results showed that the risk score model combining five-lncRNA signature predicts prognosis of GC patients well especially in stage Ⅱ-Ⅳ and may provide potential therapeutic targets in future.

## Introduction

Gastric cancer(GC) is an one of the most common digestive carcinoma around the world，especially in Asian countries ,and it is estimated that about 679,100 were newly diagnosed with gastric cancer and almost 498,000 died from it in china in 2015(Saka et al., 2011; Chen et al., 2016).So far, the standard therapy for gastric cancer are still surgery and chemotherapy. However, most patients with advanced gastric cancer will still have recurrence and metastasis after treatment, resulting in poor prognosis for them. Despite many therapeutic endeavors, the overall survival of GC patients remains bleak after treatment (Saka et al., 2011). How to identify gastric cancer patients with poor survival prognosis and taking effective treatments as early as possible is the key to improve survival time. Hence, to investigate more potential therapeutical and prognostic biomarkers in gastric cancer is of vital significance.

Long non-coding RNAs(lncRNAs) which are not less than 200 nucleotides are a class of no or limited protein-coding potential RNA transcripts. Increasing evidence showed that lncRNAs, as oncogenes or tumor suppressor genes，play crucial roles in the pathophysiological process of various human diseases, especially in the initiations and developments of tumors, For example, lncRNA-ATB disorders have been shown to contribute to cancer cell proliferation, migration, invasion, drug resistance in tumor and prompt epithelial-mesenchymal transition (EMT) through competitively bound to miRNAs(Li et al., 2017; Balas & Johnson, 2018). Moreover, some researchers suggested that lncRNAs could act as new prognostic biomarkers in cancers, such as CCAT2(Yu et al., 2017), HOXB-AS3(Huang et al., 2017) and ASLNC07322(Li et al., 2019) in colon cancer. A large number of lncRNAs closely related to the prognosis of gastric cancer have been found as well, for example MEG3(Wei & Wang, 2017)、SNHG7(Wang et al., 2017)、DANCR(Mao et al., 2017)[10](Mao et al., 2017). Furthermore, emerging increasingly risk score models has been constructed to predict prognosis of human tumors, too. Recently, among elderly non-small cell lung cancer, the prognostic difference could be identified by the 8-lncRNA signature (Miao et al., 2019). However, the lncRNA related to prognosis in patients with gastric cancer remains infant and requires long-term efforts.

In this study, for the purpose of selecting ideal differential expression lncRNAs for prognostic prediction, we analyzed 486 GC patients from The Cancer Genome Atlas (TCGA) database according to the corresponding risk score. Then the two independent Gene Expression Omnibus (GEO) dataset were applied to validate the lncRNAs selected. Next, we explored predictive

effect of five lncRNAs in different clinical subgroups combining with the clinical character of patients. Then, we further constructed a nomogram model combing the clinic factors and five lncRNAs to heighten the accuracy of prognostic prediction. Finally, we performed a pathway enrichment analysis to understand the potential function in GC.

## Materials & Methods

### Preparation of GC datasets

We acquired the training dataset of gastric cancer from TCGA, including 486 sample and 60498 gene (case: normal=450:36).The microarray data of validation set and the survival data of patients are publicly available at the GEO with accession numbers GSE62254 (300 case; 19293gene) and GSE15459 (200 case; 24438 gene).

### Normalization of GEO data

Due to the differentiated expression profile of the two GEO datasets (GSE62254, GSE15459), we performed quantile normalization on the original data and downloaded it as a probe-level CEL file. The Affymetrix U133 Plus 2.0 was used as probe matching platform. We downloaded it from Affymetrix website (http://www.affymetrix.com), and a total of 2986 lncRNA-specific probes were included.

### Statistical analysis

R software was used for all the statistics. Bayes analysis was done using limma R packages. Survival R packages was used for Kaplan–Meier survival analysis, and the statistical P values were generated by the Cox–Mantel log-rank test. During Cox survival analysis, the cutoff values of gene expression were determined by median. The significance was defined as P values being less than 0.05 and we acquired 278 statistically significant genes. After getting the common genes between TCGA and GEO(GSE62254), five lncRNAs were screened into Cox survival prediction model to get the risk scores for each patient.

We calculated the risk scores of every patient, and Zero score was used as the cutoff value to classify them into two risk score groups. The Kaplan-Meier analysis was applied for survival differention of the two groups. Overall survival(OS) was cut off at four years, and disease-free survival(DFS) was cut off at two years. Time-dependent receiver operating characteristic (ROC) curves were drawn to show the value of prediction model. We also acquired the risk scores of each patient from the validation set from GEO to compute ROC curves and plot the Kaplan-Meier survival curve to access the effect of cox survival prediction. In order to know the relationship between the risk score and clinical information, we use the risk score to estimate the hazards ratio of each subgroup of patients divided by clinical information included gender, TNM stage, histologic grade, race and age. During this analysis, the cutoff values of each clinical index were determined by median. We wanted to visualize the prognostic strength of different clinical index and the risk score in a single feature, A nomogram was established using the package of rms in R. We calculated the concordance index (C-index) and plot the calibration curve to determine its predictive accuracy and discriminatory capacity.

Linear regression analysis with five lncRNAs and protein coding genes. Then, the aberrantly activated signaling pathways and genes were screened out by Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis using Web-based Gene Set Analysis Toolkit (http://www.webgestalt.org/).

## Results

**Identification of five prognostic lncRNAs from the training series.**

After downloading raw data from TCGA database, a total of 486 samples which have complete clinic and prognostic information were included in the study as a training cohort. Then, we performed bayesian analysis (p<10^ (-5)) and univariable Cox proportional hazards regression analysis (log2|fold change| >1 and adjusted P < 0.05) to identify certain prognostic related lncRNAs. A total of 278 lncRNAs were further analyzed. Furthermore, to verify accuracy of prognosis prediction of the selected lncRNAs from TCGA database into the GEO validation set, 38 shared genes were found after intersecting the 278 lncRNAs with the validation dataset (GSE62254). After multivariable Cox proportional hazards regression analyses, we identified five lncRNAs as independent prognostic factors for gastric cancer, including LINC00205, TRHDE-AS1, OVAAL, LINC00106, MIR100HG (Table 1). The expression information of five lncRNAs in gastric cancer patients was showed in volcano and heat maps 【Figure 1A-B】.the survival curve was also plotted based on the OS and DFS of these 408 patients 【Figure 1C-D】. From the OS survival curve, we can analyze that the slope of the curve tends to be gentle at 48 months, so we take the 48-month as cuff off value and the survival time longer than 48 months is classified as a good prognosis, otherwise the prognosis is poor. Similarly, the DFS curve is bounded by 24 months and disease-free survival times greater than 24 months are classified as good prognosis, otherwise the prognosis is poor.

**Creation of a lncRNAs-based risk model from the test cohort**

According to the schematic workflow of the present study shown in Table 2, on the basis of the coefficient of the 5 lncRNAs acquired from multivariable Cox hazards analyses, we, then, created risk-score formula as followings: risk score = (0.249092 × the expression level of LINC00205) + (0.182045 × the expression level of TRHDE-AS1) + (0.271169 × the expression level of OVAAL) + (-0.20794 × the expression level of LINC00106) + (0.502539 × the expression level of MIR100HG). Among the 5 lncRNAs, a negative coefficient means a protective factor, such as LINC00106. The remaining 4 lncRNAs with positive coefficients, involving LINC00205, TRHDE-AS1, OVAAL and MIR100HG, were served as risk factors. The risk scores of each patient in test cohort were calculated based on above formula. Then, the patients in test cohort were divided into two subgroups, high risk (n = 204) and low risk group (n = 204), as zero was used as the cut-off value. Moreover, we performed Kaplan-Meier survival analysis to assess the effect of the lncRNAs-based model on the OS and DFS of GC in test cohort【Figure 2A-B】. Our results indicated that the high-risk group had a significantly worse prognosis than the low risk in both OS and DFS, and the P value were 1 × 10 ^ (-6) and 6 × 10 ^ (-6), respectively. Furthermore, the scatter plots for death and recurrence incidence of GC

patients were drawn in 【Figure 2C-D】.As showed in plots, both death and recurrence cases for GC patients in high risk group were significantly more than low-risk（P<0.001）. Finally, in order to better and more accurately evaluate the prognostic value of the five lncRNAs signatures by using the risk score model, We performed time-dependent ROC analysis by using the four-year cut-off OS and two-year cut-off DFS as the ROC ending points as demonstrated above. The area under the ROC curve (AUC) is 0.729 and 0.692, respectively, suggesting a valuable prediction of GC patients' survival【Figure 2E-F】.

**Validation of lncRNAs-based model for prognostic prediction in independent cohorts**

To assess the prognostic significance of this novel prognostic model involving five signature in GC patients, we used the other two independent validation sets from GEO database. By using established risk score formula given above, we calculated the risk score similarly. The GC patients in GSE62254 (validation group-1, n=300) and GSE15459 (validation group-2, n=200) datasets were divided into high-risk and low-risk groups as well. Kaplan-Meier survival analysis was used in two independent validation groups. Because of lack of DFS data in two validation groups, we only calculated the OS of the patients in the two validation sets. Consistent with the training group, our results showed that the GC patients in high-risk subgroup in two validation groups had a poorer OS (log-rank test P = 0.009 and 0.02, respectively)【Figure 3A-B】. The scatter plots for death events were shown in【Figure 3C-D】.Similar with training group, the incidence of death cases for GC patients in high risk group were significantly more than low risk group（P<0.01）. The AUC of those two validation cohort is 0.622 and 0.610, respectively【Figure 3E-F】. Our results further confirmed the favorable prognostic value of this risk score model in GC patients.

**The lncRNAs-based model had a favorable prognostic prediction in stageⅡ,stageⅢand stageⅣ patients**

To further investigate the potential of the lncRNAs-based model, stratified Kaplan-Meier survival analysis for OS in training group were performed based on the AJCC TNM stage, including stageⅠ,stageⅡ,stageⅢand stageⅣ【Figure 4A-D】. Similarly, the five-lncRNA signature had good predictive value for OS in these subgroups involving stageⅡ(P=0.008),stageⅢ(P=0.02)and stageⅣ(P=0.01). otherwise, it is not in stage I (P=0.3). We probably draw unauthentic conclusions due to the limitation of sample size in stage I (only 26 patients).

In addition, to estimate the hazards ratio of each subgroup of patients divided by clinical information included gender, TNM stage, histologic grade, race and age【Table 3】, the risk score model where the cutoff values of each clinical index were determined by median were used to divide the every subgroup into two risk group. Forest plot were draw in 【Figure 5】. our results indicated that the risk score model involving five-lncRNA signature may have relatively good prognosis value in some certain clinic subgroups insisting of gender, histologic grade and age. Furthermore, to improve the prognosis value of this risk score model, we combined the independent clinic related factors with this risk score model to construct nomogram model to

predict prognosis. Nomogram model and nomogram calibration curve was drawn in 【Figure 6A-B】. Moreover, to evaluate the effect of this nomogram model, we also calculate C index of this new model. C index was 0.69668 in predicting four year OS of GC patients, indicating it may have favorable potential prognosis significance.

**Potential function of five lncRNAs**

In order to investigate the functions of those five lncRNAs in GC, we calculated the Pearson correlation between the five lncRNA signatures and 19605 protein-coding genes in the TCGA dataset. A total of 421 genes positively correlated with at least one lncRNA (Pearson's coefficient> 0.8, P <10 ^ (-5)) were further selected for KEGG pathway enrichment analyses【Figure7A】. For biological processes, the co-expressed genes were mainly enriched in these pathways, for instance cGMP−PKG, Calcium signaling pathway, and cAMP signaling pathway etc. This indicated that these five lncRNAs may be related to regulating the initiation and progress of tumors【Figure7B】.

## Discussion

In this study, we identified a potential five-lncRNA signature which are differential expression from tumor tissue to normal as prognostic predictor of GC. The final five-lncRNA signature was verified to be associated with outcome in GC patients after a complicated analysis and it may become an underlying therapeutical biomolecule in future. The prognostic significance of the constructed risk score model involving five-lncRNA has been comfirmed in validation series. Moreover, stratified analysis suggested that the risk score model had a favorable prognosis predictive value in GC patients with Ⅱ-Ⅳ stage. Finally, to enhance the predictive accuracy for GC patients, we also combined clinic related factors with five-lncRNA signature to constructed a nomogram model confirmed by calibration curve and C index.

As we all known, gastric cancer is a common malignancy in the digestive system (Siegel et al., 2019). Despite the continuous improvement of treatments, the five-year survival rate of patients with advanced gastric cancer still hovers at 20% (Min et al., 2019; Misawa et al., 2019). Therefore, early diagnosis，early identification of high-risk patients and positive treatment measures as early as possible for gastric cancer patients are the key to improve survival time. Increasing attentions has been arousing increasing attentions to figure out more novel prognostic indicators for GC at the same time. Over the past few decades, a large amount of research evidence showed that protein-encoding genes(Ghoorun et al., 2019; Luo et al., 2019) and microRNAs(Li et al., 2020; Zhou et al., 2019), which acted as oncogenes or tumor suppressors , play vital role in occurrence and development of various tumors and could predict the prognosis as well. A number of nomogram model including clinic related factors were found to predict prognosis of GC. For example , Yue(Yu & Zhang, 2019).etc used tumor size and tumor site, as independent prognosis factors ,to construct OS nomogram for predicting outcome of GC patients, then the C-index for this model was 0.633 (95% CI: 0.579–0.687)，indicating the model was able to assess the prognosis of GC patients in OS. Recently, more and

more lncRNAs related to the prognosis of gastric cancer have been continuously discovered, but prognostic prediction models related to lncRNAs still lack a unified conclusion so far. We herein provide a nomogram including clinic related factors and five-lncRNA signature which may become potential therapeutic target to effectively predict prognosis of GC patients.

As a result, it is urgent to explore new biomarkers to improve the assessment of diagnosis and prognosis of GC patients due to the key limitation on the AJCC TNM staging system and some related socring systems. With the rapid development of computational technologies, a lot of lncRNAs have been identified, among which only a small proportion has been functionally annotated. However, accumulating study showed that lncRNAs, as carcinogenes or tumor suppressors ,not only participate in the tumorigenesis and development of various tumors by regulating the processes of chromatin remodeling, transcription and post-transcriptional modification(Bartonicek et al., 2016; Iyer et al., 2015), but also can be used as a underlying molecule for tumor diagnosis and prognosis. In addition, some studies have found that gastric cancer-related lncRNAs are involved in biological behaviors such as proliferation, migration, invasion, and autophagy of gastric cancer cells, affecting the formation and prognosis of GC (Mao et al., 2017; Wei & Wang, 2017). For example, lncRNA MEG3 could inhibit proliferation, metastasis and prognosis of GC through up-regulated p53 expression as a key tumor suppressor molecular (Wei & Wang, 2017). We herein found multiple lncRNAs as predictors of GC prognosis instead of a single lncRNA. a total of five lncRNAs(LINC00205, TRHDE-AS1, OVAAL, LINC00106, MIR100HG) were identified in this work and we further established a risk score model. Kaplan-Meier analysis suggested that this model has favorable prognostic effect in GC patients. Next, to narrow the bias of small-scale data, we used two independent GEO datasets as validation groups. Our results confirmed the risk score model were stable and steady in predict the prognosis of GC.

Among the five lncRNAs, including LINC00205, TRHDE-AS1, OVAAL and MIR100HG, acted as risk factors for GC patients, otherwise, the LINC00106 was a protective factor. Except for LINC00205 and MIR100HG, the other three lncRNAs have been less reported in the literatures. Furthermore, except for LINC00106, in present study, the 4 lncRNAs were identified as biomarkers and prognosis predictors for the first time in GC so far. Consistent with our result, it reported that the high expression of LINC00106 indicated a prolonged overall survival in GC (Qi et al., 2020). Nevertheless, the function of this lncRNA in gastric cancer and its specific mechanism need further study. Interestingly, in hepatocellular carcinoma (HCC), LINC00205 expression levels, as a suppressor of tumor, are positively associated with OS and recurrence-free survival by a comprehensive genome-wide analysis (Cui et al., 2017). Furthermore, a study showed that, as a competing endogenous RNA and lower expression level in tumor tissue, LINC00205 may negatively regulate the progression of HCC via miR-184/EPHX1 axis (Long et al., 2019), While another research figured that LINC00205, as a oncogene, promotes proliferation, migration and invasion of HCC cells by targeting miR-122-5p(Zhang et al., 2019). Moreover, The function of LINC00205 was acted as a protective factor in pancreatic cancer survival [HR = 0.58, p (Log rank) = 0.0091](Giulietti et al., 2018). The role and prognostic

prediction of LINC00205 in various cancer was discrepant. This might be associated with the specificity of different tumor. It reported that up-regulation of the TRHDE - AS1 could inhibits the growth of lung carcinoma through competitively combination with miRNA - 103/KLF4 axis(Zhuan et al., 2019). A study have observed that OVVAL is highly expressed in colon cancer and melanoma, and further experimental results show OVAAL promote the proliferation of cancer cells via dual mechanisms controlling RAF/MEK/ERK signaling and p27-mediated cell senescence(Sang et al., 2018). lncRNA MIR100HG has been studied as a oncogene in acute megakaryoblastic leukemia(Emmrich et al., 2014), laryngeal squamous cell carcinoma(Huang et al., 2019), and mediate cetuximab resistance via Wnt/β-catenin signaling(Lu et al., 2017) in colorectal cancer. In summary, although the role of these lncRNA in cancer need more elucidate, our results may provide a novel thinking for the study of gastric cancer.

To further investigate the function of the five lncRNAs in gastric cancer, we performed a pathway enrichment analysis result from the unclear specific pathway mechanism. The results indicated that the enrichment pathways are involved in regulation, including cGMP−PKG signaling pathway,Calcium signaling pathway, and cAMP signaling pathway etc. This suggests that five-lncRNA may play a important role in tumor occurrence and development regulation in GC patient by above molecular pathways. There has been evidence that lncRNA can promote tumorigenesis through the cGMP−PKG signaling pathway. For example, overexpression of SRRM2-AS accelerated angiogenesis of nasopharyngeal carcinoma via cGMP-PKG signaling pathway(Chen et al., 2019) . A study proved that down-expression of LINC01585 can active cAMP signaling pathway, resulting in the growth of breast cancer(Ma et al., 2019). In short, lncRNA can participate in the genesis and development of various tumors through the above pathways. Thus, our results are consistent with many recent findings.

A previous study firstly reported a 24-lncRNA signature were significantly associated with the prognosis of GC patients by using lncRNA expression profiling of GC from GEO (Zhu et al., 2016),and it difinitely provide a special perspective of novel potential targets in treating GC in the future. However, due to the limitation of amount of data in GEO dataset，the lncRNAs candidates identified may not represent the complete lncRNA populations underlying GC biological behavior. In our study, we take full advantage of TCGA and GEO data to comprehensively investigate the prognostic lncRNAs. Moreover, to evaluate the prediction effect of this five-lncRNA signature, we determine the end point of the ROC curve based on the cut-off value of the OS and DFS curve rather than the survival outcome, which may be able to more reasonably evaluate the effect of the prediction model. Finally, in order to improve the accuracy of the five gene prognosis models, we also combined the clinically relevant prognostic factors to establish a nomogram model. The results showed that the nomogram prediction model combining clinically relevant factors can effectively predict OS in patients with gastric cancer.

Of course, there are some limitations in this study. Because of lack of DFS data in two GEO validation groups, we only used two validation cohort to verified the prognosis value of five-lncRNA signature in the OS of the GC patients. Secondly, due to the limitation of the tumor data,

316 experimental research on these lncRNAs is highly needed to further understand these functions
317 in GC in the future.

## 318 Conclusions

319 We established a risk score model including five lncRNAs to predict GC patients' OS and
320 DFS, particularly in those with Ⅱ-Ⅳ stage. Our findings also provided evidence of developing
321 effective prognostic biomarkers for GC patients and potential therapeutic targets in the future.

## 322 Acknowledgements

323 At the point of finishing this paper, we would thank to Jie Peng for technical assistance with
324 the data analysis.

## 325 References

326 Balas, M.M., and Johnson, A.M. 2018. Exploring the mechanisms behind long noncoding RNAs and cancer.
327 *Noncoding RNA Res* 3:108-117. 10.1016/j.ncrna.2018.03.001
328 Bartonicek, N., Maag, J.L., and Dinger, M.E. 2016. Long noncoding RNAs in cancer: mechanisms of action and
329 technological advancements. *Molecular Cancer* 15:43. 10.1186/s12943-016-0530-6
330 Chen, S., Lv, L., Zhan, Z., Wang, X., You, Z., Luo, X., and You, H. 2019. Silencing of long noncoding RNA
331 SRRM2-AS exerts suppressive effects on angiogenesis in nasopharyngeal carcinoma via activating MYLK-
332 mediated cGMP-PKG signaling pathway. *JOURNAL OF CELLULAR PHYSIOLOGY*. 10.1002/jcp.29382
333 Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., and He, J. 2016. Cancer
334 statistics in China, 2015. *CA: A Cancer Journal for Clinicians* 66:115-132. 10.3322/caac.21338
335 Cui, H., Zhang, Y., Zhang, Q., Chen, W., Zhao, H., and Liang, J. 2017. A comprehensive genome-wide analysis of
336 long noncoding RNA expression profile in hepatocellular carcinoma. *Cancer Med* 6:2932-2941.
337 10.1002/cam4.1180
338 Emmrich, S., Streltsov, A., Schmidt, F., Thangapandi, V.R., Reinhardt, D., and Klusmann, J.H. 2014. LincRNAs
339 MONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia. *Molecular Cancer* 13:171.
340 10.1186/1476-4598-13-171
341 Ghoorun, R.A., Wu, X.H., Chen, H.L., Ren, D.L., and Wu, X.B. 2019. Prognostic Significance of FKBP14 in
342 Gastric Cancer. *Onco Targets Ther* 12:11567-11577. 10.2147/OTT.S221943
343 Giulietti, M., Righetti, A., Principato, G., and Piva, F. 2018. LncRNA co-expression network analysis reveals novel
344 biomarkers for pancreatic cancer. *CARCINOGENESIS* 39:1016-1025. 10.1093/carcin/bgy069
345 Huang, J.Z., Chen, M., Chen, Gao, X.C., Zhu, S., Huang, H., Hu, M., Zhu, H., and Yan, G.R. 2017. A Peptide
346 Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *MOLECULAR CELL* 68:171-184.
347 10.1016/j.molcel.2017.09.015
348 Huang, Y., Zhang, C., and Zhou, Y. 2019. LncRNA MIR100HG promotes cancer cell proliferation, migration and
349 invasion in laryngeal squamous cell carcinoma through the downregulation of miR-204-5p. *Onco Targets Ther*
350 12:2967-2973. 10.2147/OTT.S202528
351 Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R.,
352 Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S.M., Wu, Y.M., Robinson, D.R., Beer, D.G., Feng, F.Y., Iyer, H.K.,
353 and Chinnaiyan, A.M. 2015. The landscape of long noncoding RNAs in the human transcriptome. *NATURE*
354 *GENETICS* 47:199-208. 10.1038/ng.3192
355 Li, J., Li, Z., Zheng, W., Li, X., Wang, Z., Cui, Y., and Jiang, X. 2017. LncRNA-ATB: An indispensable cancer-
356 related long noncoding RNA. *Cell Prolif* 50. 10.1111/cpr.12381
357 Li, J.P., Zhang, H.M., Liu, M.J., Xiang, Y., Li, H., Huang, F., Li, H.H., Dai, Z.T., Gu, C.J., Liao, X.H., and Zhang,
358 T.C. 2020. miR-133a-3p/FOXP3 axis regulates cell proliferation and autophagy in gastric cancer. *JOURNAL OF*
359 *CELLULAR BIOCHEMISTRY*. 10.1002/jcb.29613
360 Li, X., Lv, X., Li, Z., Li, C., Li, X., Xiao, J., Liu, B., Yang, H., and Zhang, Y. 2019. Long Noncoding RNA
361 ASLNC07322 Functions in VEGF-C Expression Regulated by Smad4 during Colon Cancer Metastasis. *Mol Ther*
362 *Nucleic Acids* 18:851-862. 10.1016/j.omtn.2019.10.012
363 Long, X., Li, Q., Zhi, L.J., Li, J.M., and Wang, Z.Y. 2019. LINC00205 modulates the expression of EPHX1 through
364 the inhibition of miR‐184 in hepatocellular carcinoma as a ceRNA. *JOURNAL OF CELLULAR PHYSIOLOGY*
365 235:3013-3021. 10.1002/jcp.29206
366 Lu, Y., Zhao, X., Liu, Q., Li, C., Graves-Deal, R., Cao, Z., Singh, B., Franklin, J.L., Wang, J., Hu, H., Wei, T.,

Yang, M., Yeatman, T.J., Lee, E., Saito-Diaz, K., Hinger, S., Patton, J.G., Chung, C.H., Emmrich, S., Klusmann, J.H., Fan, D., and Coffey, R.J. 2017. lncRNA MIR100HG-derived miR-100 and miR-125b mediate cetuximab resistance via Wnt/beta-catenin signaling. *NATURE MEDICINE* 23:1331-1341. 10.1038/nm.4424

Luo, X., He, Y., Tang, H., Cao, Y., Gao, M., Liu, B., and Hu, Z. 2019. Effects of HER2 on the invasion and migration of gastric cancer. *American Journal of Translational Research* 11:7604-7613.

Ma, R., Zhai, X., Zhu, X., and Zhang, L. 2019. LINC01585 functions as a regulator of gene expression by the CAMP/CREB signaling pathway in breast cancer. *GENE* 684:139-148. 10.1016/j.gene.2018.10.063

Mao, Z., Li, H., Du B, Cui, K., Xing, Y., Zhao, X., and Zai, S. 2017. LncRNA DANCR promotes migration and invasion through suppression of lncRNA-LET in gastric cancer cells. *Biosci Rep* 37. 10.1042/BSR20171070

Miao, R., Ge, C., Zhang, X., He, Y., Ma, X., Xiang, X., Gu, J., Fu, Y., Qu, K., Liu, C., Wu, Q., and Lin, T. 2019. Combined eight-long noncoding RNA signature: a new risk score predicting prognosis in elderly non-small cell lung cancer patients. *Aging (Albany NY)* 11:467-479. 10.18632/aging.101752

Min, S.H., Won, Y., Kim, G., Lee, Y., Park, Y.S., Ahn, S.H., Park, D.J., and Kim, H.H. 2019. 15-year experience of laparoscopic gastrectomy in advanced gastric cancer: analysis on short-term and long-term oncologic outcome. *SURGICAL ENDOSCOPY AND OTHER INTERVENTIONAL TECHNIQUES*. 10.1007/s00464-019-07292-x

Misawa, K., Mochizuki, Y., Sakai, M., Teramoto, H., Morimoto, D., Nakayama, H., Tanaka, N., Matsui, T., Ito, Y., Ito, S., Tanaka, K., Uemura, K., Morita, S., and Kodera, Y. 2019. Randomized clinical trial of extensive intraoperative peritoneal lavage versus standard treatment for resectable advanced gastric cancer (CCOG 1102 trial). *Br J Surg* 106:1602-1610. 10.1002/bjs.11303

Qi, M., Yu, B., Yu, H., and Li, F. 2020. Integrated analysis of a ceRNA network reveals potential prognostic lncRNAs in gastric cancer. *Cancer Medicine*. 10.1002/cam4.2760

Saka, M., Morita, S., Fukagawa, T., and Katai, H. 2011. Present and future status of gastric cancer surgery. *JAPANESE JOURNAL OF CLINICAL ONCOLOGY* 41:307-313. 10.1093/jjco/hyq240

Sang, B., Zhang, Y.Y., Guo, S.T., Kong, L.F., Cheng, Q., Liu, G.Z., Thorne, R.F., Zhang, X.D., Jin, L., and Wu, M. 2018. Dual functions for OVAAL in initiation of RAF/MEK/ERK prosurvival signals and evasion of p27-mediated cellular senescence. *Proceedings of the National Academy of Sciences* 115:E11661-E11670. 10.1073/pnas.1805950115

Siegel, R.L., Miller, K.D., and Jemal, A. 2019. Cancer statistics, 2019. *CA Cancer J Clin* 69:7-34. 10.3322/caac.21551

Wang, M.W., Liu, J., Liu, Q., Xu, Q.H., Li, T.F., Jin, S., and Xia, T.S. 2017. LncRNA SNHG7 promotes the proliferation and inhibits apoptosis of gastric cancer cells by repressing the P15 and P16 expression. *Eur Rev Med Pharmacol Sci* 21:4613-4622.

Wei, G.H., and Wang, X. 2017. lncRNA MEG3 inhibit proliferation and metastasis of gastric cancer via p53 signaling pathway. *Eur Rev Med Pharmacol Sci* 21:3850-3856.

Yu, C., and Zhang, Y. 2019. Development and validation of prognostic nomogram for young patients with gastric cancer. *Annals of Translational Medicine* 7:641. 10.21037/atm.2019.10.77

Yu, Y., Nangia-Makker, P., Farhana, L., and Majumdar, A. 2017. A novel mechanism of lncRNA and miRNA interaction: CCAT2 regulates miR-145 expression by suppressing its maturation process in colon cancer cells. *Molecular Cancer* 16:155. 10.1186/s12943-017-0725-5

Zhang, L., Wang, Y., Sun, J., Ma, H., and Guo, C. 2019. LINC00205 promotes proliferation, migration and invasion of HCC cells by targeting miR-122-5p. *Pathology - Research and Practice* 215:152515. 10.1016/j.prp.2019.152515

Zhou, H.Y., Wu, C.Q., and Bi, E.X. 2019. MiR-96-5p inhibition induces cell apoptosis in gastric adenocarcinoma. *World J Gastroenterol* 25:6823-6834. 10.3748/wjg.v25.i47.6823

Zhu, X., Tian, X., Yu, C., Shen, C., Yan, T., Hong, J., Wang, Z., Fang, J., and Chen, H. 2016. A long non-coding RNA signature to improve prognosis prediction of gastric cancer. *Molecular Cancer* 15. 10.1186/s12943-016-0544-0

Zhuan, B., Lu, Y., Chen, Q., Zhao, X., Li, P., Yuan, Q., and Yang, Z. 2019. Overexpression of the long noncoding RNA TRHDE‐AS1 inhibits the progression of lung cancer via the miRNA‐103/KLF4 axis. *JOURNAL OF CELLULAR BIOCHEMISTRY*. 10.1002/jcb.29029

## Table 1(on next page)

Five lncRNAs significantly associated with prognosis of GC patients in the training group

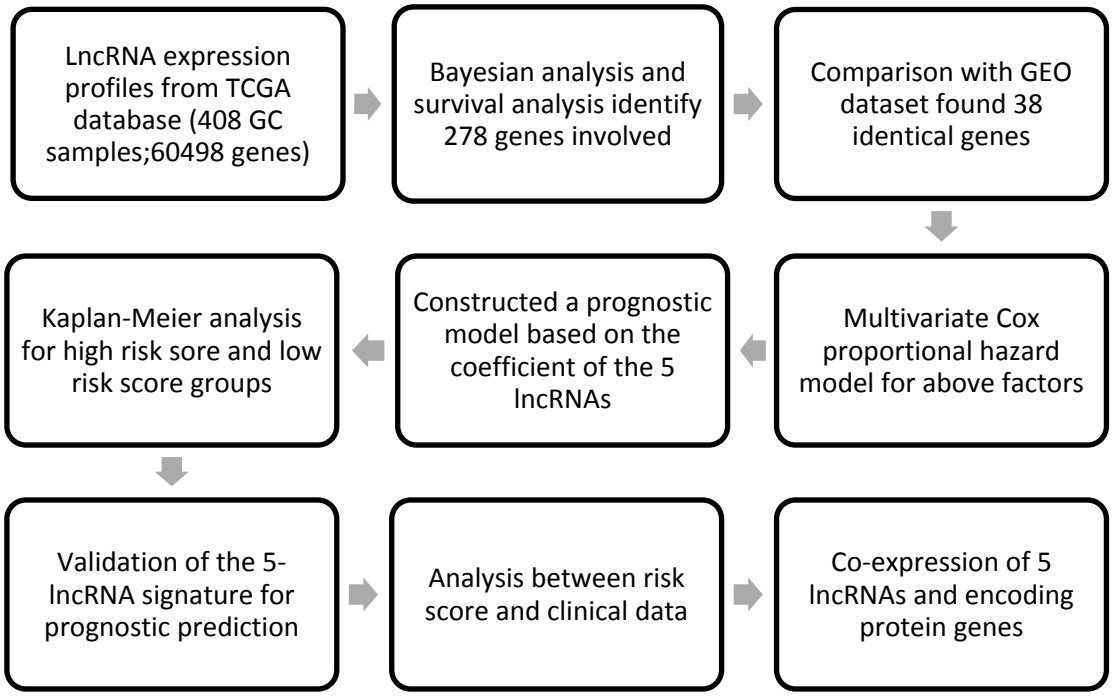
Derived from the multivariable Cox proportional hazards regression analysis in the training group

| Gene name | Ensemble ID | Chr. | Coordinate | Coefficient | Hazard ratio | P value |
|---|---|---|---|---|---|---|
| LINC00205 | ENSG00000223768.1 | 21 | 45288052-45297354 | 0.249092 | 1.373451497 | 0.047216345 |
| TRHDE-AS1 | ENSG00000236333.3 | 12 | 72253507-72273509 | 0.182045 | 1.846654514 | 0.000109193 |
| OVAAL | ENSG00000236719.2 | 1 | 180558974-180566518 | 0.271169 | 1.880897277 | 0.0000744 |
| LINC00106 | ENSG00000236871.6 | X&Y | 1397025-1399412 | -0.207942 | 0.624972486 | 0.003469142 |
| MIR100HG | ENSG00000255248.6 | 11 | 122028329-122422871 | 0.502539 | 1.396343319 | 0.036829012 |

1

## Table 2(on next page)

The schematic workflow of the present study


LncRNA expression profiles from TCGA database (408 GC samples;60498 genes)
Bayesian analysis and survival analysis identify 278 genes involved
Comparison with GEO dataset found 38 identical genes
Multivariate Cox proportional hazard model for above factors
Constructed a prognostic model based on the coefficient of the 5 lncRNAs
Kaplan-Meier analysis for high risk sore and low risk score groups
Validation of the 5-lncRNA signature for prognostic prediction
Analysis between risk score and clinical data
Co-expression of 5 lncRNAs and encoding protein genes

1

## Table 3(on next page)

The prognosis values of five-lncRNA signature in OS of GC patients in clinic subgroup.

Abbreviations: HR, Hazard ratio; 95%CI, 95% confidence interval.

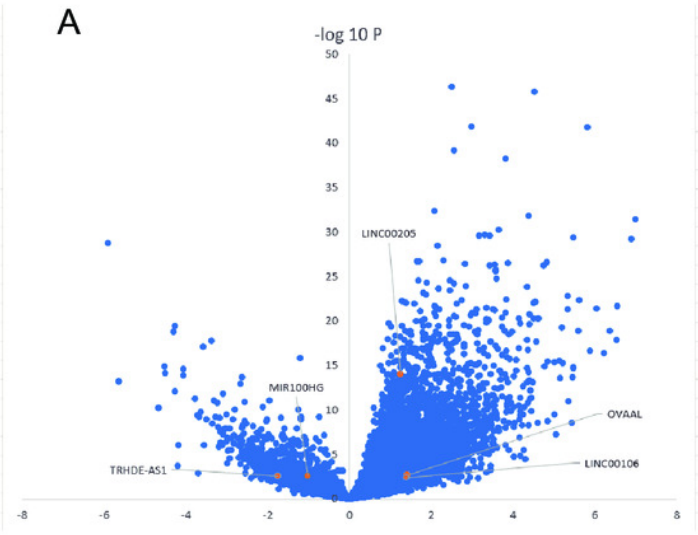
| | Number (High Risk score/Low Risk score) | HR (95%CI) | P value |
|---|---|---|---|
| **Total** | 204/204 | 2.09 (1.80, 2.44) | 0.000001 |
| **Gender** | | | |
| Male | 129/134 | 2.29 (1.53, 3.44) | 0.00002 |
| Female | 75/70 | 1.97 (1.11, 3.47) | 0.01 |
| **Histologic grade** | | | |
| G2 | 47/97 | 2.41 (1.34, 4.33) | 0.0006 |
| G3 | 146/97 | 1.68 (1.13, 2.50) | 0.02 |
| **Race** | | | |
| Asian | 44/41 | 5.47 (1.87, 16.02) | 0.001 |
| Black or african american | 4/8 | 1.78 (0.32, 9.80) | 0.6 |
| White | 138/120 | 2.16 (1.44, 3.24) | 0.0003 |
| **Age** | | | |
| Old | 186/191 | 2.04 (1.46, 2.86) | 0.00001 |
| Young | 18/13 | 5.96 (1.26, 28.17) | 0.008 |
| **TNM stage** | | | |
| Stage I | 14/41 | 2.09 (0.63, 6.93) | 0.3 |
| Stage II | 62/58 | 2.78 (1.34, 5.78) | 0.008 |
| Stage III | 87/77 | 1.68 (1.06, 2.66) | 0.02 |
| Stage IV | 25/16 | 2.04 (0.87, 4.78) | 0.01 |

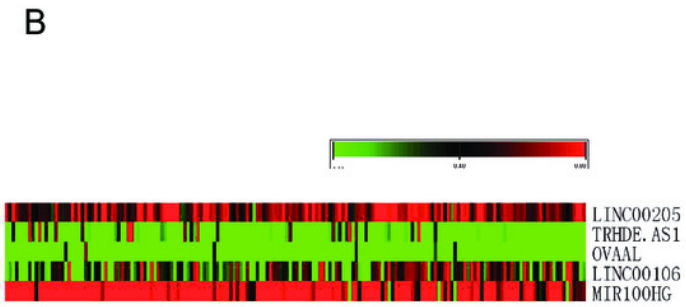
1

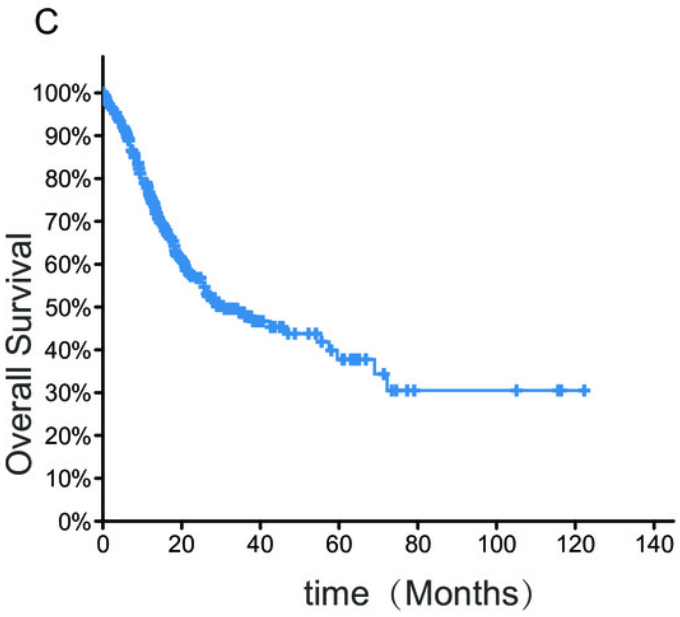
# Figure 1

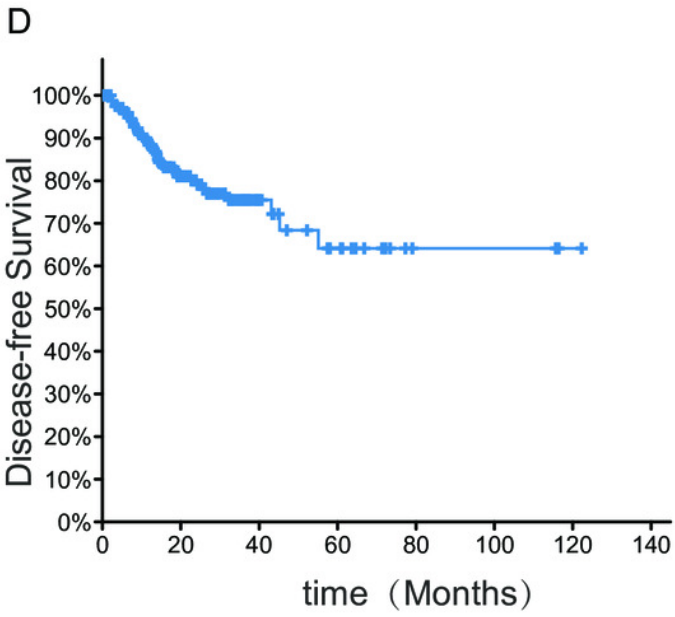
The expression information of five lncRNAs ,overall survival and disease free survival in gastric cancer patients.

(A) Volcano plot with yellow dots indicating five lncRNAs expression levels which is significantly different between tumor and normal tissue based on the criteria of an absolute log2 fold change (FC)>1 and $P < 0.05$.□B□Heatmap of the five-lncRNA expression profiles in the training set.(C-D) Kaplan-Meier analysis of patients' overall survival and disease-free survival in training group.


A
-log 10 P
LINC00205
MIR100HG
OVAAL
TRHDE-AS1
LINC00106


B
LINC00205
TRHDE.AS1
OVAAL
LINC00106
MIR100HG

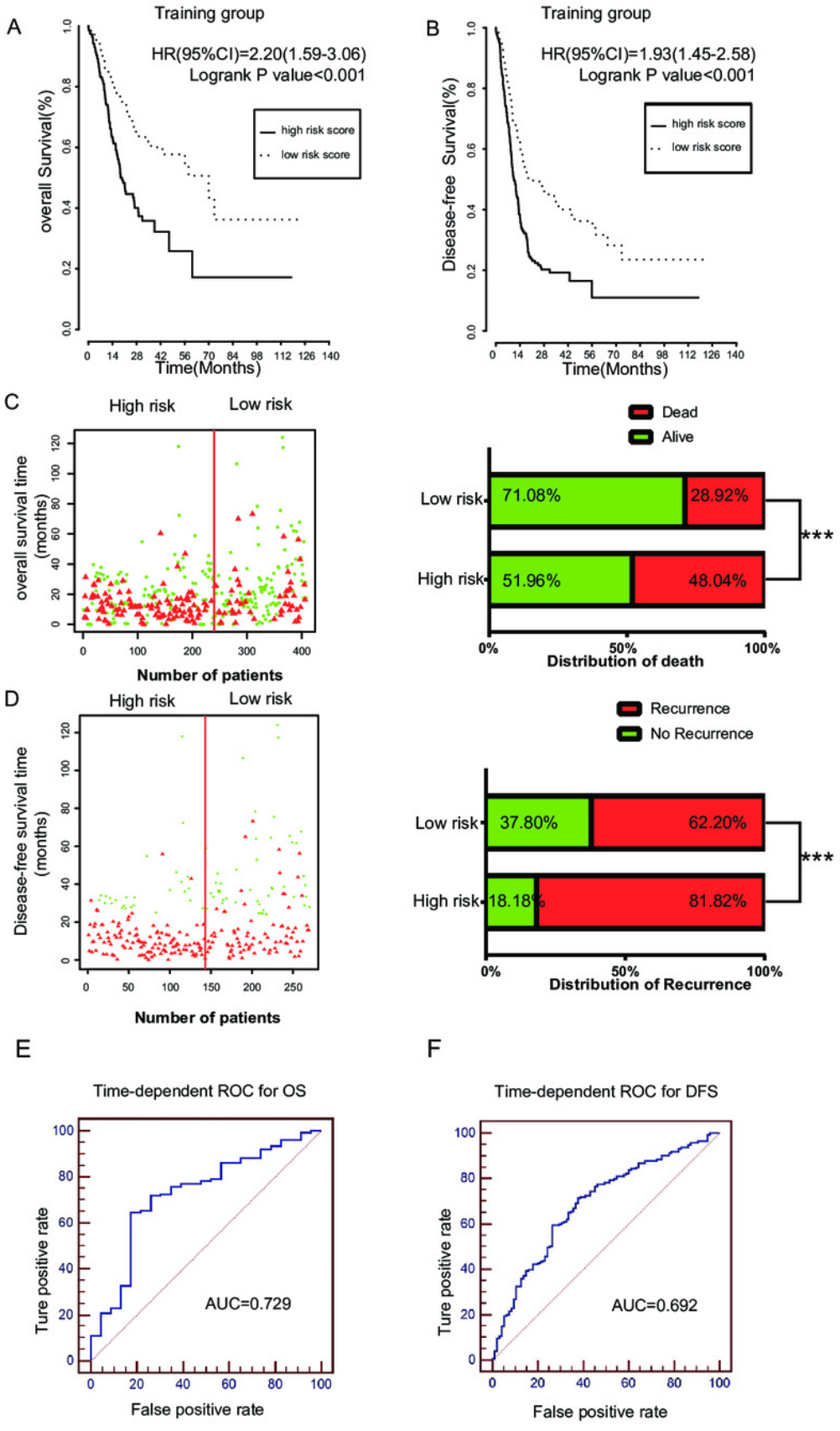

C
Overall Survival
time（Months）


D
Disease-free Survival
time（Months）

# Figure 2

The rognostic value of five-lncRNA signature in training group.

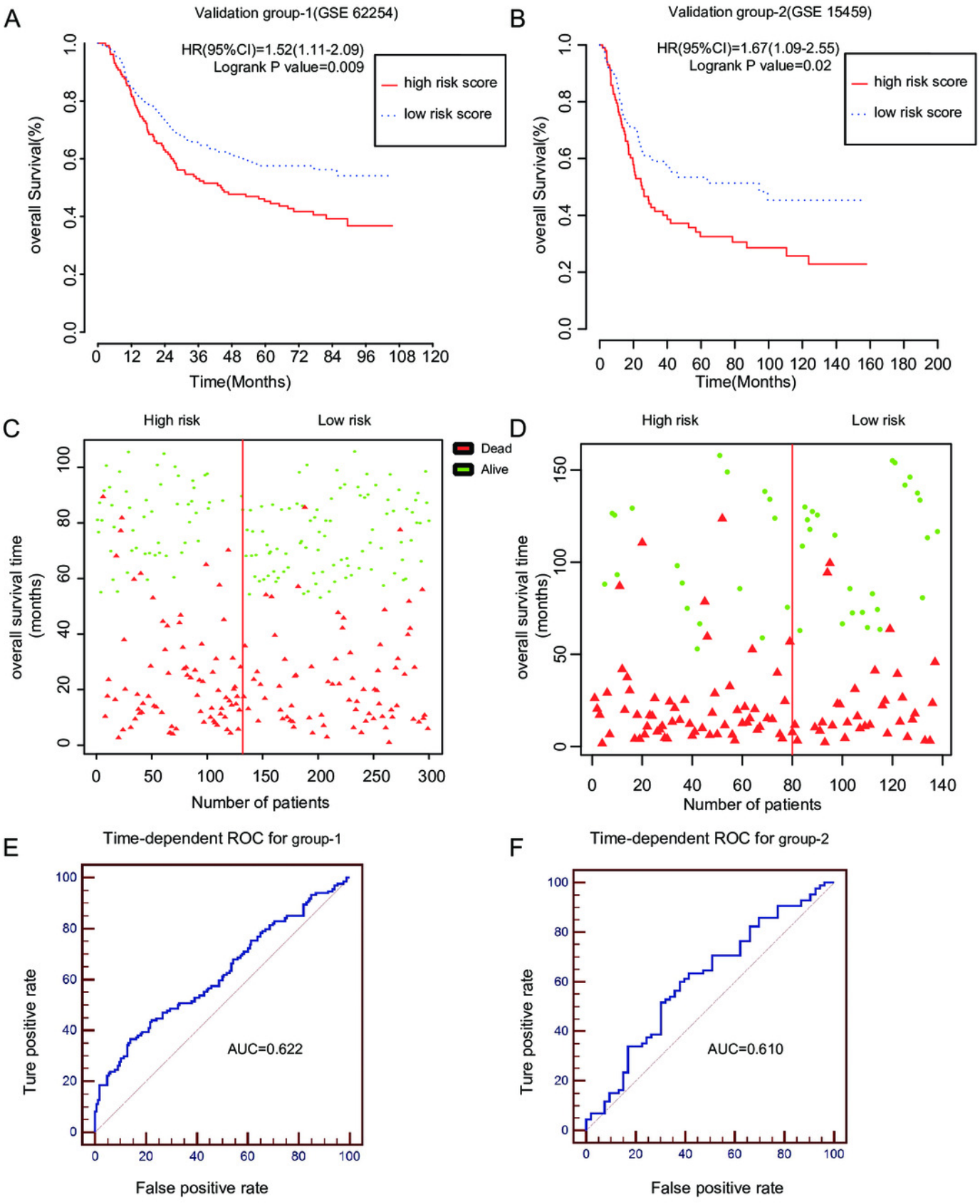
(A-B) Kaplan-Meier analysis of patients' overall survival and disease-free survival in the high-risk (n = 204) and low-risk (n = 204) subgroups of the training set.□C□The scatter plot of five-lncRNA-based risk score distribution for patient survival status (left); the percentage of patient survival status in the high-risk and low-risk subgroups of the training set (right).(D) The five-lncRNA-based risk score distribution for patient recurrence (left); the percentage of patient recurrence in the high-risk and low-risk subgroups of the training set (right). (E-F) The time-independent ROC analysis of the risk score for prediction the OS and DFS of the training set. The area under the curve was calculated for ROC curves. ***P<0.001.


A
Training group
HR(95%CI)=2.20(1.59-3.06)
Logrank P value<0.001
high risk score
low risk score
overall Survival(%)
Time(Months)
B
Training group
HR(95%CI)=1.93(1.45-2.58)
Logrank P value<0.001
high risk score
low risk score
Disease-free Survival(%)
Time(Months)
C
High risk
Low risk
overall survival time (months)
Number of patients
Dead
Alive
Low risk
71.08%
28.92%
High risk
51.96%
48.04%
***
0%
50%
100%
Distribution of death
D
High risk
Low risk
Disease-free survival time (months)
Number of patients
Recurrence
No Recurrence
Low risk
37.80%
62.20%
High risk
18.18%
81.82%
***
0%
50%
100%
Distribution of Recurrence
E
Time-dependent ROC for OS
Ture positive rate
AUC=0.729
False positive rate
F
Time-dependent ROC for DFS
Ture positive rate
AUC=0.692
False positive rate

# Figure 3

The prognostic values of five-lncRNA signature in two independent GEO validation groups.

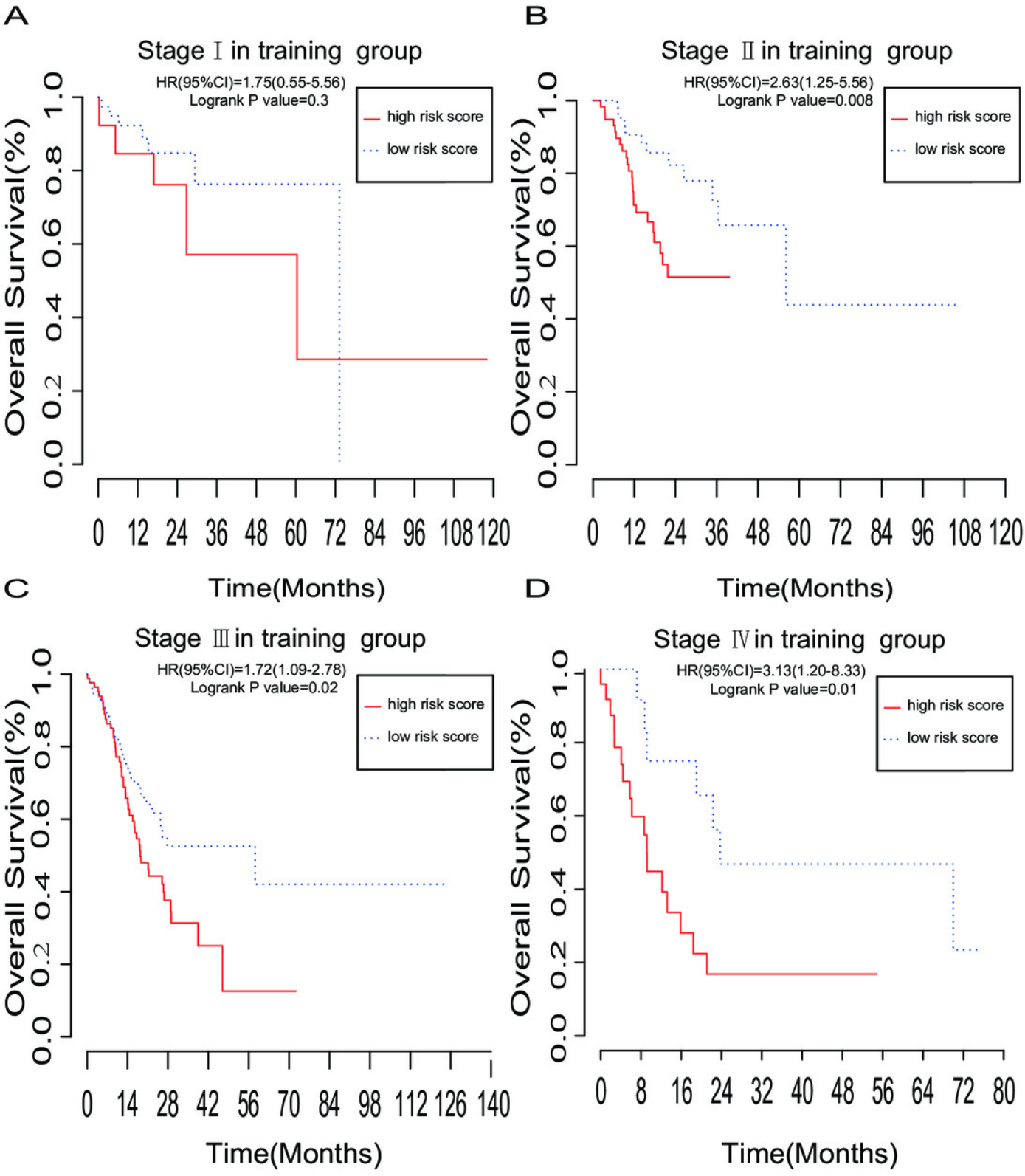
(A-B) Kaplan-Meier analysis of predicting overall survival of GC patients based on the high-risk and low-risk subgroups in two independent validation groups(GSE62254 and GSE15459). C-DThe scatter plot of five-lncRNA-based risk score distribution for patient survival status in two independent validation groups.(E-F) The time-independent ROC analysis of the risk score for prediction the OS of the two independent validation groups. The area under the curve was calculated for ROC curves.


A
Validation group-1(GSE 62254)
HR(95%CI)=1.52(1.11-2.09)
Logrank P value=0.009
high risk score
low risk score
overall Survival(%)
Time(Months)
B
Validation group-2(GSE 15459)
HR(95%CI)=1.67(1.09-2.55)
Logrank P value=0.02
high risk score
low risk score
overall Survival(%)
Time(Months)
C
High risk
Low risk
Dead
Alive
overall survival time (months)
Number of patients
D
High risk
Low risk
overall survival time (months)
Number of patients
E
Time-dependent ROC for group-1
AUC=0.622
Ture positive rate
False positive rate
F
Time-dependent ROC for group-2
AUC=0.610
Ture positive rate
False positive rate

# Figure 4

The prognostic values of five-lncRNA signature in subgroups according to the TNM stage.

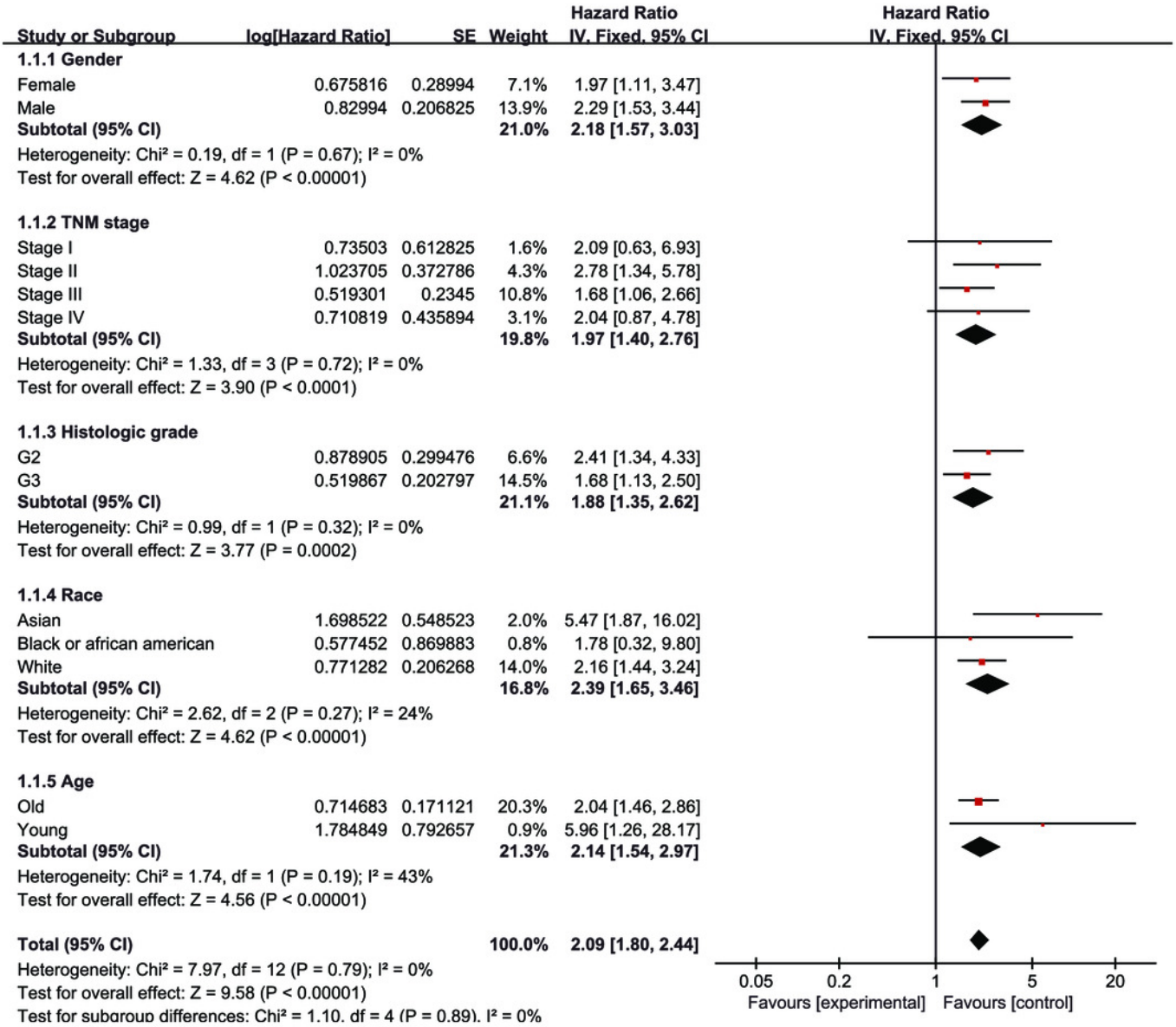
(A-D) Kaplan-Meier analysis of the overall survival of GC patients with stage I,stageⅡ,stageⅢand stageⅣ,respectively.


A
Stage Ⅰ in training group
HR(95%CI)=1.75(0.55-5.56)
Logrank P value=0.3
high risk score
low risk score
Overall Survival(%)
0.0 0.2 0.4 0.6 0.8 1.0
0 12 24 36 48 60 72 84 96 108 120
Time(Months)
B
Stage Ⅱ in training group
HR(95%CI)=2.63(1.25-5.56)
Logrank P value=0.008
high risk score
low risk score
Overall Survival(%)
0.0 0.2 0.4 0.6 0.8 1.0
0 12 24 36 48 60 72 84 96 108 120
Time(Months)
C
Stage Ⅲ in training group
HR(95%CI)=1.72(1.09-2.78)
Logrank P value=0.02
high risk score
low risk score
Overall Survival(%)
0.0 0.2 0.4 0.6 0.8 1.0
0 14 28 42 56 70 84 98 112 126 140
Time(Months)
D
Stage Ⅳ in training group
HR(95%CI)=3.13(1.20-8.33)
Logrank P value=0.01
high risk score
low risk score
Overall Survival(%)
0.0 0.2 0.4 0.6 0.8 1.0
0 8 16 24 32 40 48 56 64 72 80
Time(Months)

# Figure 5

The orest plot to evaluate prognostic values of five-lncRNA signature in subgroups divided by clinic related factors.

# Foresst plot for clinic subgroup

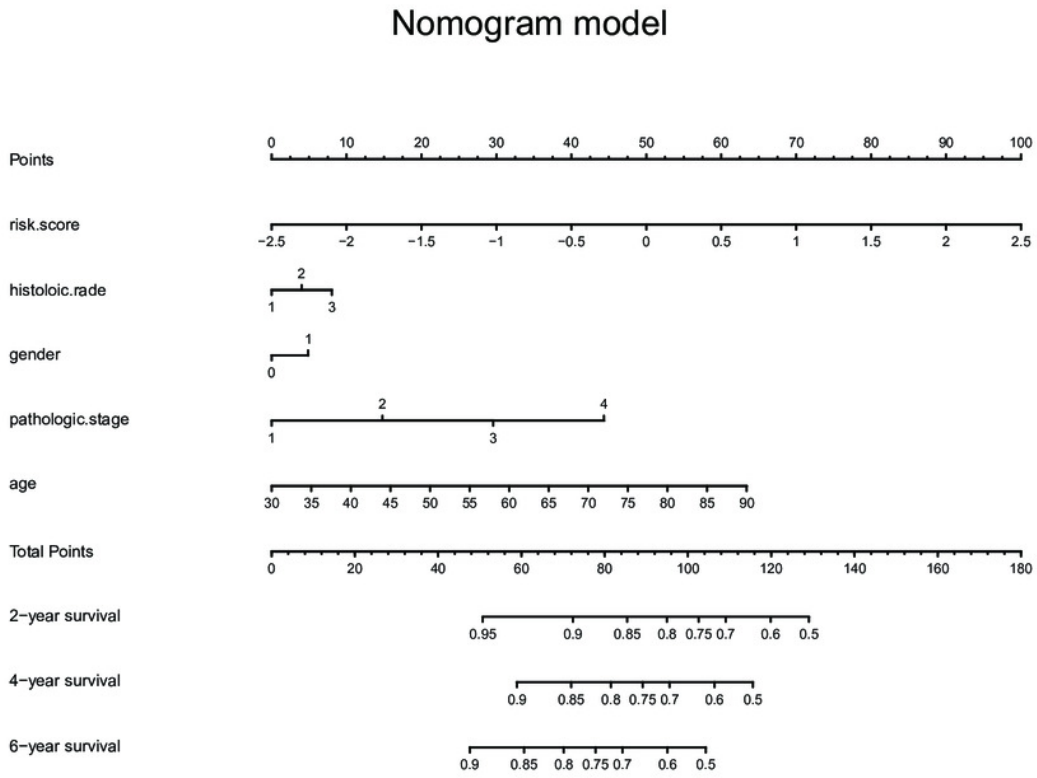


# Figure 6

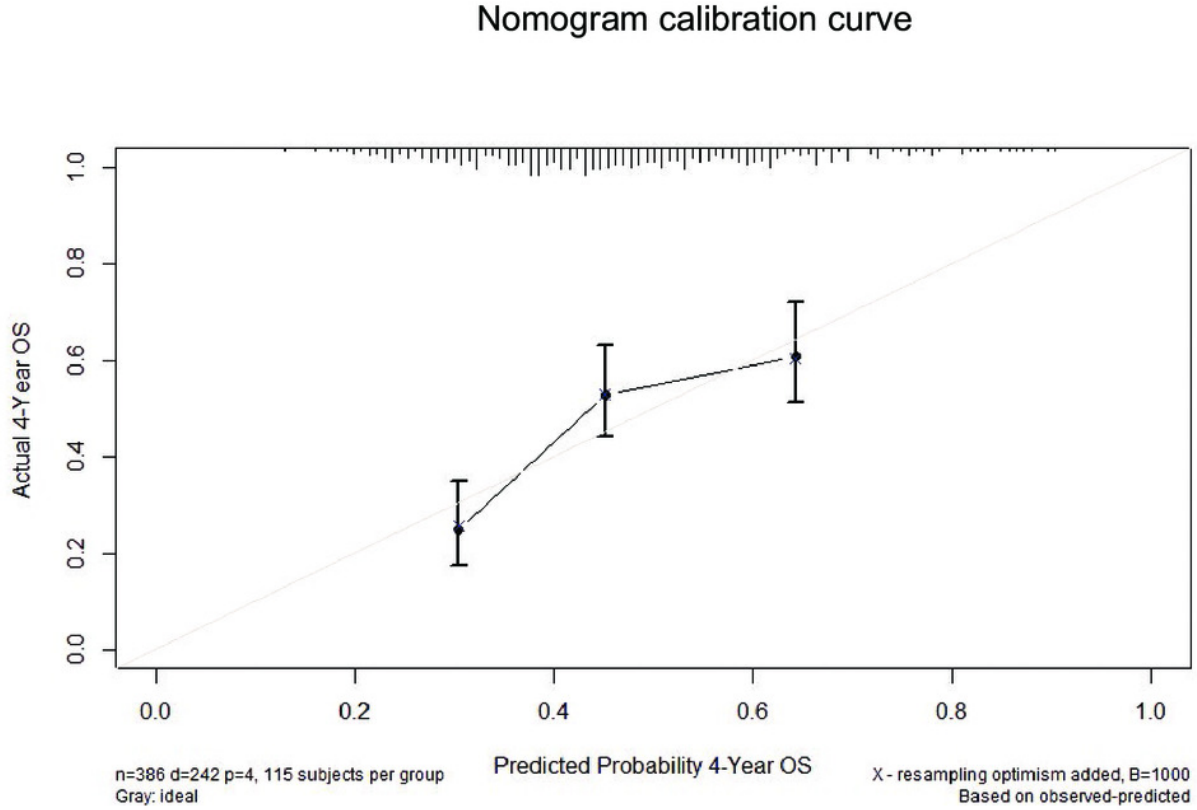
The prognosis value of a nomogram model combining five-lncRNA signature with the clinic related factors.

(A) A nomogram model combining five-lncRNA signature with the clinic related factors for predicting OS of GC patients. (B) The nomogram calibration curve to evaluate the prediction of 4-year OS of GC patients.

A


Nomogram model
Points
0 10 20 30 40 50 60 70 80 90 100
risk.score
−2.5 −2 −1.5 −1 −0.5 0 0.5 1 1.5 2 2.5
histoloic.rade
2
1 3
gender
1
0
pathologic.stage
2 4
1 3
age
30 35 40 45 50 55 60 65 70 75 80 85 90
Total Points
0 20 40 60 80 100 120 140 160 180
2−year survival
0.95 0.9 0.85 0.8 0.75 0.7 0.6 0.5
4−year survival
0.9 0.85 0.8 0.75 0.7 0.6 0.5
6−year survival
0.9 0.85 0.8 0.75 0.7 0.6 0.5

B

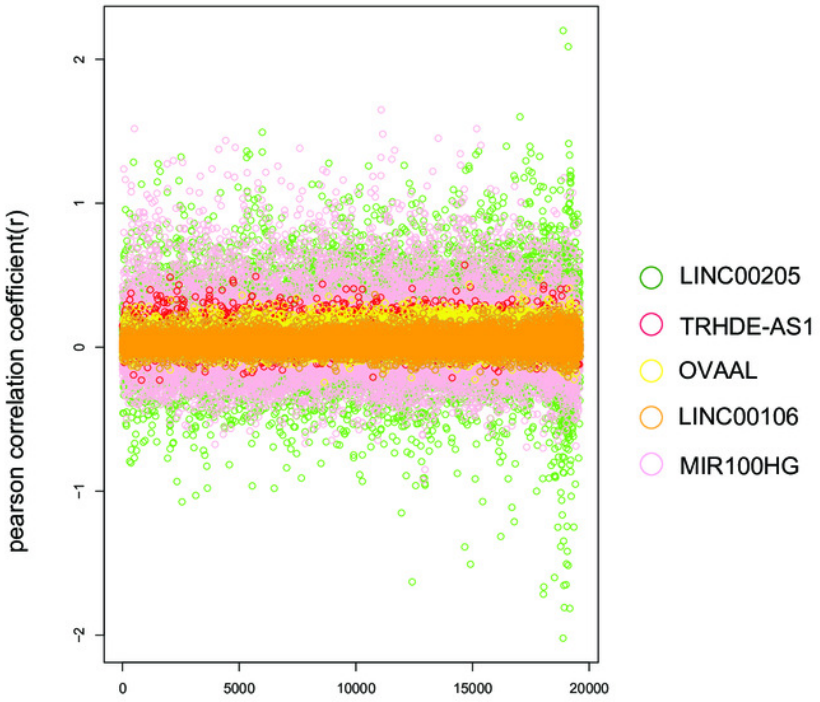

Nomogram calibration curve
Actual 4-Year OS
0.0 0.2 0.4 0.6 0.8 1.0
Predicted Probability 4-Year OS
0.0 0.2 0.4 0.6 0.8 1.0
n=386 d=242 p=4, 115 subjects per group
Gray: ideal
X - resampling optimism added, B=1000
Based on observed-predicted

# Figure 7

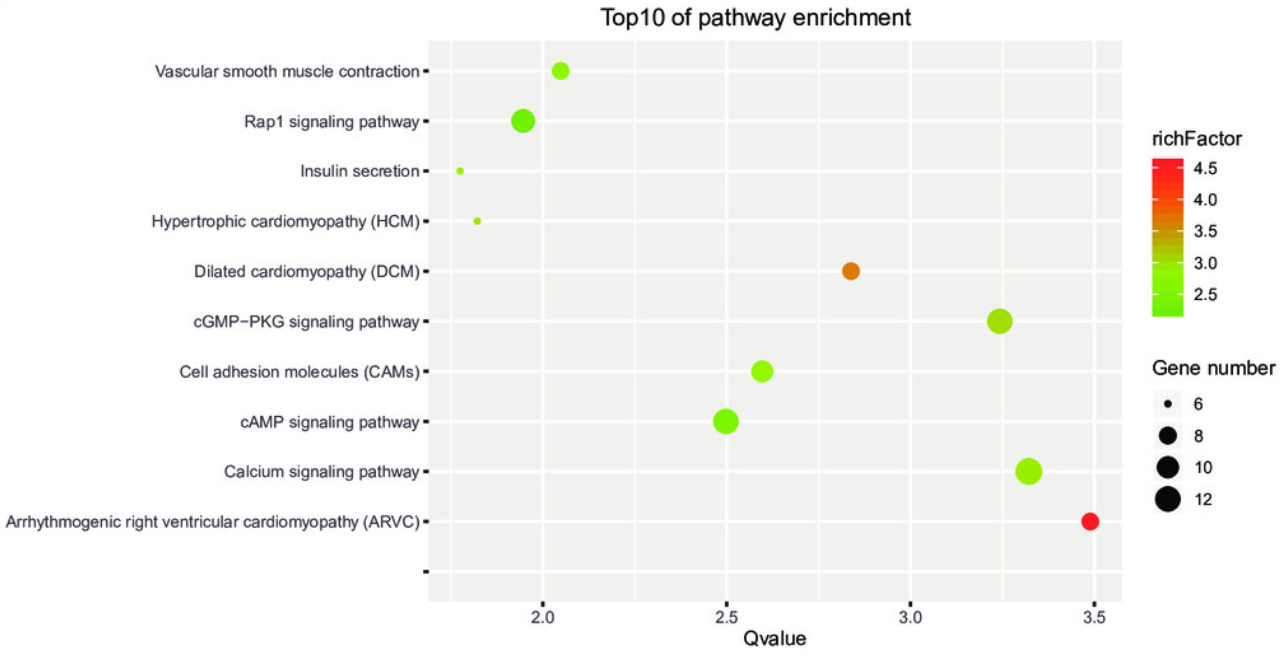
Functional enrichment results of the co-expressed protein-coding genes with five lncRNAs.

(A) the pearson correlation coefficient between 19605 protein-coding genes and five lncRNAs in TCGA database. (B) The functional enrichment bubble map of pathways by KEGG pathway analysis. bubble size represents the number of gene in the pathways.

A


pearson correlation coefficient(r)
2
1
0
-1
-2
0
5000
10000
15000
20000
LINC00205
TRHDE-AS1
OVAAL
LINC00106
MIR100HG

B


Top10 of pathway enrichment
Vascular smooth muscle contraction
Rap1 signaling pathway
Insulin secretion
Hypertrophic cardiomyopathy (HCM)
Dilated cardiomyopathy (DCM)
cGMP−PKG signaling pathway
Cell adhesion molecules (CAMs)
cAMP signaling pathway
Calcium signaling pathway
Arrhythmogenic right ventricular cardiomyopathy (ARVC)
2.0
2.5
3.0
3.5
Qvalue
richFactor
4.5
4.0
3.5
3.0
2.5
Gene number
6
8
10
12