A comprehensive analysis of the lncRNAs and genes for the gastric carcinoma

Wei Cheng
Department of Gastrointestinal Surgery
Minzhe Li
General Surgery Department
Shaofeng Lin
Department of Thoracic Surgery
Jun Xiao (xiaojunabcdef@sina.cn)
Fujian cancer hospital

Research article

Keywords: DIRC1, IQCM, MATN3, SOX14, C5orf46, CYP19A1

Posted Date: August 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-56807/v1

Abstract

Background

Gastric carcinoma (GC) in one of the most common malignant tumors in the worldwide. Despite numerous studies, the molecular mechanism of is still unclear and the prognosis of GC remains poor.

Methods

The present study obtained expression data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) database. The gene risk signature and lncRNA signature were constructed by performing the univariate cox regression analysis and least absolute shrinkage and selection operator (LASSO) analysis. The receive operator curve (ROC) analysis was applied to evaluated the specificity and sensitivity of risk signature. The potential pathway was performed by using the Gene Set Enrichment Analysis (GSEA).

Results

In total, 1641 differentially expressed genes (DEGs) and 985 differentially expressed lncRNAs (DElncRNAs) were obtained among GC samples. A 6 prognostic DEGs (DIRC1, IQCM, MATN3, SOX14, C5orf46 and CYP19A1) classifier and 9 prognostic DElncRNAs (AC007126.1, AC011352.1, AL356417.2, AP000695.1, LINC01210, LINC01614, VCAN-AS1, AC005165.1 and AC011586.2) classifier were identified using lasso-penalized multivariate survival modelling with 10 fold cross-validation. According to median risk score, patients were divided into high risk group and low risk group. The overall survival result for patients have a significant divergence between high rosk group and low risk group ($p < 0.001$). The 6 DEGs signature risk model and 9 DElncRNAs signature risk model was further verified that can served as an independent prognostic biomarkers for GC prediction among the clinical traits ($P < 0.001$). Moreover, two independent cohort GEO dataset (GSE13911 and GSE70800) was employed to evaluate the specificity and sensitivity of the prognostic gene and prognostic lncRNA. Gene set enrichment analysis (GSEA) result for the risk model revealed that these gene involved in ECM receptor interaction pathway, ERBB signaling pathway, UBIQUITIN mediated proteolysis, cell adhesion molecules CAMS, ECM receptor interaction, focal adhesion, pathways in cancer and TGF beta signaling pathway, meaning that the prognostic gene risk model and lncRNAs risk model play an crucial role and have important prognostic values.

Conclusion

These findings may have important significance in understanding the molecular mechanism of GC and potential therapeutic method for the GC patients.

Background

Gastric carcinoma (GC) is one of the most common malignant tumours and the third leading cause of cancer-associated mortality [1, 2]. About 80% of patients with GC are diagnosed at an advanced stage, which makes the patients show a poor prognosis [3]. Surgery is the most suitable curative treatment for gastric cancer by far. But for patients with recurrent or un-resectable GC, there is no satisfactory treatment, and for advanced gastric cancer, the 5-year mortality remains 30% to 50% [4-6]. Therefore, it is necessary to explore the early diagnosis methods of GC, which can help improve the prognosis of patients.

As a type of noncoding functional RNAs, long non-coding RNAs (lncRNAs) were recently attracted to show the mechanism of competing endogenous RNAs (ceRNAs) [7, 8]. Recent studies have also indicated that ceRNAs is involved in the regulation of carcinogenesis [9-11]. Especially for GC, many lncRNAs and mRNAs were reported to be related with prognosis of GC patients [12-14]. Although many non-coding RNAs were found to be associated with the prognosis of gastric cancer and can be used as potential prognostic markers for potential prognostic analysis. However, unilateral analysis of genes or lncRNAs are often not stable because the potential association between lncRNAs and genes. Thus, it is necessary to establish a model combining multiple RNAs molecular markers with prognostic effects.

Here, we comprehensively analysed two types of RNA (lncRNAs and mRNAs) in GC patients through the RNA expression profiling data. According to differential expression of these RNAs, we established six gene signature risk model and nine lncRNAs risk model using lasso-penalized multivariate survival modelling with 10 fold cross-validation analysis. The risk model further evaluated by performing ROC analysis and validated in the two independent dataset. Based on these efforts, we hope to find a variety of more effective clinical prognostic markers for gastric cancer patients.

Methods

RNA sequencing data and clinical information

RNA expression profiling data (including LncRNA, mRNA) and corresponding clinical information of GC patients were all downloaded from The Cancer Genome Atlas (TCGA) dataset (https://cancergenome.nih.gov/). In order to reduce the bias of data, some exclusion criteria were made as follows: (1) absence of prognosis information; (2) histologic diagnosis ruled out GC. Differentially expressed lncRNA and mRNA in GC and adjacent normal tissues were analyzed by using the “edgeR” package in R software. |Fold change| > 2 and adjusted P value < 0.05 were set as the thresholds to select different kind of RNA. Gene expression greater than 1 will be retained for further analysis. The DE lncRNAs and DE genes were then integrated with the clinical information with the patients’ survival information (survival time and survival stat), respectively.

Statistical analysis and validation

The DEGs and DElncRNAs expression profile signature for prediction of prognosis was obtained from the GC tumor samples. Then we filter the expression dataset and retained the DEGs and DElncRNAs with

variance in expression levels bigger than 5, and the P values calculated from univariate cox regression analysis lower than 0.05. These DEGs and DElncRNAs were then subjected to penalized multivariate cox regression survival modelling using an LASSO estimation algorithm [15, 16].With this model, the prognostic genes were selected by 10 fold cross-validation in the dataset.

According to the cox regression coefficient, a risk score formula was established and the risk score were then calculated for each patient samples. According to the median risk score, patients were further categorized into high risk group and low risk group respectively. The prognostic lncRNAs and mRNAs were subsequently evaluated by using ROC analysis and validated in the two independent dataset (GSE13911 and GSE70800). Kaplan-Meier and Log-Rank methods were used to test difference between two groups and P-value < 0.05 was taken as statistically. All statistical analyses and graphics were conducted with R software (version 3.5.1).

Gene set enrichment analysis (GSEA)

According to the median expression levels of each prognostic genes, patients were categorized into high expression level group and low expression level group. Then, Gene set enrichment analysis were performed for the prognostic genes between high expression level group and low expression level group. The pathway was considered statistically significant if its p values were less than 0.05

Results

Differentially expressed analysis and clinical factor screening.

According to the exclusion criteria, 314 patients with the expression data of mRNA and lncRNA were enrolled in this study, respectively. A total of 1641 (including 872 up-regulated genes and 769 down-regulated genes) differentially expressed genes (DEGs) and 985 (including 775 up-regulated and 210 down-regulated) differentially expressed lncRNAs (DElncRNAs) were identified between tumour tissue and normal tissue with a criterion: adjusted p value < 0.05 and |Fold change| >2. As showed in the Fig. 1, the expression of lncRNAs and genes were presenting by heatmap (Fig. 1A and 1C), and the volcano plots of the distribution of DEGs and DElncRNAs in each data set are shown in Fig. 1B and 1D.

Survival analysis for the DElncRNAs and DEGs

We excluded the DEGs and DElncRNAs with variance in expression levels lower than 5 and the rest DEGs and DElncRNAs were then integrated the corresponding clinical information respectively. By performing univariate cox regression analysis for the DE lncRNAs and DEGs dataset, we selected the significantly DEGs and DElncRNAs with p value smaller than 0.05. Then, Lasso-penalized multivariate Cox proportional hazards modeling was performing on the filter DElncRNA (N = 75) and DEGs (N = 128) dataset respectively. After 100 iterations, 6 DEGs expression signatures and 9 DElncRNAs expression signatures with optimal survival predictions in the GC cohort more than 50 times each. Based on the

regression coefficient of the DEGs and DElncRNAs, the risk formula was constructed for the 6 DEGs and 9 DElncRNAs, and the risk score for each patient was then calculated according to the following formula:

gene risk model = DIRC1 * 0.0086 + IQCM * 0.01921 + MATN3 * 0.0002 + SOX14 * 0.0022 + C5orf46 * 0.0022 + CYP19A1 * 0.0012

lncRNA risk model = AC007126.1 * 0.051 + AC011352.1 * 0.011 + AL356417.2 * 0.0037 + AP000695.1 * 0.0045 + LINC01210 * -0.0086 + LINC01614 * 0.0012 + VCAN-AS1 * 0.026 + AC005165.1 * 0.0006 + AC011586.2 * 0.023

According to median risk score, patients were classified into high risk group and low risk group respectively. Kaplan-Meier curves analysis result showed that the there is a significantly difference between low risk score group and high risk score group in the lncRNA risk model and gene risk model (p value < 0.001) (Fig. 2A and Fig. 2B). Moreover, patients with high risk score tend to have shorted survival time and more deceased cases, whereas patients have prolonged survival time were inclined to low risk group (Fig. 3A and 3B). To further evaluated the accuracy of the gene and lncRNA risk model, ROC analysis was conducted on these risk model and the five years AUC value were 0.851and 0.884, respectively, showed a good accuracy (Fig. 4A and 4B). We also compared the accuracy of the other clinical trait including age, gender, grade, stage and TNM with our risk model. As showed in the Fig. 5A and 5B, our risk model is superior to the other clinical trait in accuracy. In addition, we performed a univariate and multivariate cox regression analysis for the risk model and clinical information, as showed in the Fig. 6A to 6D, the P value of the lncRNA risk model and gene risk model were significantly among the clinical trait, suggesting that these risk models can served as an independent predictor. We subsequently investigate the correlation between prognostic genes and lncRNAs, as the Fig. 7 showed, C5orf46 was highly correlated with AC011352.1, SOX14 was highly correlated with LINC01210.

Validation of the risk model

To further investigated the sensitivity and specificity of our risk model, we downloaded two GEO dataset including GSE13911 and GSE70800 to validate it. The GSE13911 dataset including 31 normal samples and 38 tumor sample, and including five prognostic genes (C5orf46, CYP19A1, DIRC1, MATN3 and SOX14). ROC analysis result showed that the CYP19A1 have a highest AUC value (0.735) and the AUC value of the DIRC1 is the lowest (0.466) (Fig. 8A). The lncRNA expression dataset GSE70800 were re-annotated based on the latest annotation revision. In the dataset, six prognostic lncRNA were identified and ROC analysis result showed that the AUC range from 0.558 to 0.841 (Fig. 8B).

Gene Enrichment analysis for the prognostic genes

In order to explore the potential function of the prognostic gene and lncRNA risk model, we performed GSEA enrichment analysis for the risk group that obtained according to the median risk score. In the gene risk model, as showed in the Fig. 9A, ECM receptor interaction pathways was significantly enriched in the high risk group, while ERBB signaling pathway and UBIQUITIN mediated proteolysis were significantly

enriched in low risk group in the risk gene model. In addition, In the lncRNA risk model, cell adhesion molecules CAMS, ECM receptor interaction, focal adhesion, pathways in cancer and TGF beta signaling pathway were enriched in the high risk group (Fig. 9B). All the result suggesting that the risk model for gene and lncRNA have play important role in the Pathogenesis of GC.

Discussion

A growing number of researches suggests that many complex diseases, especially cancer, can rarely be attributed to a single genetic mutation [17–18]. Many studies had reported that lncRNAs and genes were related with prognosis of GC. Cao et al identified a set of lncRNAs differentially expressed in gastric cancer, providing useful information for discovery of new biomarkers and therapeutic targets in gastric cancer [19]. Lin et al found 10 differentially expressed lncRNAs potentially regulating the p53 signalling pathway from large scale expression profiling of lncRNA and mRNA [20].

GC is a painful experience for patients due to its poor prognosis [21]. Therefore, identification of effective prognostic biomarkers and exploration of potential regulatory networks are indispensable steps in the development of effective treatments.

In the present study, in order to gain insights into the molecular events relevant to GC prognosis, we took advantage of the molecular resolution provided by TCGA database and downloaded the expression profile of lncRNA and genes firstly. By performing cox regression analysis and LASSO analysis, we then identified six prognostic genes including C5orf46, CYP19A1, DIRC1, MATN3, SOX14 and IQCM, and nine prognostic lncRNAs namely AC007126.1, AC011352.1, AL356417.2, AP000695.1, LINC01210, LINC01614, VCAN-AS1, AC005165.1 and AC011586.2 that associated with the survival of GC. According to their coefficient derived from the LASSO analysis, we constructed six gene signature risk score model and nine lncRNA signature risk score model, respectively. We further categorized patients into high risk score group and low risk group based on the median risk score. Patients with a high risk score tend to have a significantly shorter survival time, corresponding with more death cases of GC. ROC analysis was used to evaluate the accuracy of the risk model for the gene and lncRNA, separately. The high AUC value for the 5 year (gene = 0.851, lncRNA = 0.884) revealed that the risk model are reliable for the prognosis of the GC. Further cox regression analysis between risk model score and clinical trait suggesting that the gene risk model and lncRNA risk model can act as an independent predictor for the GC survival. In addition, we also compared the accuracy between risk model and clinical trait and the result showed that the AUC value of the risk model was better than clinical trait, indicated that our risk model was more accuracy than clinical trait. In order to validate the prognostic genes and lncRNAs, we downloaded a gene expression dataset (GSE13911) and lncRNA expression dataset from GEO database (GSE70800). High confidence AUC value demonstrate that the genes and lncRNAs have important prognostic value. GSEA enrichment analysis result revealed that the gene risk model and lncRNA risk model have important function in the molecular pathogenesis and progression of GC.

However, there were several limitations towards our work. First, though the comparison between our results and published articles has suggested the validity of our result, it was still a limitation that we did not proved an external validation for our results. Second, several novel lncRNAs, with significant clinical significance in GC need to be explored further to determine the underlying molecular mechanism. Finally, despite the limited power of detecting individual events, the model that we proposed has promising implications in clinical practice.

Conclusion

In a word, our study delineates two prognosis models based on lncRNAs and mRNAs that may improve the poor prognosis of GC. This finding provided some new potential prognostic markers, and identified novel therapeutic targets for GC.

Declarations

Data Availability

The dataset performed in this study are available from the corresponding author on reasonable requests.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Funding Statment

This study was supported by the Fujian Provincial Health Technology Project (2017-ZQN-18, 2019J01200 and 2019J05139) and Science and Technology Program of Fujian Provincial (2018Y2003 and 2017Y0020).

Acknowledgements

We appreciated TCGA database for providing the original study data. This study was supported by the Fujian Provincial Health Technology Project (2017-ZQN-18, 2019J01200 and 2019J05139) and Science and Technology Program of Fujian Provincial (2018Y2003 and 2017Y0020).

Authors' contributions

Jun Xiao designed the study, Wei Cheng, Minzhe Li and Shaofeng Lin collected the clinical information and lncRNA expression data. Wei Cheng analysis data and wrote the manuscript, Jun Xiao revised and offered advice about the manuscripts.

References

[1] Van Cutsem E, Sagaert X, Topal B, Haustermans K and Prenen H: Gastric cancer. Lancet 388: 2654-2664, 2016.

[2] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J and Jemal A: Global cancer statistics, 2012. CA Cancer J Clin 65: 87-108, 2015.

[3] Zong L, Abe M, Seto Y and Ji J: The challenge of screening for early gastric cancer in China. Lancet 388: 2606, 2016.

[4] Cidon EU, Ellis SG, Inam Y, Adeleke S, Zarif S, Geldart T. Molecular targeted agents for gastric cancer: a step forward towards personalized therapy. Cancers (Basel). 2013;5(1):64–91.

[5] Song Z, Wu Y, Yang J, Yang D, Fang X. Progress in the treatment of advanced gastric cancer. Tumour Biol. 2017;39(7):1010428317714626.

[6] Hao N B, He Y F, Li X Q, et al. The role of miRNA and lncRNA in gastric cancer[J]. Oncotarget, 2017, 8(46):81572-81582.

[7] Zhang M, Du X. Noncoding RNAs in gastric cancer: Research progress and prospects:[J]. World Journal of Gastroenterology, 2016, 22(29):6610.

[8] Zhixia Z, Zhijuan L, Yuqi H , et al. The Long Noncoding RNA D63785 Regulates Chemotherapy Sensitivity in Human Gastric Cancer by Targeting miR-422a[J]. Molecular Therapy - Nucleic Acids, 2018, 12:405-419.

[9] Xian-Zi Y, Tian-Tian C, Qing-Jun H, et al. LINC01133 as ceRNA inhibits gastric cancer progression by sponging miR-106a-3p to regulate APC expression and the Wnt/β-catenin pathway[J]. Molecular Cancer, 2018, 17(1):126.

[10] Hai-Feng Z, Hai-Yan Q, Fa-Bo F. Long Noncoding RNA LINC01133 Sponges miR-422a to Aggravate the Tumorigenesis of Human Osteosarcoma[J]. Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics, 2017.

[11] Zhang X F, Ye Y, Zhao S J . LncRNA Gas5 acts as a ceRNA to regulate PTEN expression by sponging miR-222-3p in papillary thyroid carcinoma[J]. Oncotarget, 2018, 9(3).

[12] Li C Y, Liang G Y, Yao W Z, et al. Integrated analysis of long non-coding RNA competing interactions reveals the potential role in progression of human gastric cancer[J]. International Journal of Oncology, 2016.

[13] Lin X C, Zhu Y, Chen W B, et al. Integrated analysis of long non-coding RNAs and mRNA expression profiles reveals the potential role of lncRNAs in gastric cancer pathogenesis[J]. International Journal of Oncology, 2014, 45(2):619-628.

[14] Li F, Huang C, Li Q, et al. Construction and Comprehensive Analysis for Dysregulated Long Non-Coding RNA (lncRNA)-Associated Competing Endogenous RNA (ceRNA) Network in Gastric Cancer[J]. Medical Science Monitor, 2018, 24:37-49.

[15] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building.[J]. Statistics in Medicine, 2010, 26(30):5512-5528.

[16] GSVA: gene set variation analysis for microarray and RNA-Seq data

[17] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research, 43(7), e47. doi: 10.1093/nar/gkv007.

[18] Zhang X, Zhang W, Jiang Y , et al. Identification of functional lncRNAs in gastric cancer by integrative analysis of GEO and TCGA data[J]. Journal of Cellular Biochemistry, 2019(2).

[19] Nibbe RK, Chowdhury SA, Koyuturk M, Ewing R, Chance MR: Protein-protein interaction networks and subnetworks in the biology of disease. Wiley Interdiscip Rev Syst Biol Med 2011;3:357-367.

[20] Cao W J. Analysis of long non-coding RNA expression profiles in gastric cancer[J]. World Journal of Gastroenterology, 2013, 19(23).

[21] Lin X C, Zhu Y, Chen W B, et al. Integrated analysis of long non-coding RNAs and mRNA expression profiles reveals the potential role of lncRNAs in gastric cancer pathogenesis[J]. International Journal of Oncology, 2014, 45(2):619-–628.

Figures

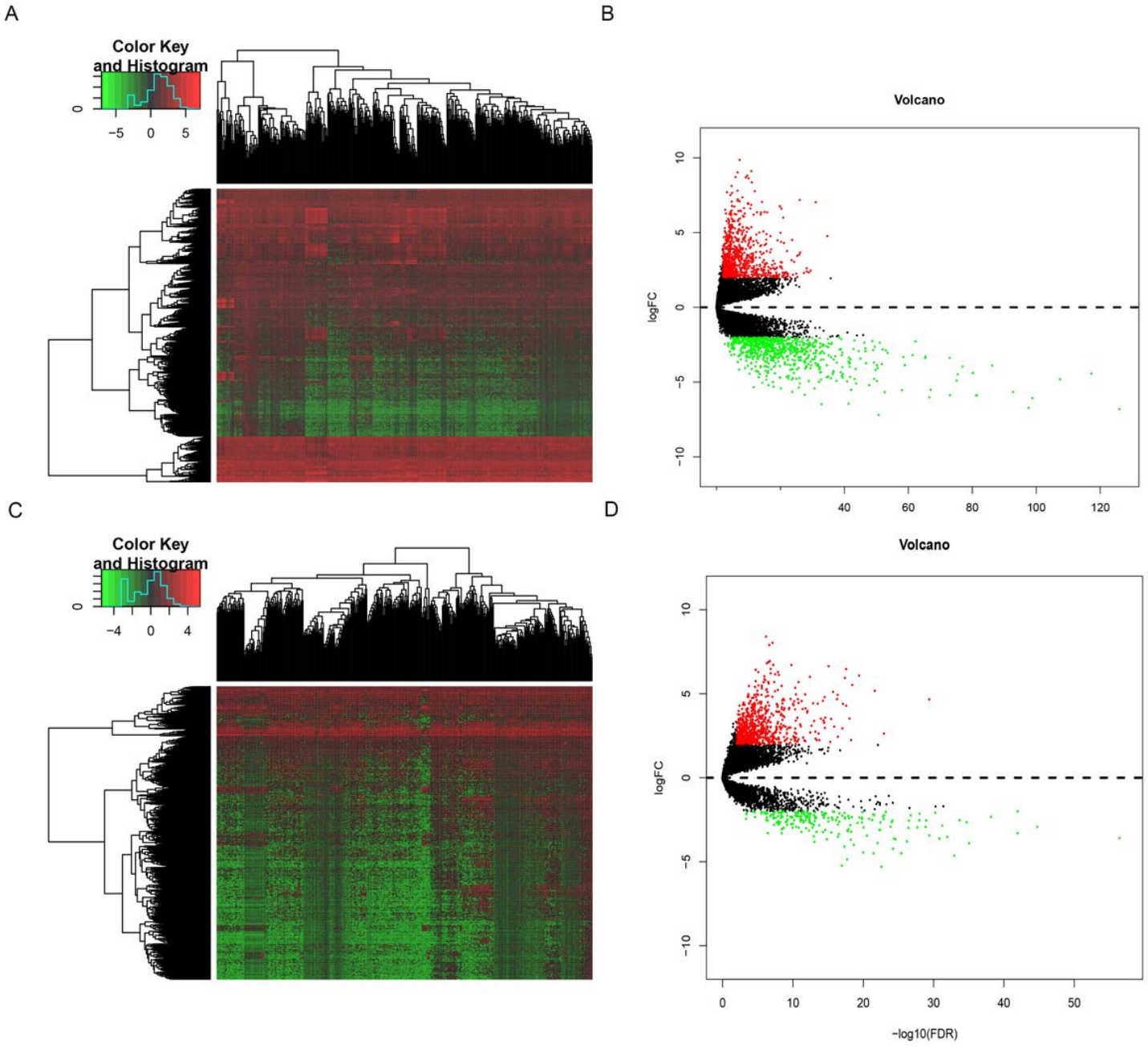

Figure 1

Volcano plot of differentially expressed lncRNAs (a) and DEGs (c). The red point in the plot represents up-regulated RNAs and blue point represents down-regulated RNAs with statistical significance; Heatmap of differentially expressed DElncRNAs (b) and DEGs (d) between GC and non-tumour tissues.

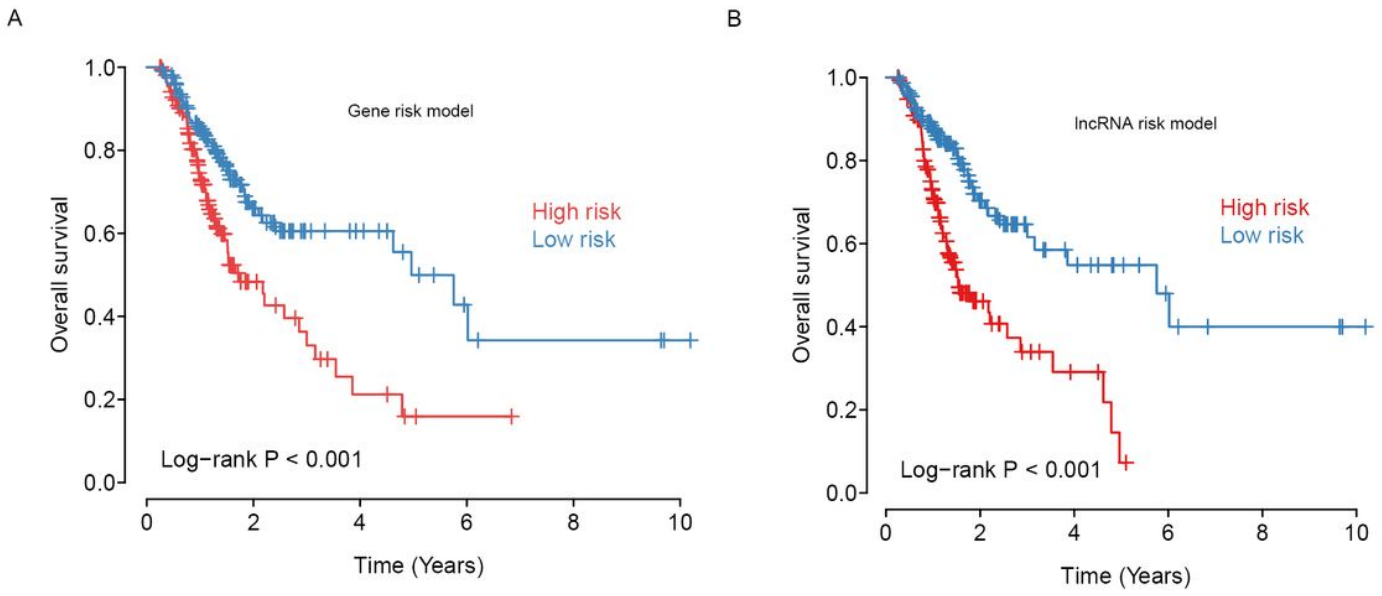

Figure 2

K-M curves analysis for the gene risk model (A) and lncRNA risk model (B). The p value was calculated using log-rank test analysis.

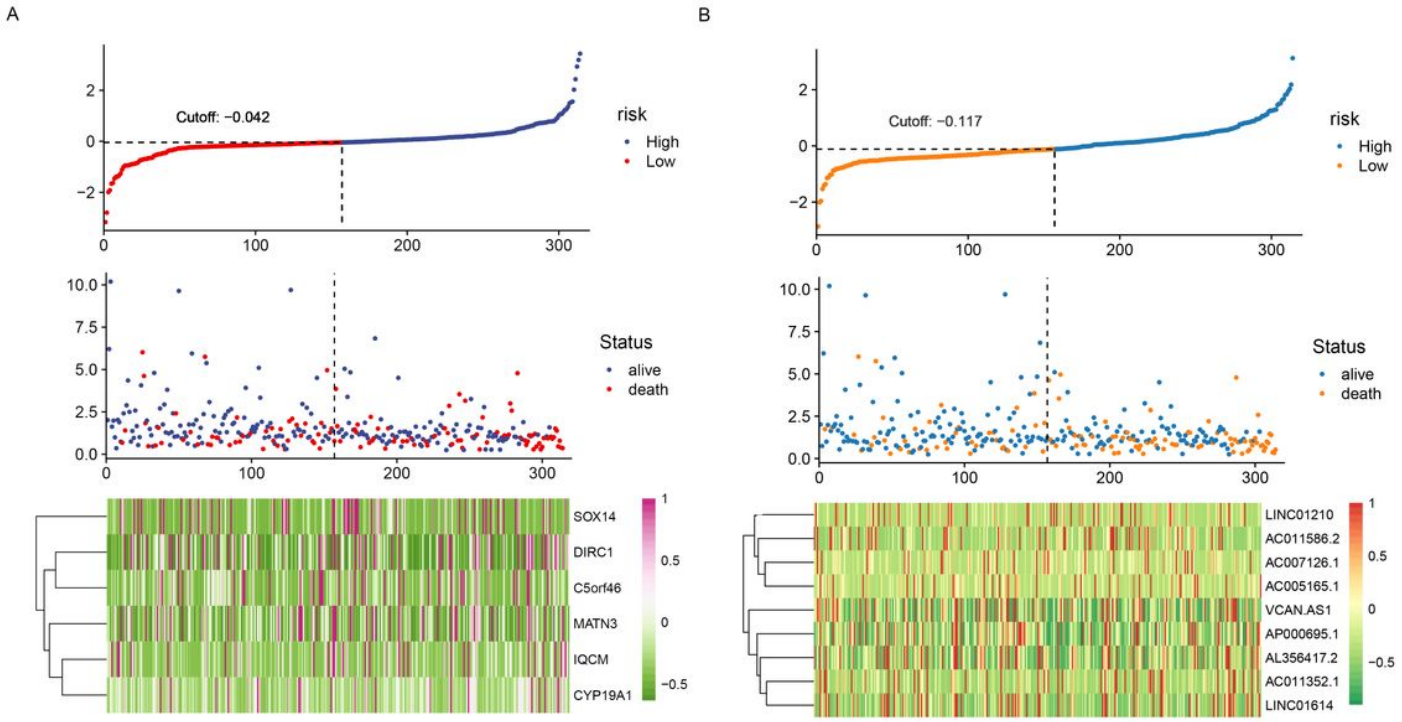

Figure 3

Risk plot for the GC patients samples. (A) Gene risk model dataset (B) lncRNA risk model dataset. Each panel consists of three rows: top rows showed a risk score distribution for the high risk score group and

low risk score group; middle rows represent the GC patients distribution and survival status; the bottom rows showed that the heatmap of prognostic gene or lncRNA expression.

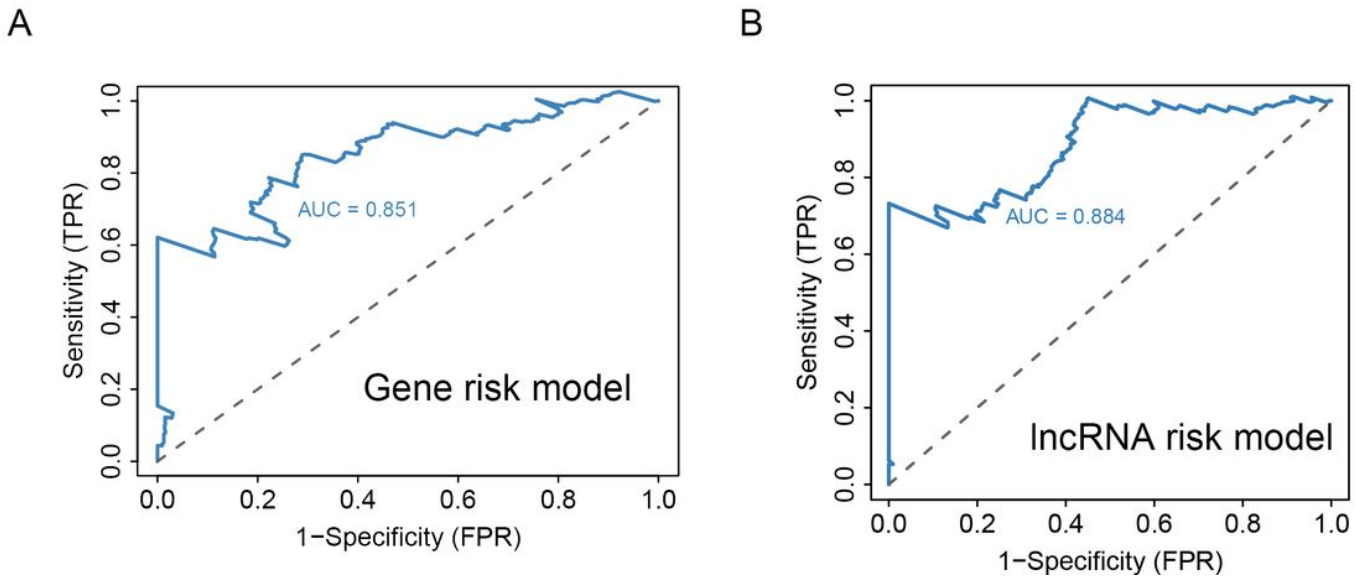

Figure 4

The ROC analysis of the five years for the gene risk model (A) and lncRNA risk model (B).

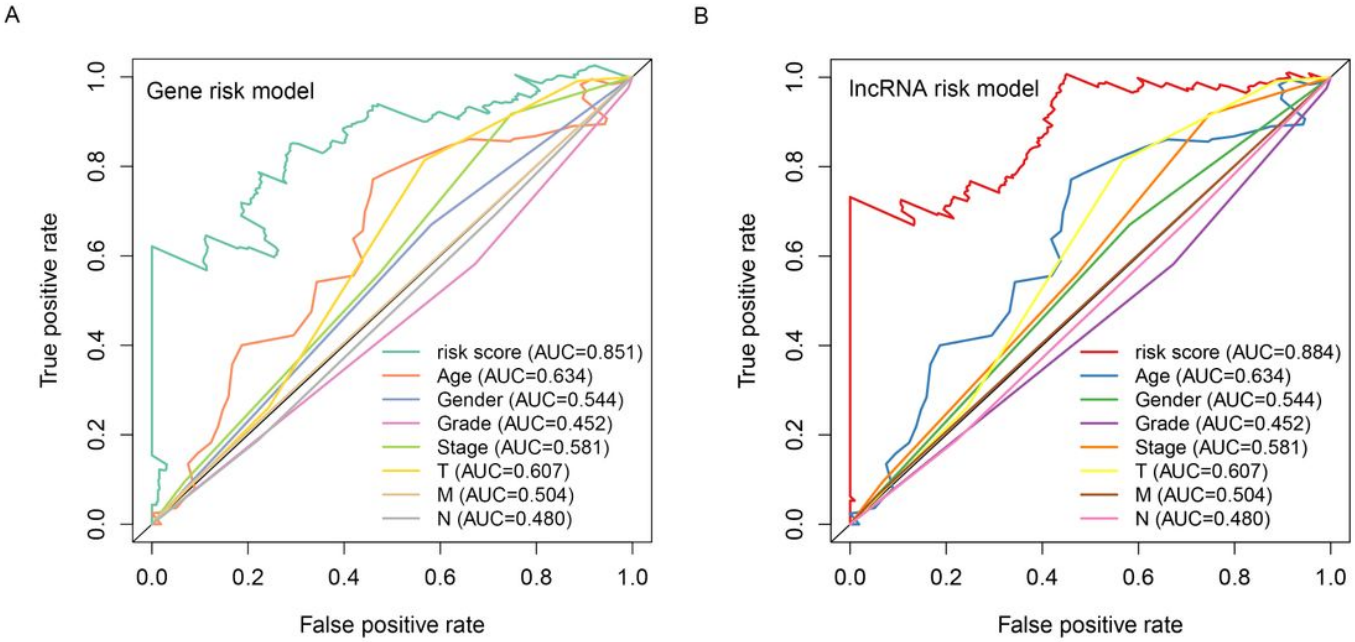

Figure 5

A comparison between risk model score and clinical trait in gene (A) and lncRNA (B).

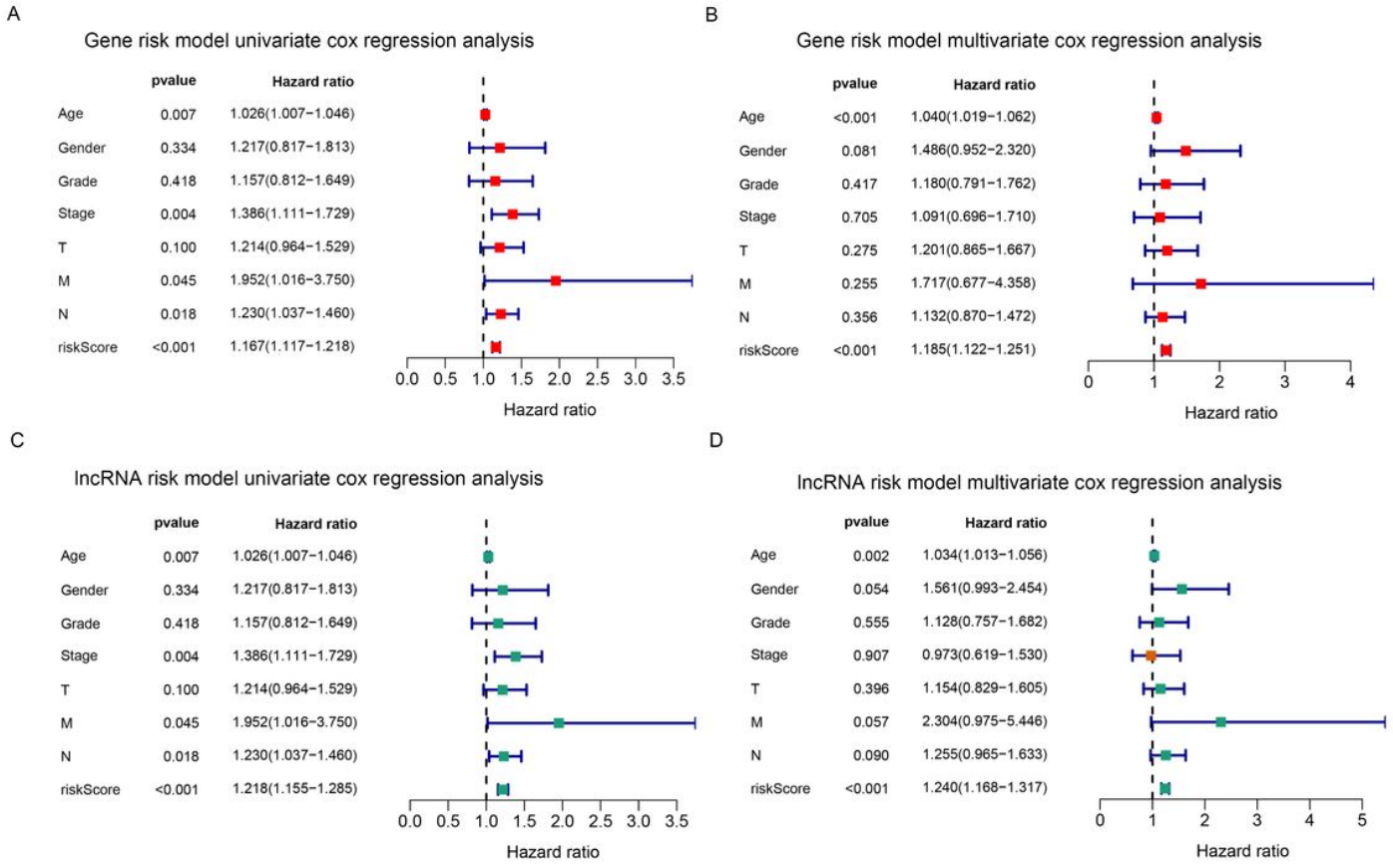

Figure 6

Univariate cox regression analysis and multivariate cox regression analysis for the gene (A and B) or lncRNA risk model (C and D) and clinical trait.

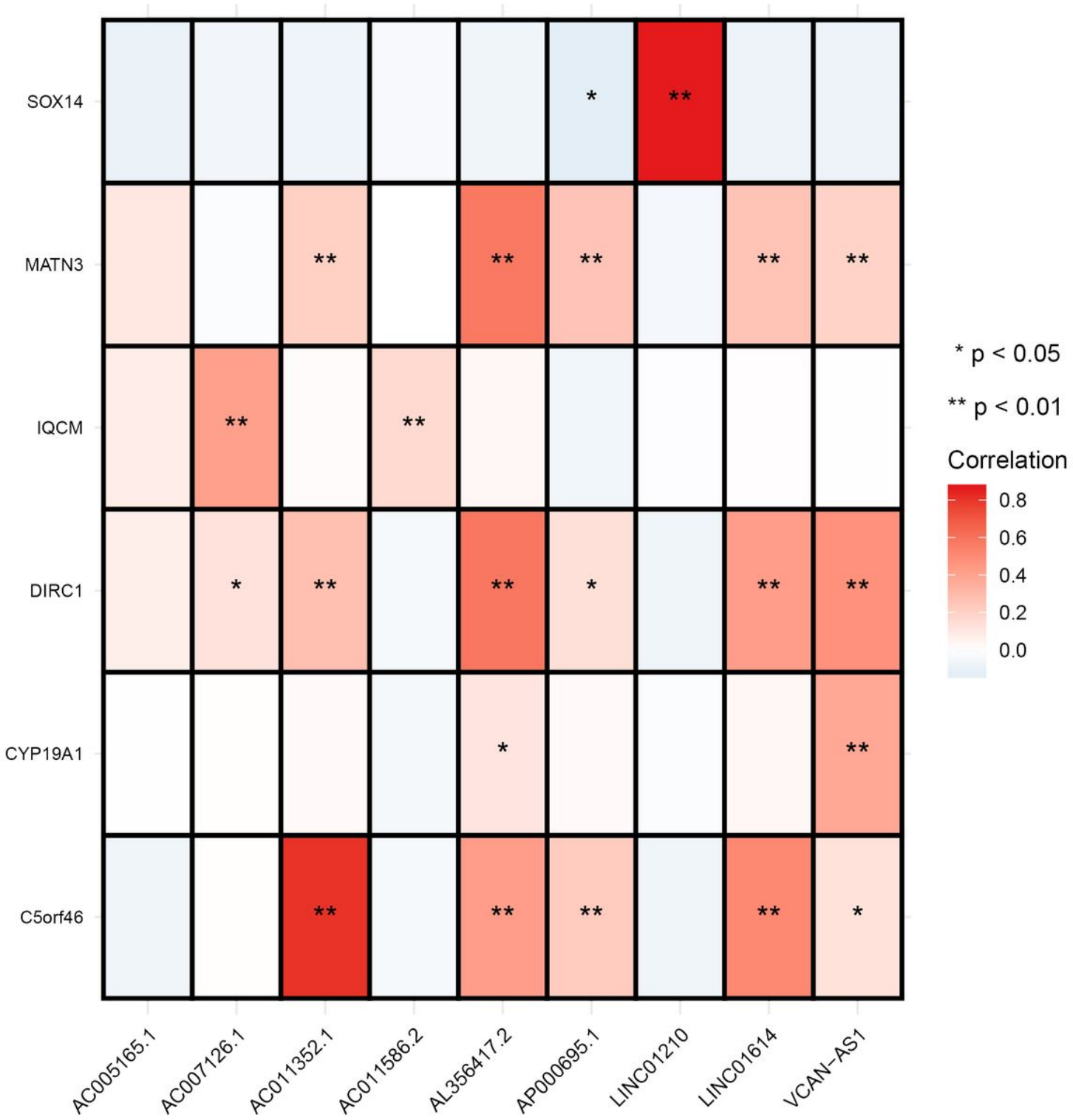

Figure 7

Correlation analysis was performed between prognostic genes and lncRNAs

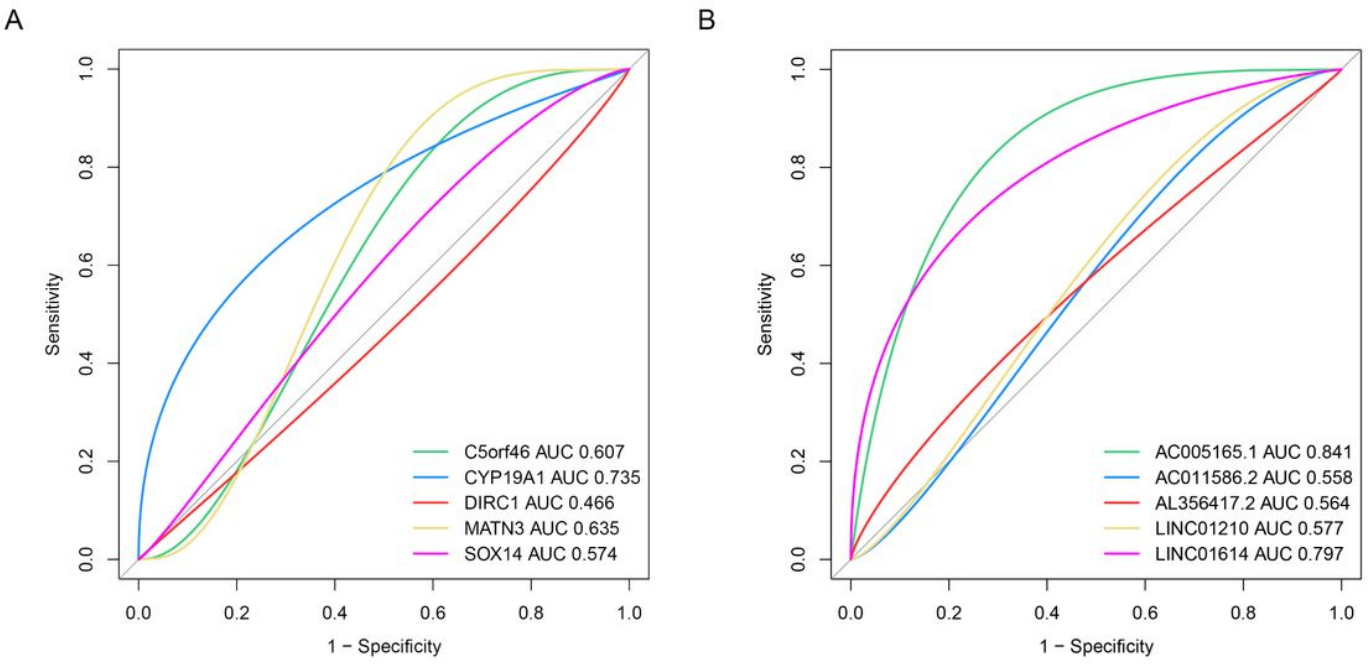

Figure 8

ROC analysis for the prognostic genes (A) and lncRNAs (B) in the GSE13911 and GSE70800, respectively.

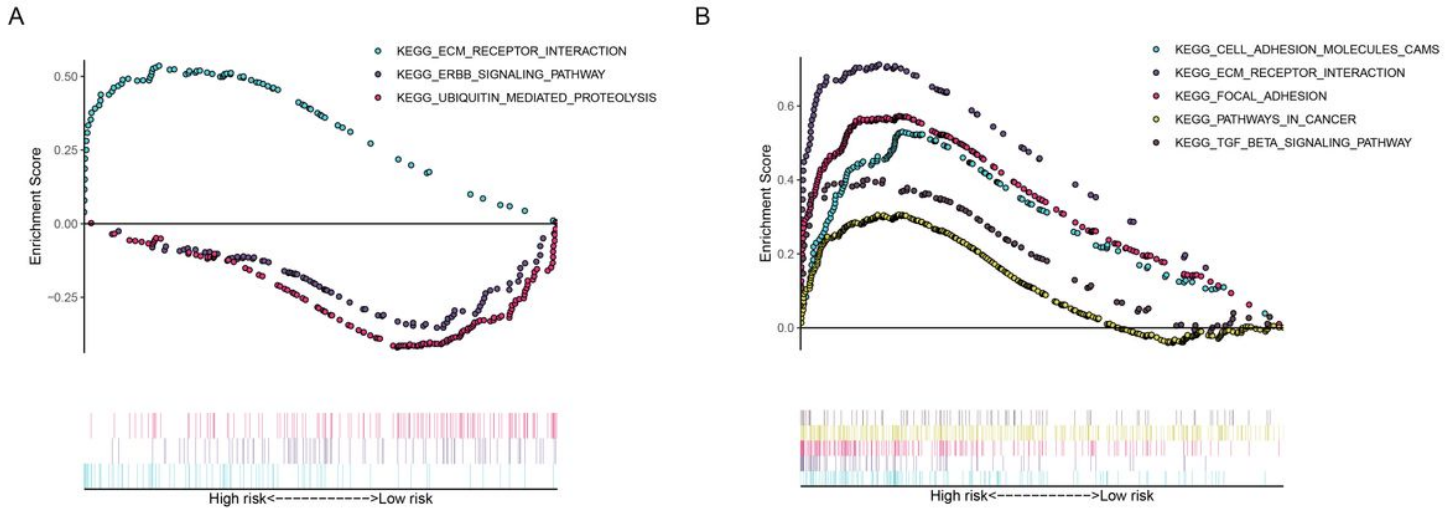

Figure 9

GSEA enrichment analysis for the risk group of gene model (A) and lncRNA model (B).