Identification of a Six-lncRNA Prognosis Model for Predicting Progression-Free Survival in Patients with Thyroid Cancer

Pei Ji
The First Affiliated Hospital of Nanjing Medical University
Tingyu Xu
The First Affiliated Hospital of Nanjing Medical University
Yifang Hu
The First Affiliated Hospital of Nanjing Medical University
Dai Cui
The First Affiliated Hospital of Nanjing Medical University
Zhongmin Wang (wangzhongmin@njmu.edu.cn)
The First Affiliated Hospital of Nanjing Medical University

Research Article

Keywords: thyroid cancer, long non-coding RNAs, progression-free survival, prognosis model

Posted Date: February 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1338110/v1

Abstract

Background: Thyroid cancer is the most common malignant tumor of the endocrine system. Long non-coding RNAs (lncRNAs) have been demonstrated as novel biomarkers for cancer prognosis.

Methods: In this study, we performed differential expression analysis of lncRNA expression profiles in GEO datasets. LASSO regression analysis was conducted to identify lncRNA-based prognosis model that can predict progression-free survival in thyroid cancer patients from The Cancer Genome Atlas (TCGA). Further functional analysis revealed the potential biological functions of the lncRNAs.

Results: A risk score model based on six lncRNA biomarkers were established after LASSO Cox regression analysis. The prognostic value of the 6-lncRNA prognosis model was successfully validated and ROC curves was analysed. Patients were classified into high- and low-risk groups using the 6-lncRNA signature-based risk score. Patients in the low-risk group had significantly better progression-free survival than high-risk group. The result of multivariate analysis showed that the six-lncRNA signature was independent from clinical features such as age, gender and stage. GO and KEGG enrichment analysis and estimation of immune infiltration suggested that the lncRNAs might closely associated with tumorigenesis.

Conclusion: Our study has constructed a novel six-lncRNA prognosis model to improve progression-free survival prediction in patients with thyroid cancer.

Background

Thyroid cancer (TC) is the most common endocrine malignancy cancerand its incidence is increasing [1-3]. It is estimated that about 90,000 new thyroid cancer cases and 6,800 thyroid cancer death in China in 2015 [4]. Papillary thyroid carcinoma (PTC) is the most common thyroid cancer subtype, usually characterized by slow proliferation and rare distant metastatic spread, causing a favorable prognosis for most patients [5]. In recent decades, there has been a huge improvement in early diagnosis and raising survival rate with the development of molecular diagnoses and targeted therapies. However, nearly 5% of thyroid cancers behave aggressively and are associated with distant metastasis. The tumorigenic processes of thyroid cancer have not yet been completely unraveled. The major clinical challenge remains to find effective molecular biomarkers that correlate with thyroid cancer prognosis to extend survival time of thyroid cancer patients.

Additionally, overdiagnosis and overtreatment are common problems associated with indolent diseases. The screening and identification of indolent TC and the treatment of these over-diagnosed cancers can increase the risk of injury to patients [6, 7]. Therefore, the use of effective and sensitive biomarkers to identify specific thyroid cancer patients and provide personalized treatment has become an urgent need. In the last decade, a growing body of studies have demonstrated the utility and superiority of long non-coding RNAs (lncRNAs) as novel biomarkers for cancer diagnosis, prognosis, and therapy [8]. However, our understanding of the prognostic value of lncRNAs is still very limited.

The purpose of this study was to explore the clinical value of lncRNA risk prediction model in thyroid cancer. We performed differently expression analysis of lncRNA expression profiles in two GEO datasets. Then we systematically investigate the prognostic value of lncRNAs in The Cancer Genome Atlas (TCGA) project. In the training cohort, we identified a set of six lncRNAs demonstrating an ability to stratify patients into high- and low-risk groups with significantly different progression-free survival (PFS) in 244 TCGA TC patients. The six-lncRNA signature was successfully validated on TCGA cohort of 244 patients, and we evaluated the specificity and sensitivity of the model by ROC curve analysis. We further used multivariate analysis which suggested that the six-lncRNA risk prediction model is independent of clinical features. The results of KEGG and GO pathway enrichment analysis showed the potential pathways these lncRNAs may affect. We also investigated the differences in immune cell proportions between high- and low-risk score groups.

Material And Methods

Patient Data Sets and screening of differentially expressed lncRNAs

The GSE33630 and GSE29265 datasets was downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). GSE33630 contains 60 thyroid carcinomas and 45 normal thyroids, while GSE29265 contains 29 thyroid carcinomas and 20 normal thyroids. Both two datasets was based on GPL570 platform ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array). The expression value of lncRNAs was compared between tumor and normal tissues by LIMMA package. |log2 fold-change (FC)| > 1 and *P*-value <0.05 were used as cut-off criteria.

Clinical information of patients with thyroid cancer was assessed from The Cancer Genome Atlas (TCGA) project (https://cancergenome.nih.gov/). A total of 488 TCGA thyroid cancer patients with both lncRNA expression profiles and clinical follow-up information were utilized in our study. The PFS data was downloaded from the UCSC Xena website (https://xena.ucsc.edu/). The detailed clinical features of cohorts were listed in Table 1.

Table 1: Patient characteristics from TCGA database.

Characteristics	Training dataset (n=244)	Validation dataset (n=244)	Overall (n=488)
Age			
≤60	188	200	388
>60	56	44	100
Survival status			
DiseaseFree	218	223	441
Recurred/Progressed	26	21	47
Gender			
Female	185	173	358
Male	59	71	130
pT			
T1	60	91	141
T2	86	75	161
T3	83	77	166
T4	9	9	18
pN			
N0	106	119	225
N1	112	104	216
pM			
M0	130	143	273
M1	4	4	8
pStage			
Stage I	129	150	279
Stage II	27	23	50
Stage III	59	47	106
Stage IV	28	23	51

Construction of prognostic biomarkers

We carried out Lasso-penalized Cox regression based on differently expression lncRNAs to construct the model for prediction of PFS. All the thyroid cancer patients in TCGA cohort were partitioned into a training

cohort including 244 samples for identifying prognostic lncRNA signature and building prognostic risk model, and a testing cohort including other 244 samples for validating its prognostic value. The study flowchart is shown in Figure 1.

The LASSO Cox regression model was analysed using the 'glmnet' package and the optimal λ was chosen through cross-validation routine with 10-fold cross-validation (Supplementary Figure 1A). Based on the cut-off of the median risk score, the patients were divided into high- and low-risk groups.

Survival analysis and cox regression

To assess the independence of our risk prediction model, we performed univariate and multivariate Cox regression analyses. The receiver-operator characteristics (ROC) and nomogram were established based on the results of risk model. Kaplan-Meier (K-M) curve analysis was used to evaluate the PRS between high- and low-risk groups. Kaplan-Meier method was also performed in the patients stratified by age, gender and stage to verify the predictive value of risk model and clinical features.

Function enrichment analysis

The differently expressed mRNAs (DEmRNAs) was identified in the same patient group of TCGA datasets which satisfied the threshold criteria of log2|fold change|>1.0 and P-value<0.05. The co-expression relationships between the lncRNAs in risk model and DEprotein-coding genes were calculated using Pearson correlation coefficients with the cut-off value was >0.6 or <-0.6. Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of the DEmRNAs were performed to explore the potential biological function using the 'clusterProfiler' package (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html).

Estimation of immune infiltration score

To explore the differences of immune cell subtypes, we assess the proportions of 22 types of infiltrating immune cells using CIBERSORT algorithm [9]. The percentage of each type of immune cell in the samples was calculated based on expression data. Mann-Whitney U test was used to compare immune cell subtypes in the high- and low-risk groups.

Statistical analysis

R software (version = 3.6.1) was used for all statistical tests. T-test for continuous variables and $\chi 2$ test for categorical variables were performed to assess the relationship between the clinical features. Univariate and multivariate Cox regression analyses were performed to verify if the risk score calculated from the expression level of lncRNAs were independent of other clinical characteristics. $P<0.05$ was considered statistically significant.

Results

Indentification of DElncRNAs between thyroid tumor and normal tissues. A flow chart of the analysis procedure was developed to describe our study (Figure 1). In this study, differently expressed lncRNAs (DElncRNAs) were identified from two datasets of GSE33630 and GSE29265 after standardization of the microarray data. Based on the cutoff criteria of |log2 FC| > 1 and *P*-value <0.05, a total of 30 DElncRNAs were identified, including 9 up-regulated and 21 down-regulated lncRNAs. The results were presented in heatmap and volcano plots as showed in Supplementary Figure 2.

Establishment of a six-lncRNA risk model associated with TC survival

We adopted Lasso-penalized Cox analysis to narrow the lncRNAs for prediction of the PFS among the pool of DElncRNAs. Six candidate lncRNAs were selected to build the predictive model. Accordingly, the risk score was calculated with the coefficients weighted by the LASSO penalized regression, as risk score=(-0.0146×expression of LINC01204) + (0.0661×expression of AC100810.1) + (0.1470×expression of AC009120.2) + (-0.1420×expression of AC090617.5) + (-0.0537×expression of LINC00511) + (0.0514×expression of AJ009632.2). The risk score for each patient was calculated and all patients were classified into either high- or low-risk score group based on the median value of risk score. Patients in the lower-risk group had significantly better PFS than higher-risk group. The accuracy of the 6-lncRNA risk prediction model was assessed by calculating the AUCs in the training, validation datasets and all TCGA samples (Figure 2).

The AUCs for the lncRNA risk model in training and validation stage were 0.617 and 0.570, while AUC of all TCGA samples was 0.594. We also constructed a nomogram to forecast the PRS in thyroid cancer patients which integrating conventional risk factors like age, gender and stage (Supplementary Figure 1B).

Moreover, univariate Cox regression analysis was performed on the training and validation set respectively to evaluate the independent prognostic value of the risk score. According to the results (Figure 3), the risk score based on 6 lncRNAs was able to effectively predict the PRS of TC patients. Multivate Cox regression analysis demonstrated that the 6-lncRNA risk model was an independent risk factor when adjusting for the other clinical features (Figure 3).

To further access the predictive value of the 6-lncRNA risk model and clinical features, the patients were stratified by age, gender and stage, the lasso-score was still a significant prognostic model for patients in the high-risk group with poorer prognosis (Figure 4).

Functional Analysis

To explore the function of the 6-lncRNA model, we examined the differently expressed mRNAs in the same patient group with the threshold criteria of |log2FC| > 1.0 and *P*-value < 0.05. The co-expressed lncRNA-mRNA pairs was identified by calculating the Pearson correlation coefficients. DEmRNAs correlated with at least one of the lncRNAs in the model were included in the following analysis (Pearson correlation coefficient>0.60). We performed gene ontology (GO) enrichment analysis and Kyoto

encyclopedia of genes and genomes (KEGG) pathway analysis on these DEmRNAs to investigate the biological roles of these lncRNAs. The top 20 GO and KEGG terms were visualized in Figure 5.

High and low risk score groups showed differential immune cells expression (Supplementary Figure 3). Compared with low-risk patients, high-score group contained higher proportion of CD8 T cells, NK cells, Monocytes, Macrophages M2. The proportions of naive B cells, CD4 T cells, active and resting Dendritic cells were relatively higher in low-risk patients.

Discussion

The incidence of thyroid cancer has been increasing all over the world. Although most thyroid cancer patients have a benign prognosis but some patients suffer from locoregional or distant tumor recurrence [10]. In recent years, long non-coding RNAs have been reported that they have complex biological functions and may regulate proteins involved in tumorigenesis [11, 12]. An increasing number of studies proved that LncRNAs can act as a tumor suppressor genes or oncogenes and play an important role in tumor progress, proliferation, migration and epigenetic regulation [13, 14]. Various lncRNAs were demonstrated to be associated with thyroid cancer such as MALAT1, H19, BANCR, HOTAIR, which can be used as novel biomarkers for early diagnosis and treatment [15].

In this study, we established and validated a 6-lncRNA risk prediction model, including LINC01204, AC100810.1, AC009120.2, AC090617.5, LINC00511 and AJ009632.2, to predict PFS for patients with thyroid cancer. Based on microarray data from GEO database, we identified aberrantly expressed lncRNAs between tumor and normal tissues. After LASSO regression analysis in training set, six lncRNAs were selected to construct the prognostic prediction model for PFS. Patients with lower lasso-score have better survival prognosis than those with high lasso-score. To further illustrate the robustness of the six-lncRNA risk model, the prognostic value was further validated using the rest TCGA patients. The result of validation dataset confirmed the good reproducibility in predicting patients' outcome. The six-lncRNA signature was independent of age, gender and stage according to the results of multivariate analysis in training, validation and all TCGA samples. When patients were stratified by clinical features, K-M survival analysis showed that the risk model remains the same predictive power in different groups.

Our study established 6-lncRNA-based risk model for PFS which is associated with the prognosis of thyroid cancer. However, most of the lncRNAs in this model have not been functionally annotated and the biological functions remain unclear. Only AC100810.1 and AC009120.2 were reported to be associated with prognosis of cancer patients in previous studies. AC100810.1 was identified as one of immune-related lncRNAs with prognostic value of breast cancer patients by Zheng et al [16]. Wang et al. developed and verificated an immune-related lncRNA signature including AC009120.2 which can accurately determine the prognosis of patients with bladder cancer [17]. Since the infiltration of immune cells in TME plays a crucial role on cancer progression and patient prognosis [18, 19], we performed CIBERSORT algorithms and found significant difference of the immune cell infiltration between high- and low- risk score groups.

We conducted correlation analysis between lncRNAs in our prognosis model and the differently expressed mRNAs in TCGA. The KEGG pathway enrichment analysis suggested that the co-expressed mRNAs were enriced in Spliceosome, RNA polymerase and TNF signaling pathway. The GO enrichment analysis demonstrated that mRNAs associated with the six lncRNAs mainly involved in RNA splicing, spliceosomal complex and DNA-templated transcription. These findings might generate a cheap molecular test but the sample size of our study was limited. The value of this model need to be assessed in further cohort studies.

Conclusion

we developed a 6-lncRNAs prognostic model to predict progression-free survival for patients with thyroid cancer. Our results may lead to a useful predictive model to classify risk subgroups for thyroid cancer patients.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors approved the manuscript and the submission to BMC Medical Genomics.

Data Availability

All data could be downloaded from public databases (TCGA and GEO). The GSE33630 and GSE29265 datasets was downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). Clinical information and the lncRNA expression profiles of 488 patients with thyroid cancer was assessed from The Cancer Genome Atlas (TCGA) project (https://cancergenome.nih.gov/).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Funding

None.

Authors' contributions

JP performed data analyses and wrote the manuscript. XTY and HYF contributed significantly in data analyses and manuscript revision. WZM and CD conceived and designed the study. All authors have read and approved the final manuscript.

Acknowledgements

None.

References

1. Cabanillas ME, McFadden DG, Durante C. Thyroid cancer. Lancet. 2016;388(10061):2783-2795.

2. Li N, Du XL, Reitzel LR, Xu L, Sturgis EM. Impact of enhanced detection on the increase in thyroid cancer incidence in the United States: review of incidence trends by socioeconomic status within the surveillance, epidemiology, and end results registry, 1980-2008. Thyroid. 2013;23(1):103-110.

3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7-30.

4. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ, He J. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115-132.

5. Cabanillas ME, McFadden DG, Durante C. Thyroid cancer. Lancet. 2016;388(10061):2783-2795.

6. Lin JS, Bowles E, Williams SB, Morrison CC. Screening for Thyroid Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA. 2017;317(18):1888-1903.

7. Kitahara CM, Sosa JA. The changing incidence of thyroid cancer. Nat Rev Endocrinol. 2016;12(11):646-653.

8. Qu L, Wang ZL, Chen Q, Li YM, He HW, Hsieh JJ, Xue S, Wu ZJ, Liu B, Tang H *et al*. Prognostic Value of a Long Non-coding RNA Signature in Localized Clear Cell Renal Cell Carcinoma. Eur Urol. 2018;74(6):756-763.

9. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453-457.

10. Ho AS, Luu M, Barrios L, Chen I, Melany M, Ali N, Patio C, Chen Y, Bose S, Fan X *et al*. Incidence and Mortality Risk Spectrum Across Aggressive Variants of Papillary Thyroid Carcinoma. Jama Oncol. 2020;6(5):706-713.

11. Bhan A, Soleimani M, Mandal SS. Long Noncoding RNA and Cancer: A New Paradigm. Cancer Res. 2017;77(15):3965-3981.

12. Sedaghati M, Kebebew E. Long noncoding RNAs in thyroid cancer. Curr Opin Endocrinol Diabetes Obes. 2019;26(5):275-281.

13. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet. 2014;15(1):7-21.

14. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011;43(6):904-914.

15. Mahmoudian-Sani MR, Jalali A, Jamshidi M, Moridi H, Alghasi A, Shojaeian A, Mobini GR. Long Non-Coding RNAs in Thyroid Cancer: Implications for Pathogenesis, Diagnosis, and Therapy. Oncol Res Treat. 2019;42(3):136-142.

16. Li Z, Li Y, Wang X, Yang Q. Identification of a Six-Immune-Related Long Non-coding RNA Signature for Predicting Survival and Immune Infiltrating Status in Breast Cancer. Front Genet. 2020;11:680.

17. Wang J, Shen C, Dong D, Zhong X, Wang Y, Yang X. Identification and verification of an immune-related lncRNA signature for predicting the prognosis of patients with bladder cancer. Int Immunopharmacol. 2021;90:107146.

18. Fridman WH, Pages F, Sautes-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat Rev Cancer. 2012;12(4):298-306.

19. Fridman WH, Zitvogel L, Sautes-Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. Nat Rev Clin Oncol. 2017;14(12):717-734.

Figures

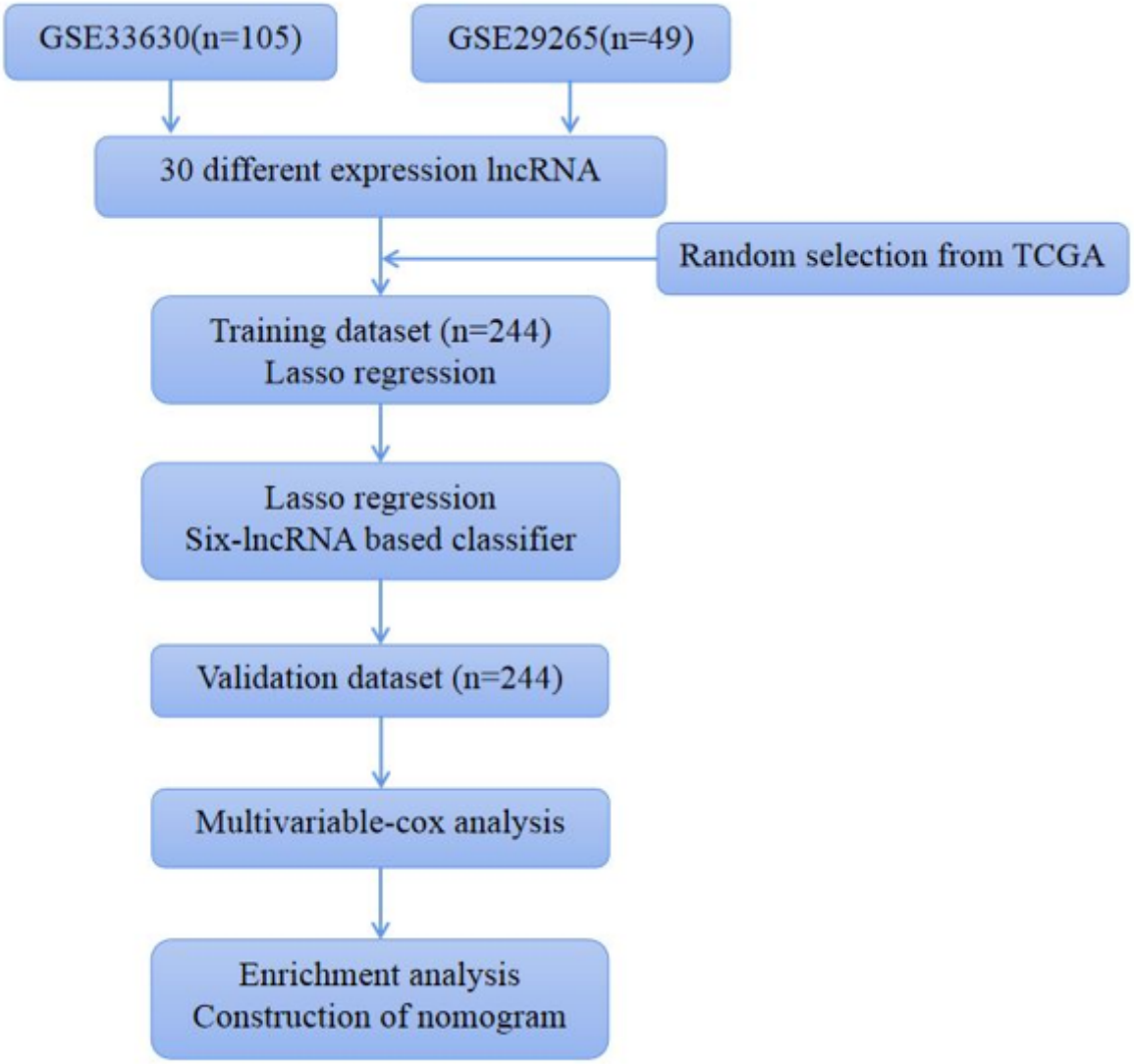

Figure 1

Flowchart of the identification of the six-lncRNA expression signature.

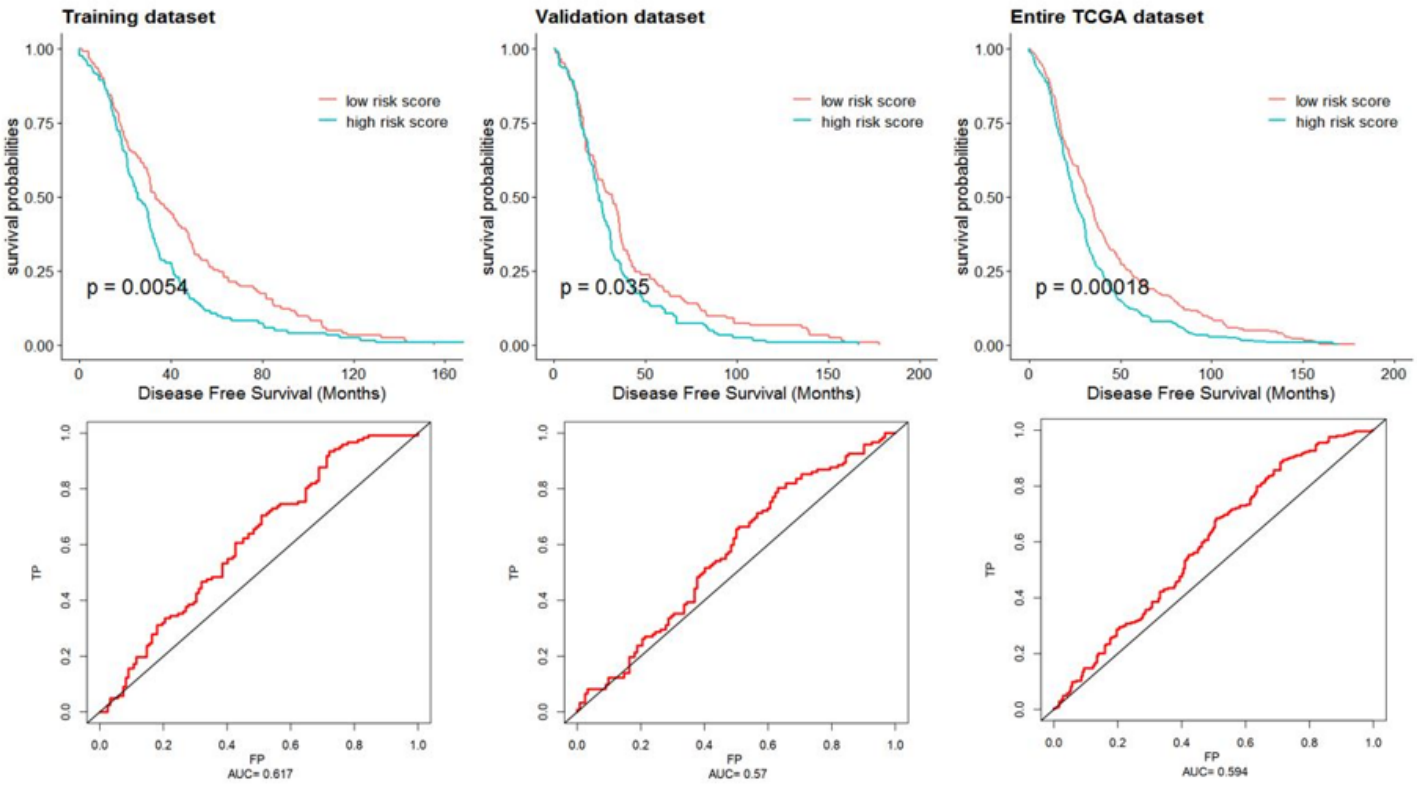

Figure 2

Kaplan-Meier and ROC analyses in the training, testing groups and entire TCGA dataset.

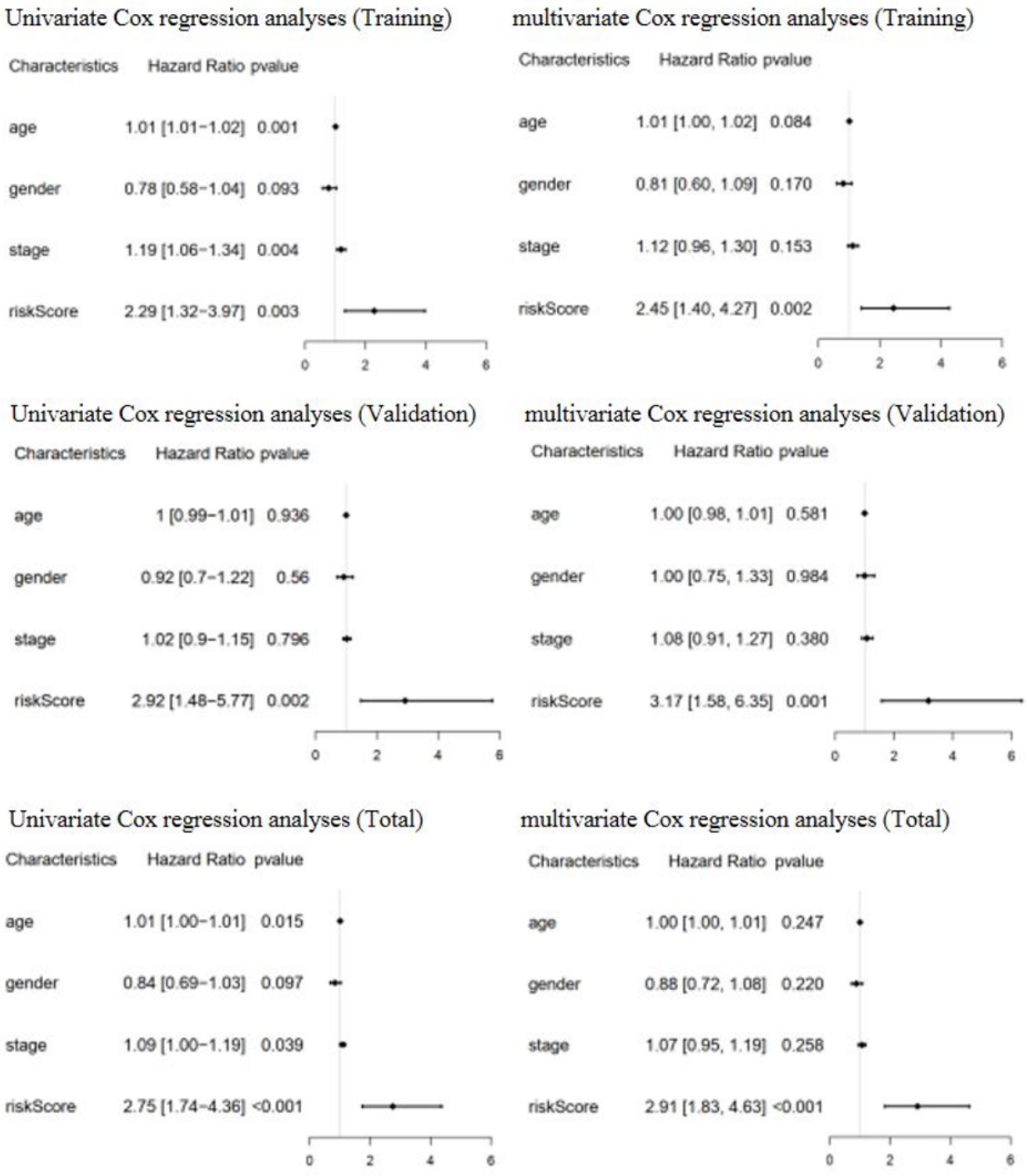

Figure 3

The Cox regression analysis for evaluating the independent prognostic value of the risk score.

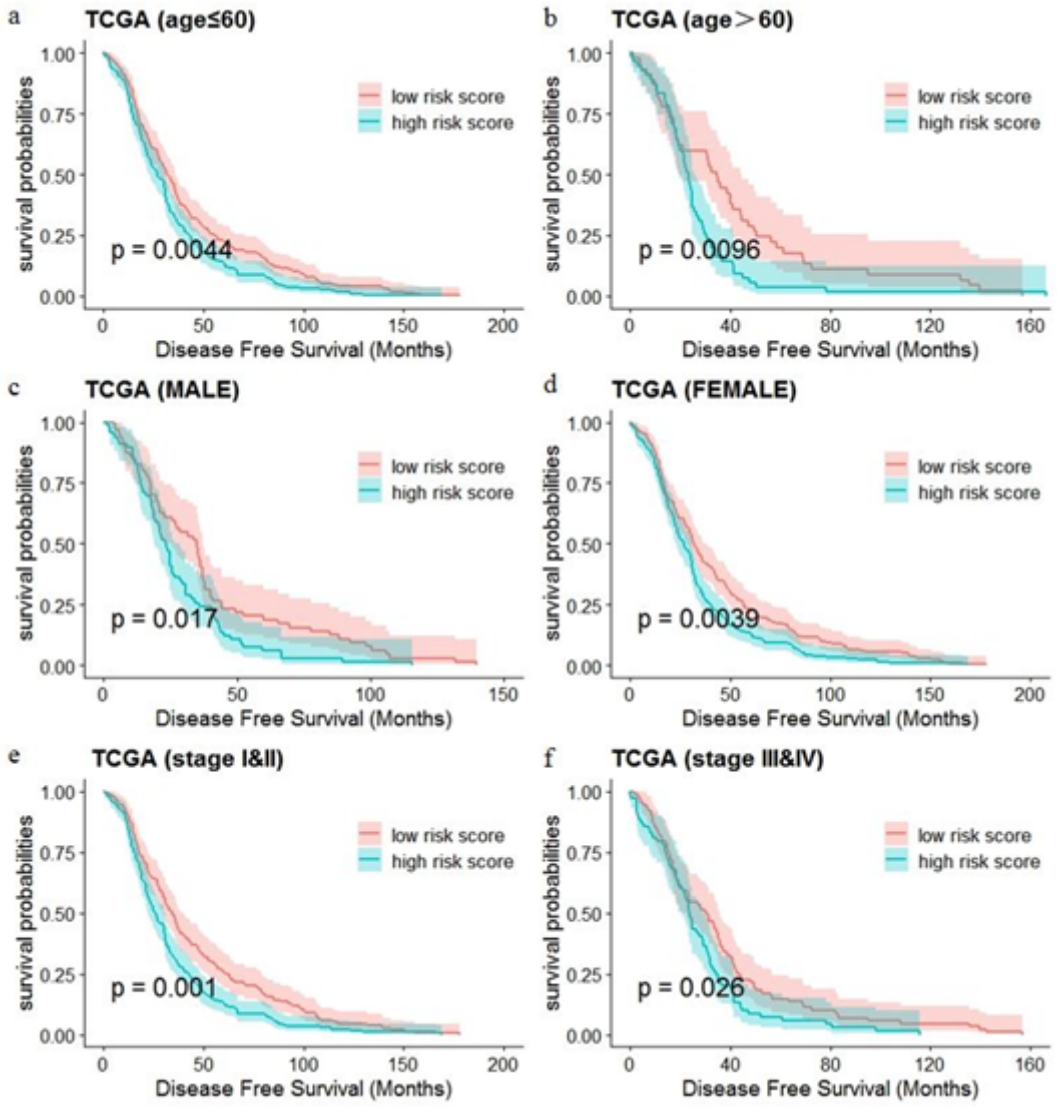

Figure 4

Stratification analysis by age and gender. Kaplan-Meier curve analysis of PRS in high- and low-risk groups for young patients (a) and old patients (b). Kaplan-Meier curve analysis of PRS in high- and low-risk groups for male patients (c) and female patients (d). Kaplan-Meier curve analysis of PRS in high- and low-risk groups for patients with different clinical stage (e,f).

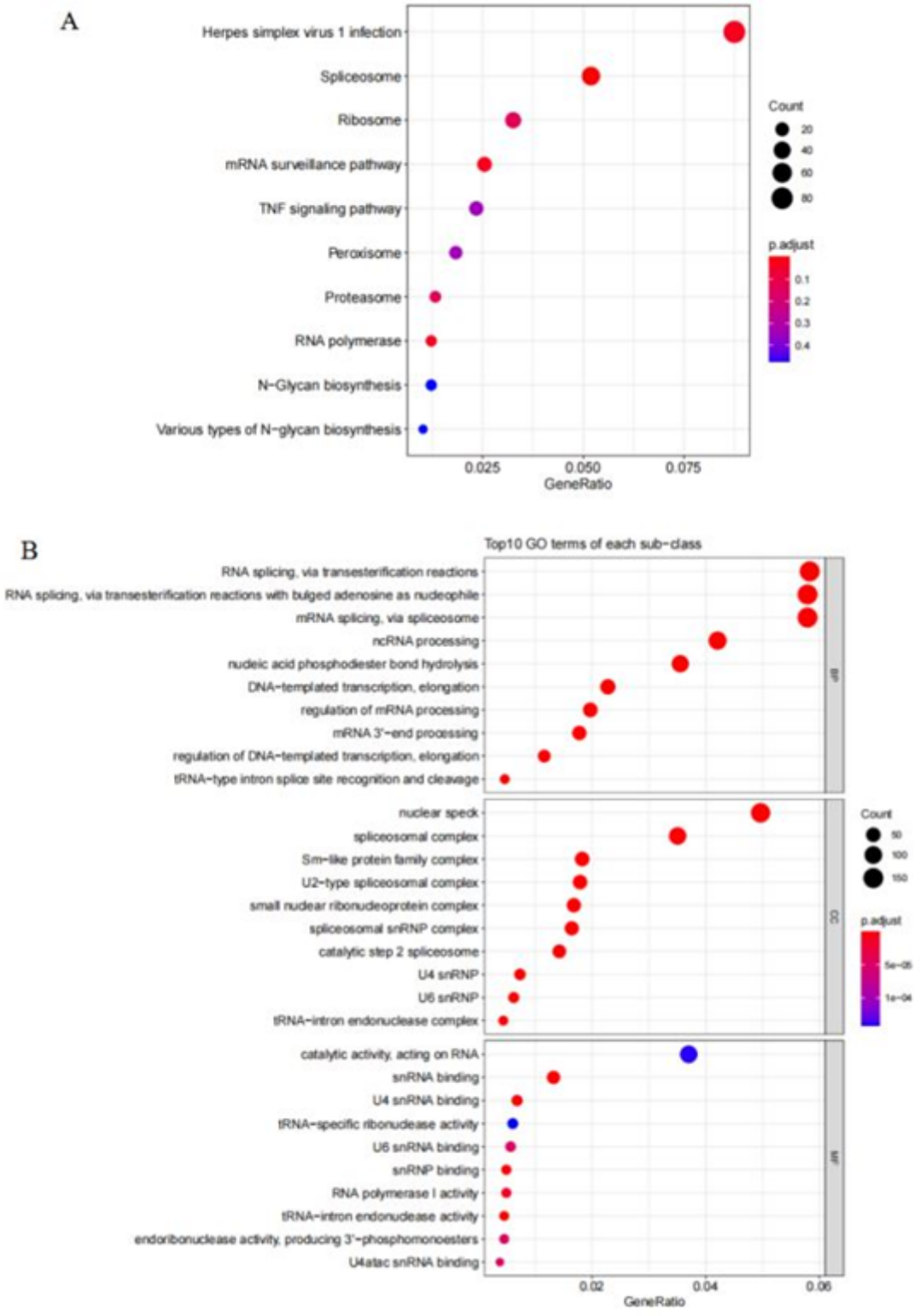

Figure 5

Functional enrichment analysis of the six-lncRNA risk model based on co-expressed mRNAs. A The top 10 Significantly enriched KEGG pathways. B Top10 GO terms of each sub-class in GO enrichment analysis.

Supplementary Files

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

- SupplementaryMaterials.docx