Tumor microenvironment related novel signature predict lung adenocarcinoma survival

Juan Chen [1], **Rui Zhou** [Corresp. 1]

[1] Respiratory Medicine, The Second Xiangya Hospital of Central South University, Changsha, Hunan, China

Corresponding Author: Rui Zhou
Email address: zhourui2355@csu.edu.cn

Background:Lung adenocarcinoma (LUAD) is the most common histological type of lung cancers, which is the primary cause of cancer-related mortality worldwide. Growing evidence has suggested that tumor microenvironment (TME) plays a pivotal role in tumorigenesis and progression. Hence, we investigate the correlation of TME related genes with LUAD prognosis.

Method:The information of LUAD gene expression data was obtained from The Cancer Genome Atla(TCGA). According to their immune/stromal scores calculated by the ESTIMATE algorithm, differentially expressed genes(DEGs) were identified. Then, we performed univariate Cox regression analysis on DEGs to obtain genes that are apparently bound up with LUAD survival (SurGenes). Functional annotation and protein-protein interaction (PPI) was also conducted on SurGenes. By validating the SurGenes with another data sets of lung cancer from Gene Expression Omnibus (GEO), 106 TME related SurGenes were generated. Further, intersection analysis was executed between the 106 TME related SurGenes and hub genes from PPI network, PTPRC and CD19 were obtained. And Gene Set Enrichment Analysis and CIBERSORT analysis were performed on PTPRC and CD19. Based on the TCGA LUAD dataset, we conducted factor analysis and Step-wise multivariate Cox regression analysis for 106 TME related SurGenes to construct the prognostic model for LUAD survival prediction. LUAD dataset in GEO (GSE68465) was used as the testing dataset to confirm the prognostic model. Multivariate Cox regression analysis was used between risk score from the prognostic model and clinical parameters.

Result:106 TME related genes were collected in our research totally, which were markablely correlated with the overall survival(OS) of LUAD patient. Bioinformatics analysis suggest them mainly concentrated on immune response, cell adhesion, and extracellular matrix. More importantly, among 106 TME related SurGenes, PTPRC and CD19 were highly interconnected nodes among PPI network and correlated with immune activity, exhibiting significant prognostic potential. The prognostic model was a weighted linear combination of the 106 genes, by which the low-OS LUAD samples could be separated from the high-OS samples with success. This model was also able to rebustly predict the situation of survival (training set: p-value<0.0001,area under the curve (AUC) =0.649; testing set: p-value=0.0009, AUC=0.617). By combining with clinical parameters, the prognostic model was optimized. The AUC achieved 0.716 for 3 year and 0.699 for 5 year.

Conclusion:A series of TME-related prognostic genes were acquired in this research, which could reflect immune disorders within tumor microenvironment, and PTPRC and CD19 show the potential to be an indicator for LUAD prognosis and tumor microenvironment modulation. The prognostic model constructed base on those prognostic genes presented a high predictive ability, and may have clinical implications in the overall survival prediction of LUAD.

Tumor microenvironment related novel signature predict lung adenocarcinoma survival

Juan Chen[1]

[1]Respiratory Medicine, The Second Xiangya Hospital of Central South University. Changsha, Hunan, China

Corresponding Author:

Rui Zhou[1]

No.139 Middle Renmin Road, Changsha, Hunan, 410011, China.

Email address: zhourui2355@csu.edu.cn

Abstract

Background: Lung adenocarcinoma (LUAD) is the most common histological type of lung cancers, which is the primary cause of cancer-related mortality worldwide. Growing evidence has suggested that tumor microenvironment (TME) plays a pivotal role in tumorigenesis and progression. Hence, we investigate the correlation of TME related genes with LUAD prognosis.

Methods: The information of LUAD gene expression data was obtained from The Cancer Genome Atlas (TCGA). According to their immune/stromal scores calculated by the ESTIMATE algorithm, differentially expressed genes (DEGs) were identified. Then, we performed univariate Cox regression analysis on DEGs to obtain genes that are apparently bound up with LUAD survival (SurGenes). Functional annotation and protein-protein interaction (PPI) was also conducted on SurGenes. By validating the SurGenes with another data sets of lung cancer from Gene Expression Omnibus (GEO), 106 TME related SurGenes were generated. Further, intersection analysis was executed between the 106 TME related SurGenes and hub genes from PPI network, and PTPRC and CD19 were obtained. And Gene Set Enrichment Analysis (GSEA) and CIBERSORT analysis were performed on PTPRC and CD19. Based on the TCGA LUAD dataset, we conducted factor analysis and Step-wise multivariate Cox regression analysis for 106 TME related SurGenes to construct the prognostic model for LUAD survival prediction. LUAD dataset in GEO (GSE68465) was used as the testing dataset to confirm the prognostic model. Multivariate Cox regression analysis was used between risk score from the prognostic model and clinical parameters.

Results: 106 TME related genes were collected in our research totally, which were markablely correlated with the overall survival(OS) of LUAD patient. Bioinformatics analysis suggest them mainly concentrated on immune response, cell adhesion, and extracellular matrix. More importantly, among 106 TME related SurGenes, PTPRC and CD19 were highly interconnected nodes among PPI network and correlated with immune activity, exhibiting significant prognostic potential. The prognostic model was a weighted linear combination of the 106 genes, by which the low-OS LUAD samples could be separated from the high-OS samples with success. This model was also able to rebustly predict the situation of survival (training set: p-value<0.0001,area under the curve (AUC) =0.649; testing set: p-value=0.0009, area under the curve (AUC) =0.617). By combining with clinical parameters, the prognostic model was optimized. The AUC achieved 0.716 for 3 year and 0.699 for 5 year.

Conclusion: A series of TME-related prognostic genes were acquired in this research, which could reflect immune disorders within tumor microenvironment, and PTPRC and CD19 show the potential to be an indicator for LUAD prognosis and tumor microenvironment modulation. The prognostic model constructed base on those prognostic genes presented a high predictive ability, and may have clinical implications in the overall survival prediction of LUAD.

Introduction

Lung cancer is the most common cause of cancer-related deaths worldwide.[1] Non-small cell lung cancer (NSCLC) represents 85% of lung cancers, mainly including lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LSCC).[2] Notably, the incidence of LUAD significantly increased and surpassed of LSCC and it constitutes nearly 40% of all lung malignancies.[3] Although with multiple kinds treatment methods, including surgery, chemotherapy, radiotherapy, and dramatic treatment shift of some targeted therapeutic agents, the prognosis for LUAD patients remains poor worldwide, with 5-year relative survival currently at 18%.[4-6] Therefore, understanding the mechanism of carcinogenesis and therapeutics of lung cancer is quietly important.

At present, the pathogenesis of LUAD has not been adequately described, however, there are increasing study support the view of tumor microenvironment critically influence gene expression of tumor tissues, and the clinical outcomes further. The tumor microenvironment (TME) is a complicated mixture, immune and stromal cells are two major types of non-tumor factors. They have been certified to promote the development of diagnostic and prognostic assessment process of lung cancer[7-9]. Hence, exploring the molecular composition and function of the TME is critical to effectively manage cancer progression and immune response[10-11].

Previous studies have solved the problem of the complexity of tumor infiltrating immune cells in LUAD. Most of these studies assessed tumor-infiltrating immune cells by immunohistochemical analysis of a single marker.[14-16] Fortunately, computational analysis of tumor immune cell

interactions is now available by bioinformatics tools. In 2013, Yoshihara firstly reported an algorithm called ESTIMATE (Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data). The infiltration level of stromal and immune cells was predicted by calculating stromal and immune score based on data from the Cancer Genome Atlas (TCGA) data sets by ESTIMATE.[17] Subsequent reports applied the innovative algorithm to different cancer, such as prostate cancer,[18] breast cancer,[19] and colon cancer,[20] which further confirmed the effectiveness of the algorithm based on big database.

We analyzed immune/stromal scores of LUAD cohorts from TCGA, which was derived from ESTIMATE algorithm, and extracted a series of TME associated prognostic gene in LUAD. Among those TME associated prognostic gene, PTPRC and CD19 were particular interested for they are highly interconnected in PPI network, and closely correlated with immune-related activity by GSEA analysis and CIBERSORT analysis. Combine those prognostic genes, we constructed a prognostic model which could provide a moderate OS prediction for LUAD alone, and provide a robust prediction with clinical parameters.

Materials & Methods

Data acquisition

From TCGA database, including 576 samples, gene expression RNAseq and clinical information such as pathological stage, survival of LUAD were downloaded and prepared for the analysis of differential expression. The data set was submitted by University of North Carolina TCGA genome characterization center based on the Illumina HiSeq 2000 RNA Sequencing platform (Oct 13, 2017). We downloaded the estimate, immune and stromal scores of 517 LUAD samples from TCGA in ESTIMATE website. Stromal score ~~that~~ captures the presence of stroma in tumor tissue, immune score ~~that~~ represents the infiltration of immune cells in tumor tissue, and estimate score ~~that~~ infers tumor purity. ~~And~~ the Estimate score is equal to the Immune score plus the Stromal score. Fifty-nine samples were deleted for lack of immune and stromal scores in ESTIMATE website. From GEO database with the accession number GSE68465, consisting of 442 samples, gene expression profile and LUAD clinical data were obtained to validate the differential expression gene and test the prognostic model. The data set was submitted by Mervi Heiskanen based on Affymetrix HG-U133A (May 01, 2015).

Distinguishing of differentially expressed genes (DEGs)

The raw data of TCGA was pre-processed by limma algorithm.[21] The adjusted P-values (adj. p) < 0.05 and|Log2 (FC)|>1 were set as the cut-offs to screen for differentially expressed genes.

Heatmaps and clustering analysis

ClustVis web tool was used to create heatmaps and clustering.[22]

Enrichment analysis of DEGs.

DAVID conducted functional enrichment analysis, including biological process, molecular function and cell component to analyze DEGs,[23] an essential foundation for visualization, annotation, and integrated discovery. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways was also performed by the database of DAVID. False discovery rate (FDR) < 0.05 was used as the cut-off. Biological process and KEGG pathway analysis were also conducted on module gene, P<0.05 was considered to indicate a statistically significant difference.

Screening of survival-related DEGs

Kaplan–Meier plots and univariate Cox regression were used to visualize the association between the genes expression and overall survival of patients to explore its prognostic value. Statistical significance was examined using the Log-rank test. P < 0.05 were considered statistically significant.

Integration of PPI network and screening of modules

STRING database was used to retrieve predicted PPIs. Only experimentally validated interactions with a combined score >0.4 of all associations obtained in STRING were selected to construct the PPI network using Cytoscape software.[24] The Molecular Complex Detection (MCODE) plugin in Cytoscape was then utilized to find clusters of the PPI network. Degree more than 90 was set as a cut-off criterion and identified as hub genes.

Gene Set Enrichment Analysis

From Molecular Signatures Database, Hallmark and C7 gene sets v6.2 collections were downloaded as the target sets. GSEA performed using the software gsea-3.0. 517 samples from TCGA were used for GSEA, and only gene sets with NOM p < 0.05 and FDR < 0.25 were considered as signifificant.

TICs Profile

CIBERSORT computational method was applied for estimating the TIC abundance profifile in all tumor samples, which followed by quality fifiltering that only 517 tumor samples with p < 0.05 were selected for the following analysis.

Construction of the risk assessment model

To reduce the dimensionality and eliminate collinearity, factor analysis was performed on 106 prognostic genes. Then we used the Step-wise multivariate Cox regression analysis to obtain factors correlated with overall survival significantly(SurFactors). By combining the score coefficient of factors weighted by their regression coefficients and standard deviation, the risk index of each gene was calculated as follows:

$$\alpha_i = \frac{\sum_{j=1}^{m} \beta_j \times fac_{ij}}{std_i}$$

where α_i was the risk index of gene i , m the number of SurFactors, β_j the regression coefficient of factor j in multivariate Cox regression analysis, fac_{ij} the factor score coefficient of j-th factor over gene i and std_i the standard deviation of gene i.

By combining the expression values of prognostic genes weighted by their risk index, the following risk scores can be established for each patient:

$$Risk\_score = \sum_{i=1}^{n} \exp_i * \alpha_i$$

where n was the number of prognostic genes, and $\exp_i$ the expression value of gene i.

Survival and statistical analysis

We calculated risk score (RS) for every sample from testing data set based on the prognostic model and divided the specimens into two different groups based on the median RS. Kaplan–Meier survival curves were drawn and compared the two subgroups via log-rank tests. We divided the samples into two group according to their overall survival, receiver operating characteristic (ROC) curves were drawn by IBM SPSS statistics 22, and area under the curve (AUC) was calculated. Moreover, the univariate and multivariate analyses of survival were conducted to identify the prognostic factors for LUAD patients from TCGA data set. A nomogram and Calibration plots were established based on the TCGA LUAD cohort. The nomogram and calibration plot analysis were conducted by using the R package “rms” and “rmda.” All tests were two-tailed, and $P < .05$ was considered statistically significant. All statistical analyses were conducted using R software version 3.4.2.

Results

ESTIMATE algorithm of LUAD

Transcriptional expression profiles and clinical information of 517 LUAD patients were collected from TCGA database. Among them, 277 (53.6%) patients were female, 240 (46.4%) were male. Pathological stage included 277 patients (53.6%) of stage Ⅰ,122 (23.6%) stage Ⅱ,84 (16.3%)stage Ⅲ,26 (5%) stage Ⅳ and 8 (1.5%) unknown. Based on ESTIMATE algorithm, the scores of stromal and immune were calculated, ranging from-1355.85 to 3286.67 and -1959.31 to2098.77, respectively. The Estimate score was significantly associated with Pathological stage (Figure 1A, p=0.0436). The lowest Estimate score was at the most advanced pathological stage, the stage IV. Immune score showed remarkable prognostic potential, correlated with the pathological stage (Figure 1B, p=0.026), while Stromal score showed no correlation with the pathological stage (Figure 1C, p=0.145).

To explore whether the potential correlation existing between survival benefits and immune/stromal scores, 517 LUAD patients were divided into high and low score groups based on their scores. Kaplan-Meier survival curves (Figure 1D) showed that immune score significantly correlated with overall survival. Compared to the cases in the low score group, patients with high immune scores have longer median overall survival (1725 d vs. 1235 d, p = 0.0152 in log-rank test). Although there was no statistically significant correlation between the stromal score and overall survival, patients with higher stromal scores had longer median overall survival (Figure 1E, 1830 d vs. 1293 d, p= 0.0599 in log-rank test).

Identification of DEGs

To identify DEGs profiles according to immune/stromal scores, we obtained the gene expression array of 517 LUAD patients obtained TCGA database. Based on the comparison of immune scores of high group and low group, after analysis with the limma software package algorithm, there were 903 genes up-regulated and 56 genes down-regulated. We draw the heatmap with the top highly variate gene to (Figure 2A). Similarly, 1007 up-regulated genes and 30 down regulated genes were obtained for high stromal score compared with low score (Figure 2B). It is obvious that up-regulated DEGs, no matter for comparison based on immune scores nor stromal scores, take the major part in total DEGs. Therefore, we decide to focus on up-regulated DEGs in further analysis.

Enrichment analysis of DEGs

To reveal the biological function of the DEGs, GO and KEGG pathway enrichment analysis were performed for the up-regulated DEGs. GO analysis showed that immune up-regulated DEGs were remarkably enriched in immune response by biological processes, receptor activity

by molecular functions, and plasma membrane by cellular component, respectively (Figure 3A-C). Similarly, stromal up-regulated DEGs mainly enriched in cell adhesion by biological processes, extracellular matrix binding by molecular functions, and extracellular space by cellular component, respectively (Supplementary Figure 1A-C). KEGG pathway were both mainly enriched in cytokine-cytokine receptor interaction and cell adhesion molecules (CAMs) pathway(Figure 3D, Supplementary Figure 1D).

Overall survival analysis of DEGs

To make further efforts to elucidate if up-regulated DEGs could give benefits to the LUAD patient survival, Kaplan–Meier survival and univariate Cox proportional hazards regression analyses were performed on the up-regulated DEGs. The cases whose overall survival <30 days were excluded. The results showed that 446 immune related DEGs correlated with patient survival($p<0.05$), 387 stromal related DEGs correlated with patient survival($p<0.05$). Among all those genes, there were 291 duplicates of immune and stromal genes, a total of 542 genes (prognostic DEGs) associated with overall survival in patients with LUAD (Table S1).

Construction of PPI network by prognostic gene

Based on the STRING database, PPI network was obtained by using Cytoscape software to clarify the interaction between prognostic DEGs. The network was constructed by 13 modules, and we choose the top three significant modules ranked by mcode score for further study (Figure 4). A total of 527 nodes and 4661 edges were screened from the PPI network. According to the criteria above, eight nodes (PTPRC, ITGAM, LCP2, CTLA4, CD80, ITGAX, CD19, CCR5) were identified as hub genes. Consistently, these hub genes serve crucial roles in maintaining the top three modules.

Prognostic gene validate and intersection analysis with PPI network

To test whether these prognostic genes have prognostic significance in other LUAD cases. Microarray expression profile dataset GSE68465 from GEO database was downloaded. After using the Robust Multichip Averaging (RMA) method to normalize the sequence matrix data, the data were subjected to prognostic gene selection. There are 176 genes among 542 prognostic genes cannot find in GSE68465 dataset because of platform differences. Of the remaining 366 genes, 106 genes were confirmed involving in LUAD patient survival. By intersection analysis the 106 prognostic genes with hub genes of PPI network, only PTPRC and CD19 were existed both.

The correlation of PTPRC and CD19 expression with clinicopathological factors

Based on the study above, PTPRC and CD19 expression level were correlated with overall survival of LUAD patient, and PTPRC and CD19 high-expression group with longer survival. To explore the correlation of PTPRC and CD19 expression with clinical characteristic in LUAD

patient, we analysis the PTPRC and CD19 expression level with TNM stage. The result indicated that the PTPRC and CD19 expression were negative correlated with the TNM stage of LUAD patient, and with the TNM stage rising, the expression of PTPRC and CD19 declined(Figure 5).

PTPRC and CD19 could be an indicator in TME status

To further elaborate the role of PTPRC and CD19 in LUAD, we implement GSEA in the high and low expression group of PTPRC and CD19. The result display that, in the high expression level whether PTPRC or CD19, the genes were basically enriched in immune relative activities including allograft rejection, complement and inflammatory response. In the low expression level of PTPRC, the genes were basically enriched in glycolysis, and typical tumor pathway including MYC-targets-V1 and MYC-targets-V2. As to the low expression group of CD19, genes were gathered in metabolic pathway. In C7 collection, many genes were concentrated in the high expression group of no matter PTPRC nor CD19, and both low expression group enrich few gene (Figure 6).

Further,we performed correlation analysis between PTPRC and CD19 and 22 kind of immune infiltration cell. The result reveal that 16 kinds of immune infiltration cell were related with the expression of PTPRC(Figure 7), 15 kinds associated with CD19(Figure S 2).

Construction and validation of the prognostic model

Factor analysis was used to determine common axes (or dimensions) of patterns and structures, which were measured by a reduced set of 106 predicted genes. The first twelve factors explained about 81.808% of the variation in the 106 prognostic genes. After multivariate Cox regression analysis, five factors correlated with survival were obtained ($p<0.05$). Through the formula mentioned before, the final prognostic model combined with 106 candidate gene were constructed. Further, to test the predictive ability of the prognostic model, the microarray expression profile dataset GSE68465 was downloaded and we calculate the risk score for each patient. Based on the median risk score, we divided the patients into high- and low-risk group. We choose the log-rank test to identify the differences of OS between subgroups. The results showed that the overall survival of the high-risk patients is shorter than those in the low-risk ones in both TCGA dataset and GSE68465 ($p<0.0001$, $p=0.0009$ respectively, Figure 8A-B). Then, to estimate whether the prognostic model is predictive of relapse free survival (RFS). We divided the patients into high- and low-risk group compared with the median risk score, log-rank test indicate that the prognostic model we build could predict the RFS of LUAD. ROC curves were also applied to evaluate the sensitivity of the prognostic model,and the ROC curves showed that the AUC value of the prognostic model reached 0.649, 0.617 in TCGA dataset and GSE68465 respectively. The result indicates a substantially effective performance of the prognostic model for overall survival prediction(Figure 8C-D).

To optimize the model with clinical charateristics

Further, univariate and multivariate Cox regression analysis have been executed between some of clinical pathological parameters and risk score of the prognostic model (Table1). The result showed that the risk assessment model was an independent prognostic factor for prognostic. As the constructed risk assessment model with great prognostic value, we intended to improve the prognostic accuracy by intergrating with LUAD clinicopathological factors. We designed a nomogram to predict the survival of LUAD patient by combing T stage, lymph nodes metastasis, recurrence and risk score (Figure 9A). The AUC of the model was achieved 0.716 for 3 year and 0.699 for 5 year (Figure 9D-E). Figure 9C-D show the nomogram calibration plots for predicting the overall survival of 3 years and 5 years of patient.

Discussion

Lung-cancer-related deaths is quite a large scale in the world which could be the most in patients who suffer cancer. And LUAD, as the most common type of lung cancer, is able to account for more than half of the morbidity and mortality of the lung cancer patient.[25,26] Currently, growing evidence has suggested that TME plays a pivotal role in tumor initiation and progression,[27,28] especially the recent immunocheckpoint inhibitors have noticeable effects on lung adenocarcinoma. However, patient prognosis and disease progression involved with TME related genes in LUAD have not been elucidated clearly. In the present study, 1910 genes involved in immune response and cell adhesion were identified by comparing different immune/stromal scores of LUAD samples from the TCGA database. Besides, we performed survival analysis and revealed that 542 of them were associated with overall survival in LUAD patients. After performing cross validation through GSE68465, we obtained 106 TME related genes in which prominent correlation is found compared with prognosis situation. Of these survival-associated gene, 59 genes have been reported with the preliminary result. More importantly, among 106 TME related SurGenes, PTPRC and CD19 were highly interconnected nodes among PPI network and correlated with immune activity, exhibiting significant prognostic potential. Meanwhile, to explore the significance of these genes simultaneous changes for LUAD, we constructed and validated a risk assessment model that predicted survival of LUAD based on 106 genes.The prognostic model achieved robust predictive ability by combing with the clinical parameter we filtrated by multivariate Cox regression analysis.Recently, we notice that there are other prognostic model build by immune gene in LUAD[28], however, compared with their prognostic model, the prognostic model we build have some advantages. First of all, we constructed the model by factor analysis. Factor analysis is a kind of algorithms in biometrics. It represents a complex array of structure-analyzing procedures used to identify the interrelationships among a large set of observed variables and then, through data reduction, to

group a smaller set of these variables into dimensions or factors that have common characteristics. It is a tool to reduce multidimensional data to lower dimensions while retaining most of the information[29]. Secondly, our prognostic model was constructed by immune gene and stromal gene rather than only included immune gene, considering the effect of stromal gene on tumor microenvironment. Thirdly, before used to constructed the model, these 106 genes have been performed univariate cox regression analysis, and all of these gene were correlated with the prognosis of LUAD. At last, our prognostic model is also predictive not only to overall survival, but also to relapse free survival.

Among 106 TME related genes to be associated with LUAD, we are especially interested in PTPRC and CD19, as in the PPI network, they are highly interconnected nodes. And GSEA enrichment analysis showed the genes in PTPRC or CD19 high-expression group were mainly enriched in immune-related activities, such as allograft rejection, complement, and interferon response, and enriched in metabolic pathways in low expression of them, including glycolysis, oxidative phosphorylation, and typical tumor pathways. These result indicate that the expression level of PTPRC and CD19 might correlate with the status of immune and microenviroment in tumor. Further study analysised by CIBERSORT algorithm showed many tumor infiltrating immune cell were correlated with the expression of PTPRC and CD19. All of these result suggest PTPRC and CD19 participating in tumor microenvironment, and could be an indicator of TNM status.

Conclusions

In summary, after analysis of immune/stromal scores by using the ESTIMATE algorithm in TCGA database, a series of TME related genes was obtained and validated by an independent LUAD cohort. Further, the prognostic model combined with those identified genes was constructed and proved could provide a moderate OS prediction for LUAD. By integrating with clinical characteristics, testing by multivariate Cox regression analysis, the predictive ability have been optimized. Among 106 TME related genes, PTPRC and CD19 are highly interconnected nodes in the PPI network. Meanwhile, bioinformatic evidence suggest the levels of them affected the immune activity of TME. These two prognostic genes might be a candidate target to monitor and treat lung cancer, attributing to their immune nature and prognostic significance, thereby improving the clinical outcomes. Finally, our study might provide a novel insight on the potential correlation between monitoring and manipulating the TME with LUAD prognosis and precision immunotherapies.

Acknowledgements

Not applicable.

References

1. 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: 115-132.

2. Losanno T, Gridelli C. Safety profiles of first-line therapies for metastatic non-squamous non-small-cell lung cancer. Expert Opin Drug Saf. 2016; 15: 837-851.

3. Nakamura H, Saji H. Worldwide trend of increasing primary adenocarcinoma of the lung. Surg Today. 2014; 44: 1004-1012.

4. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65:5-29.

5. William WN Jr, Lin HY, Lee JJ, Lippman SM, Roth JA, Kim ES. Revisiting stage IIIB and IV non-small cell lung cancer: analysis of the surveillance, epidemiology, and end results data. Chest. 2009; 136:701-709.

6. Shaw AT, Yeap BY, Solomon BJ, Riely GJ, Gainor J, Engelman JA, Shapiro GI, Costa DB, Ou SH, Butaney M, Salgia R, Maki RG, Varella-Garcia M, Doebele RC, Bang YJ, Kulig K, Selaru P, Tang Y, Wilner KD, Kwak EL, Clark JW, Iafrate AJ, Camidge DR. Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. Lancet Oncol. 2011; 12:1004-1012.

7. Cui Y, Liu J, Yin HB, Liu YF, Liu JH. Fibulin-1 functions as a prognostic factor in lung adenocarcinoma. Jpn J Clin Oncol. 2015; 45:854–859.

8. Zhan J, Wang P, Niu M, Wang Y, Zhu X, Guo Y, Zhang H. High expression of transcriptional factor HoxB9 predicts poor prognosis in patients with lung adenocarcinoma. Histopathology. 2015; 66:955–965.

9. Kai Zhao, Zulei Li, Hui Tian. Twenty-gene-based prognostic model predicts lung adenocarcinoma survival. Onco Targets Ther. 2018; 11: 3415–3424.

10. Wood SL, Pernemalm M, Crosbie PA, Whetton AD. The role of the tumor-microenvironment in lung cancer-metastasis and its relationship to potential therapeutic targets. Cancer Treat Rev. 2014; 40:558-566.

11. Li B, Cui Y, Diehn M, Li R. Development and Validation of an Individualized Immune Prognostic Signature in Early-Stage Nonsquamous Non-Small Cell Lung Cancer. JAMA Oncol. 2017; 3:1529-1537.

12. Meyers DE, Bryan PM, Banerji S, Morris DG. Targeting the PD-1/PD-L1 axis for the treatment of non-small-cell lung cancer. Curr Oncol. 2018; 25:324-334.

13. Liu SY, Wu YL. Ongoing clinical trials of PD-1 and PD-L1 inhibitors for lung cancer in China. J Hematol Oncol. 2017; 10:136.

14. Mansuet-Lupo A, Alifano M, Pécuchet N, Biton J, Becht E, Goc J, Germain C, Ouakrim H, Régnard JF, Cremer I, Laurent-Puig P, Dieu-Nosjean MC, Blons H, Damotte D. Intratumoral Immune Cell Densities Are Associated with Lung Adenocarcinoma Gene Alterations. Am J Respir Crit Care Med. 2016; 194:1403-1412.

15. Driver BR, Ross A Miller, Tara Miller, Michael Deavers, Blythe Gorman, Dina Mody, Yimin Ge, Roberto Barrios, Eric Bernicker, Min Kim, Philip T Cagle. Programmed Death Ligand-1 (PD-L1) Expression in Either Tumor Cells or Tumor-Infiltrating Immune Cells Correlates with Solid and High-Grade Lung Adenocarcinomas. Arch Pathol Lab Med. 2017; 141:1529-1532.

16. Connolly KA, Nader M, Joshi N. Investigating Tumor-Associated Tertiary Lymphoid Structures in Murine Lung Adenocarcinoma. Methods Mol Biol. 2018; 1845:259-273.

17. Yoshihara K, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, Scott L Carter, Gad Getz, Katherine Stemke-Hale, Gordon B Mills, Roel G W Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013; 4:2612.

18. Shah N, Ping Wang, John Wongvipat, Wouter R Karthaus, Wassim Abida, Joshua Armenia, Shira Rockowitz, Yotam Drier, Bradley E Bernstein, Henry W Long, Matthew L Freedman, Vivek K Arora, Deyou Zheng, Charles L Sawyers. Regulation of the glucocorticoid receptor via a BET-dependent enhancer drives antiandrogen resistance in prostate cancer. Elife. 2017; 6:27861.

19. Priedigkeit N, Watters RJ, Lucas PC, Basudan A, Bhargava R, Horne W, Kolls JK, Fang Z, Rosenzweig MQ, Brufsky AM, Weiss KR, Oesterreich S, Lee AV. Exome-capture RNA sequencing of decade-old breast cancers and matched decalcified bone metastases. JCI Insight. 2017;2: 95703.

20. M Henar Alonso, Susanna Aussó, Adriana Lopez-Doriga, David Cordero, Elisabet Guinó, Xavier Solé, Mercè Barenys, Javier de Oca, Gabriel Capella, Ramón Salazar, Rebeca Sanz-Pamplona, Victor Moreno. Comprehensive analysis of copy number aberrations in microsatellite stable colon cancer in view of stromal component. Br J Cancer. 2017; 117:421-431.

21. Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43: e47.

22. Metsalu T, Vilo J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res.2015; 43: W566-70.

23. Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4:44–57.

24. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003; 4:2.

25. William D Travis, Elisabeth Brambilla, Masayuki Noguchi, Andrew G Nicholson, Kim Geisinger, Yasushi Yatabe, Charles A Powell, David Beer, Greg Riely, Kavita Garg, John H M Austin, Valerie W Rusch, Fred R Hirsch, James Jett, Pan-Chyr Yang, Michael Gould, American Thoracic Society. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society: international multidisciplinary classification of lung adenocarcinoma: executive summary. Proc Am Thorac Soc. 2011; 8:381-385.

26. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011; 61:69–90.

27. Liotta LA, Kohn EC. The microenvironment of the tumour-host interface. Nature. 2001; 411:375-9.

28. Lingyan Wang, Bijun Zhu, Miaomiao Zhang, Xiangdong Wang.

Roles of immune microenvironment heterogeneity in therapy-associated biomarkers in lung cancer. Semin Cell Dev Biol. 2017;64:90-97. doi: 10.1016

29. Barber DF, Faure M, Long EO. LFA-1 contributes an early signal for NK cell cytotoxicity. J Immunol. 2004; 173:3653-3659.

Table 1(on next page)

The univariate and multivariate Cox regression analysis of the patients from The Cancer Genome Atlas (TCGA)

Parameters	Univariate			Multivariate		
	HR	95%CI	P-value	HR	95%CI	P-value
Recurrence(yes VS no)	2.405	1.709-3.385	<0.001	2.449	1.738-3.450	<0.001
Age(≤65y VS >65y)	1.157	0.860-1.558	0.336			
Gender（male VS female）	1.047	0.782-1.400	0.759			
Risk score	2.216	1.635-3.004	<0.001	1.897	1.329-2.709	<0.001
T stage	2.455	1.697-3.553	<0.001	1.827	1.827-1.156	<0.01
N stage	2.546	1.900-3.411	<0.001	2.139	1.517-3.017	<0.001
M stage	1.027	0.747-1.414	0.868			
TNM stage(I-II VS III-IV)	2.686	1.973-3.658	<0.001	1.963	1.579-3.141	0.155
Number-pack-years-smoked	1.028	0.716-1.475	0.881			

Figure 1

The relationship between Estimate/immune/stromal score and the prognosis of LUAD samples from TCGA.

(A) Estimate score was significantly associated with pathological stage (p=0.0436). (B) Immune score showed predictive potential to pathological stage (p=0.026). (C) Similarly, the lowest stromal score was found in the most progressive clinicopathological stage IV, however, it was not statistically significant (p=0.145). (D) Immune score was significantly correlated with overall survival of LUAD samples. (p = 0.015). (E) Stromal score was not statistically significant correlated with overall, but the median overall survival of cases with higher stromal scores also showed longer than the patients with lower stromal scores (1830 d vs. 1293 d, p= 0.0599,p=0.145).

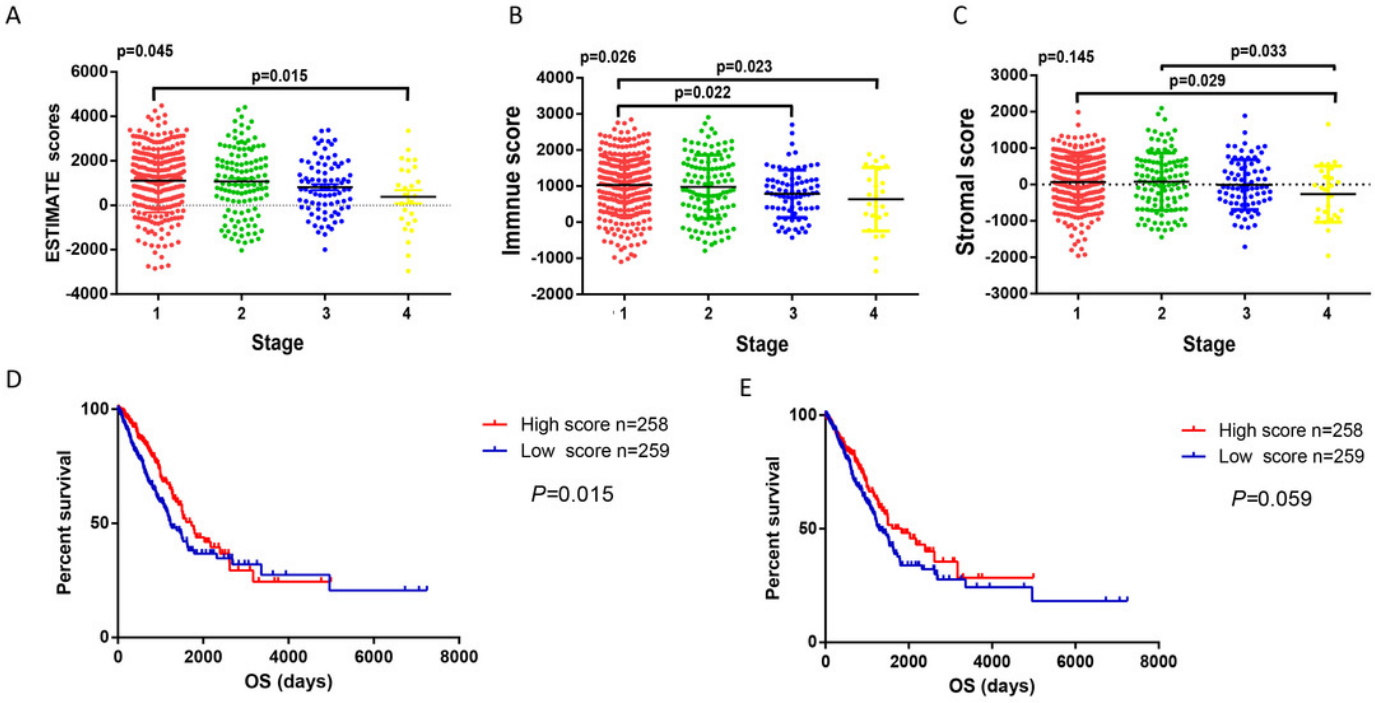


Figure 2

Differential expression genes derived from the comparision of immune/stromal high score group versus low score group.

(A-B) Heatmap of top 50 highly variaed gene from immune score /stromal score group.Genes with higher expression are shown in red, lower expression are shown in green, genes with same expression level are in black. (C-D) Venn diagram of common up /down regulated of immuneDEGs and stromal DEGs.

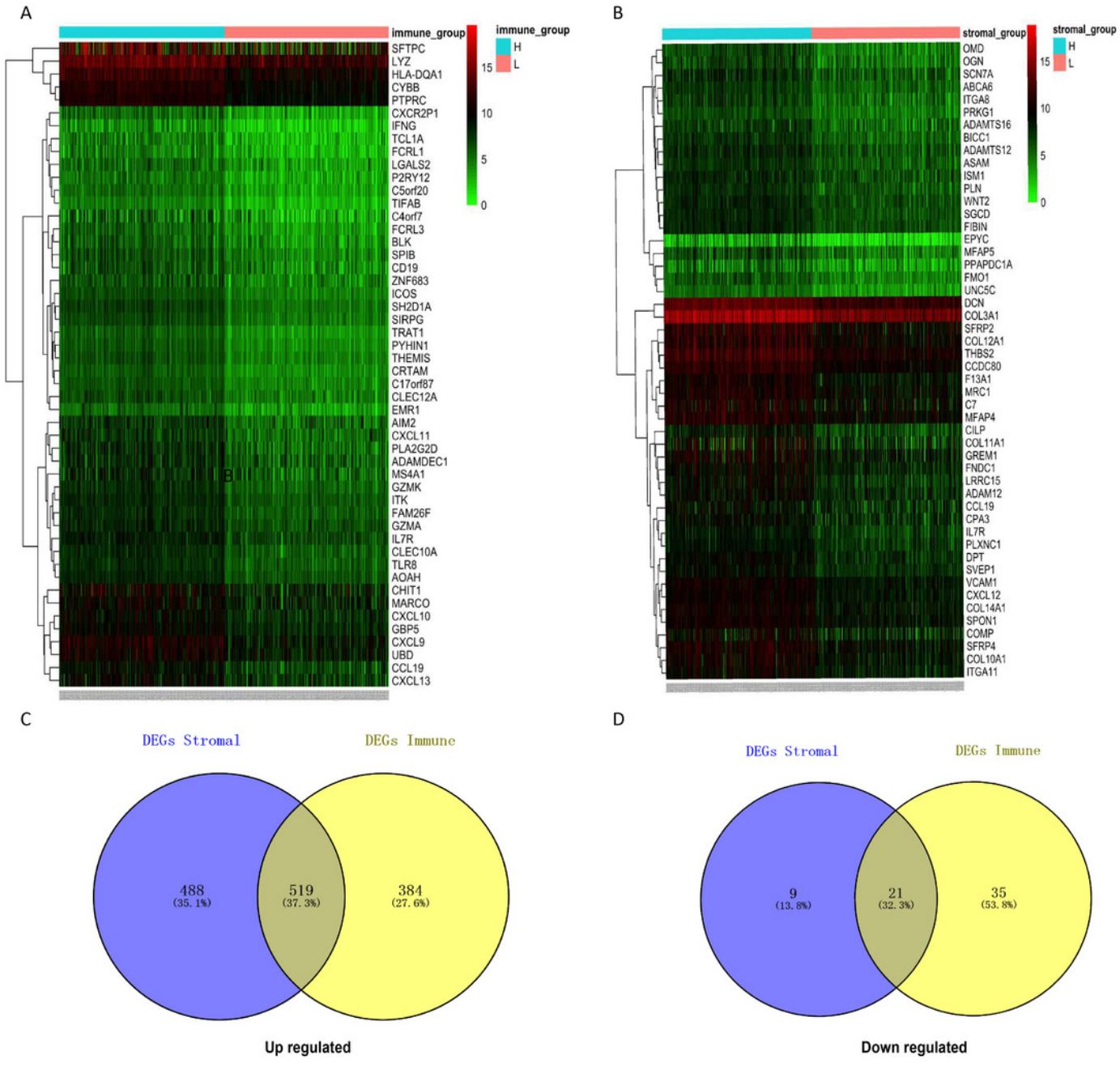


Figure 3

GO term and KEGG pathway analysis for immune up-regulated DEGs.

Top 10 pathways. False discovery rate (FDR) of GO analysis was acquired from DAVID functional annotation tool. $p < 0.05$. (A) biological process, (B) cellular component, (C) molecular function, (D) KEGG pathway.

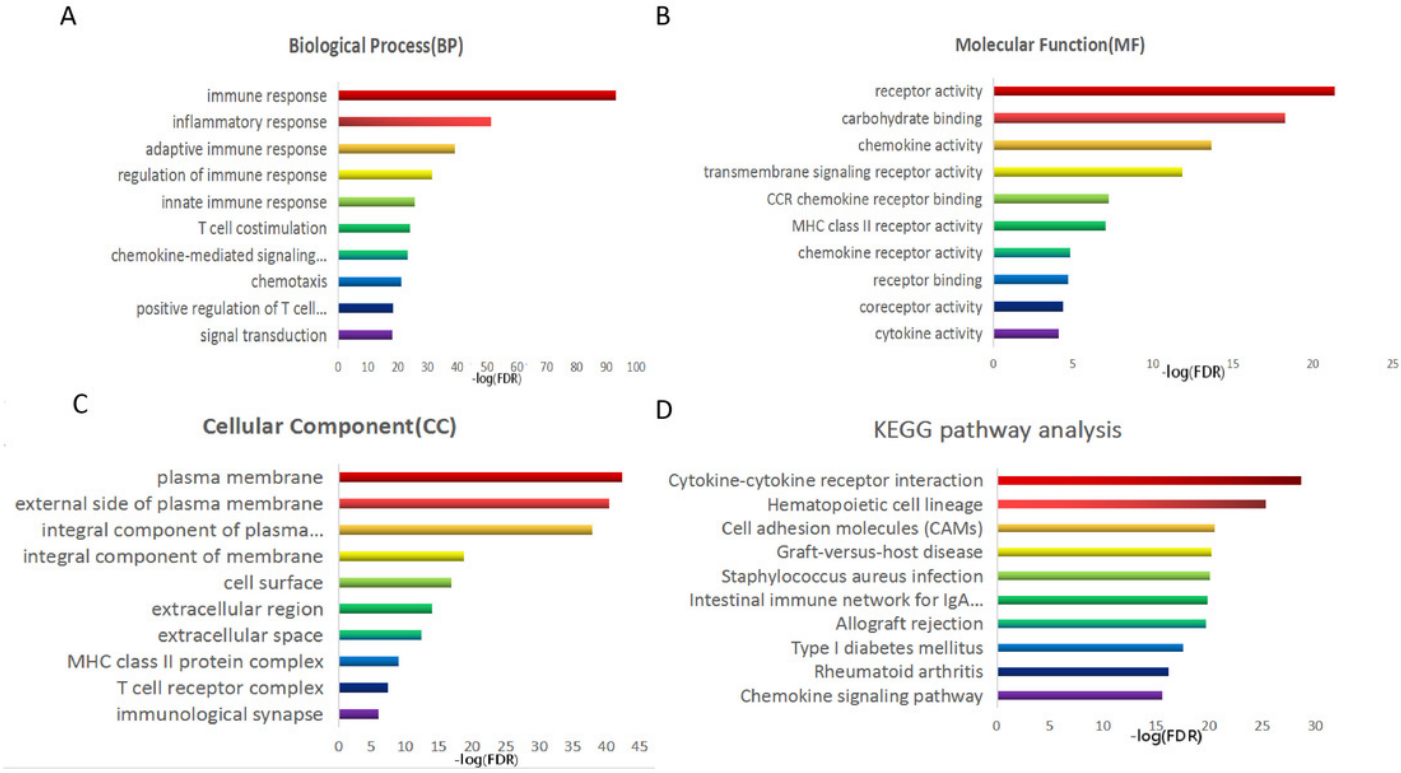


Figure 4

PPI network of top three modules, ranked by mcode score.

The color of a node in the PPI network reflects the number of interacting proteins with the designated protein, and the size of node indicates the log (FC) value of the Z score of gene expression.

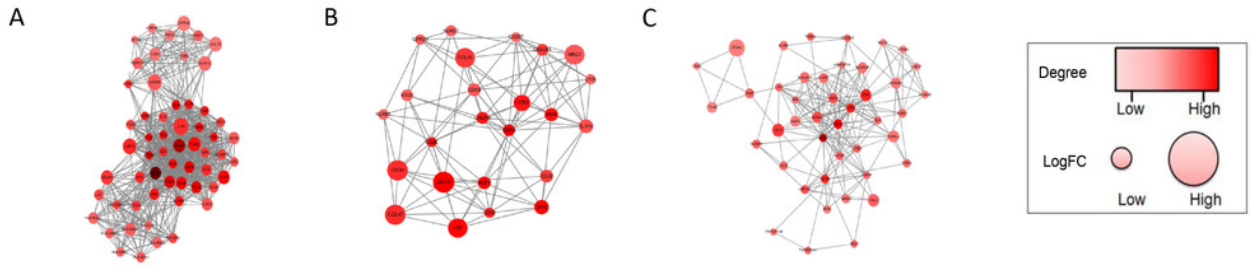


Figure 5

The correlation of PTPRC and CD19 with clinicopathological stage and survival of LUAD patients.

(A-B) The correlation of PTPRC and CD19 expression with clinicopathological stage. Wilcoxon rank sum test served as the statistical signifiificance test.

(C-D) Survival analysis for LUAD patients with different PTPRC and CD19 expression. Patients were labeled with high expression or low expression depending on the comparison with the median expression level. Log-rank test served as the statistical signifiificance test.

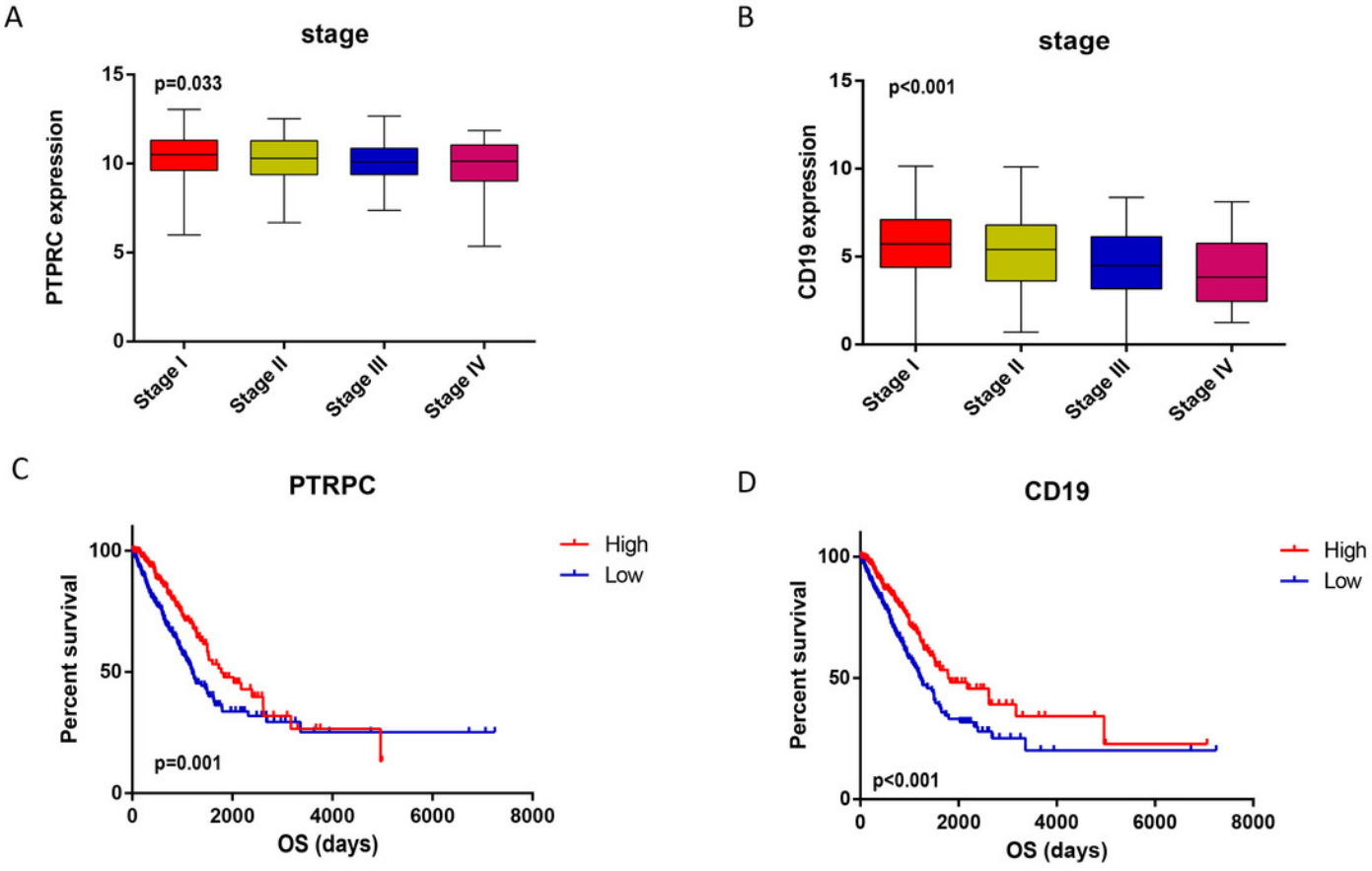


Figure 6

GSEA for samples with high PTPRC/CD19 expression and low expression.

(A-C) The enriched gene sets in HALLMARK collection by the high PTPRC expression sample. Up-regulated genes located in the left approaching the origin of the coordinates, by contrast the down-regulated lay on the right of x-axis. Only gene sets with NOM $p < 0.05$ and FDR $q < 0.25$ were considered signifificant. And only several leading gene sets were displayed in the plot. (D-F) The enriched gene sets in HALLMARK by samples with low PTPRC expression. (G-I) Enriched gene sets in C7 collection, the immunologic gene sets, by samples of high PTPRC expression. Only several leading gene sets are shown in plot. (J-K) Enriched gene sets in C7 by the low PTPRC expression. (L-N) The enriched gene sets in HALLMARK collection by the high CD19 expression sample. (O-P) The enriched gene sets in HALLMARK by samples with low CD19 expression. (Q-S) Enriched gene sets in C7 by the high CD19 expression.

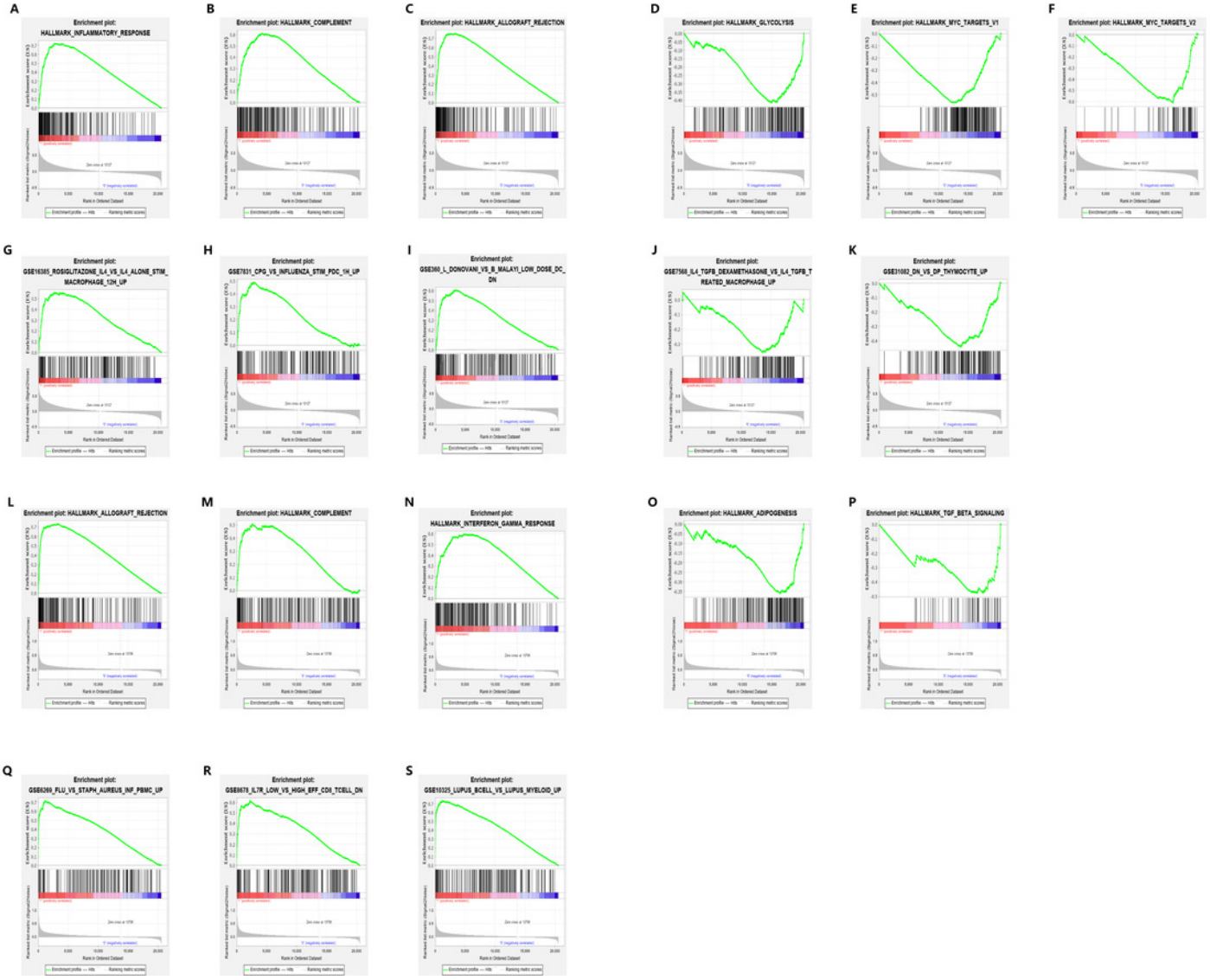

A
Enrichment plot:
HALLMARK_INFLAMMATORY_RESPONSE
B
Enrichment plot: HALLMARK_COMPLEMENT
C
Enrichment plot: HALLMARK_ALLOGRAFT_REJECTION
D
Enrichment plot: HALLMARK_GLYCOLYSIS
E
Enrichment plot: HALLMARK_MYC_TARGETS_V1
F
Enrichment plot: HALLMARK_MYC_TARGETS_V2
G
Enrichment plot:
GSE16385_ROSIGLITAZONE_IL4_VS_IL4_ALONE_STIM_MACROPHAGE_12H_UP
H
Enrichment plot:
GSE7831_CPG_VS_INFLUENZA_STIM_PDC_1H_UP
I
Enrichment plot:
GSE360_L_DONOVANI_VS_B_MALAYI_LOW_DOSE_DC_DN
J
Enrichment plot:
GSE7568_IL4_TGFB_DEXAMETHASONE_VS_IL4_TGFB_TREATED_MACROPHAGE_UP
K
Enrichment plot:
GSE13082_DN_VS_DP_THYMOCYTE_UP
L
Enrichment plot: HALLMARK_ALLOGRAFT_REJECTION
M
Enrichment plot: HALLMARK_COMPLEMENT
N
Enrichment plot:
HALLMARK_INTERFERON_GAMMA_RESPONSE
O
Enrichment plot: HALLMARK_ADIPOGENESIS
P
Enrichment plot: HALLMARK_TGF_BETA_SIGNALING
Q
Enrichment plot:
GSE6269_FLU_VS_STAPH_AUREUS_INF_PBMC_UP
R
Enrichment plot:
GSE8678_IL7R_LOW_VS_HIGH_EFF_CD8_TCELL_DN
S
Enrichment plot:
GSE10325_LUPUS_BCELL_VS_LUPUS_MYELOID_UP

Figure 7

TIC profifile in tumor samples and correlation of TICs proportion with PTPRC expression.

(A) Barplot showing the proportion of 22 kinds of TICs in LUAD tumor samples. Column names of plot were sample ID. (B) Heatmap showing the correlation between 10 kinds of TICs and numeric in each tiny box indicating the p value of correlation between two kinds of cells. The shade of each tiny color box represented corresponding correlation value between two cells, and Pearson coeffificient was used for signifificance test. (C-R) Scatter plot showed the correlation of 16 kinds of TICs proportion with the PTPRC expression ($p < 0.05$). The red line in each plot was fifitted linear model indicating the proportion tropism of the immune cell along with PTPRC expression, and Pearson coeffificient was used for the correlation test.

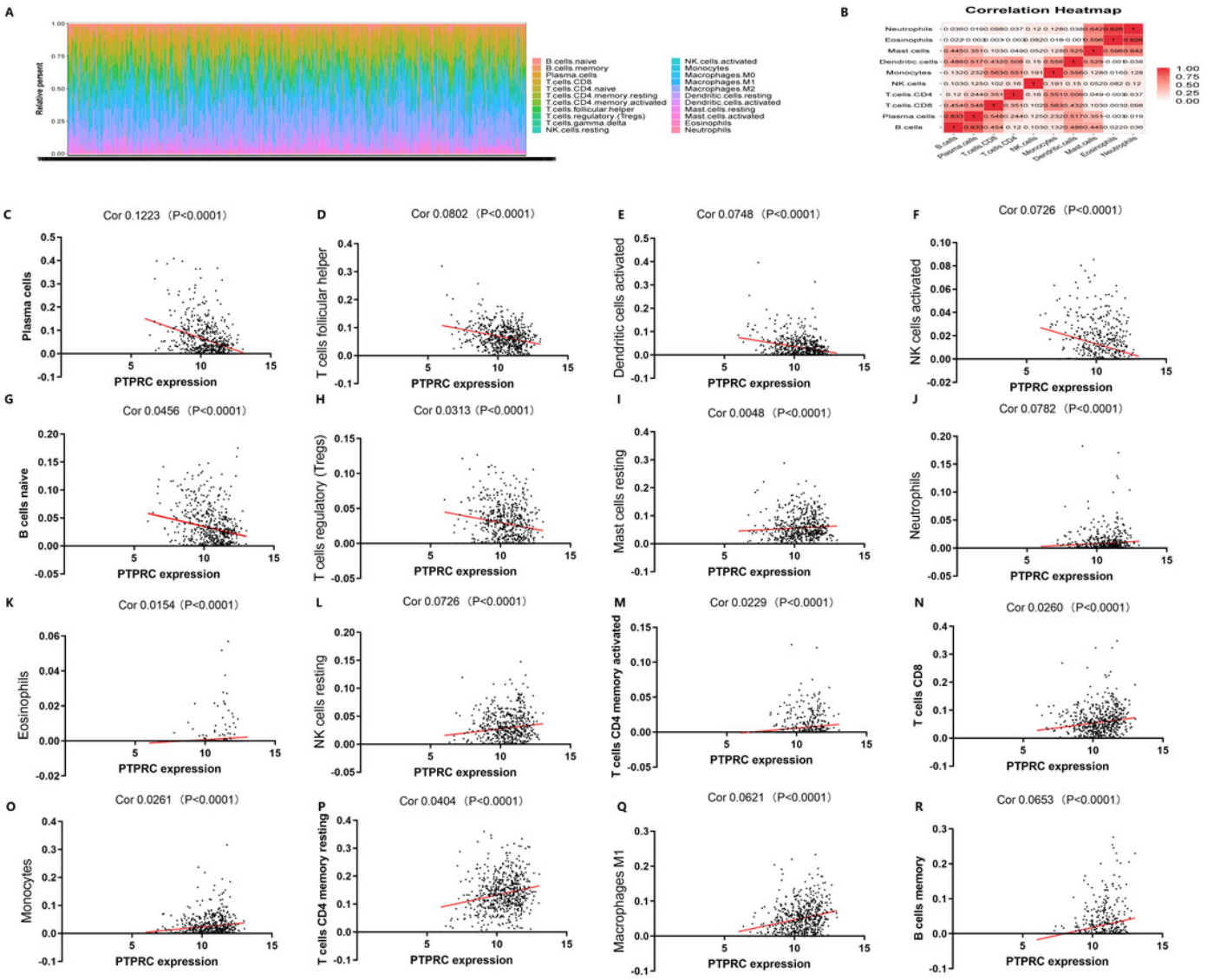

A
Relative percent
B.cells.naive
B.cells.memory
Plasma.cells
T.cells.CD8
T.cells.CD4.naive
T.cells.CD4.memory.resting
T.cells.CD4.memory.activated
T.cells.follicular.helper
T.cells.regulatory.(Tregs)
T.cells.gamma.delta
NK.cells.resting
NK.cells.activated
Monocytes
Macrophages.M0
Macrophages.M1
Macrophages.M2
Dendritic.cells.resting
Dendritic.cells.activated
Mast.cells.resting
Mast.cells.activated
Eosinophils
Neutrophils
B
Correlation Heatmap
C Cor 0.1223 (P<0.0001) Plasma cells
D Cor 0.0802 (P<0.0001) T cells follicular helper
E Cor 0.0748 (P<0.0001) Dendritic cells activated
F Cor 0.0726 (P<0.0001) NK cells activated
G Cor 0.0456 (P<0.0001) B cells naive
H Cor 0.0313 (P<0.0001) T cells regulatory (Tregs)
I Cor 0.0048 (P<0.0001) Mast cells resting
J Cor 0.0782 (P<0.0001) Neutrophils
K Cor 0.0154 (P<0.0001) Eosinophils
L Cor 0.0726 (P<0.0001) NK cells resting
M Cor 0.0229 (P<0.0001) T cells CD4 memory activated
N Cor 0.0260 (P<0.0001) T cells CD8
O Cor 0.0261 (P<0.0001) Monocytes
P Cor 0.0404 (P<0.0001) T cells CD4 memory resting
Q Cor 0.0621 (P<0.0001) Macrophages M1
R Cor 0.0653 (P<0.0001) B cells memory
PTPRC expression

Figure 8

Validation of the prognostic model.

(A) The Kaplan-Meier survival analysis of the prognostic model for LUAD samples from TCGA. (B) ROC curve of the prognostic model in LUAD samples from TCGA. (C) The Kaplan-Meier survival analysis of the prognostic model for LUAD samples from GSE68465.(D) ROC curve of the prognostic model in LUAD samples from gse68465. (E) The Kaplan-Meier survival analysis of the prognostic model with relapse free survival for LUAD samples from TCGA.

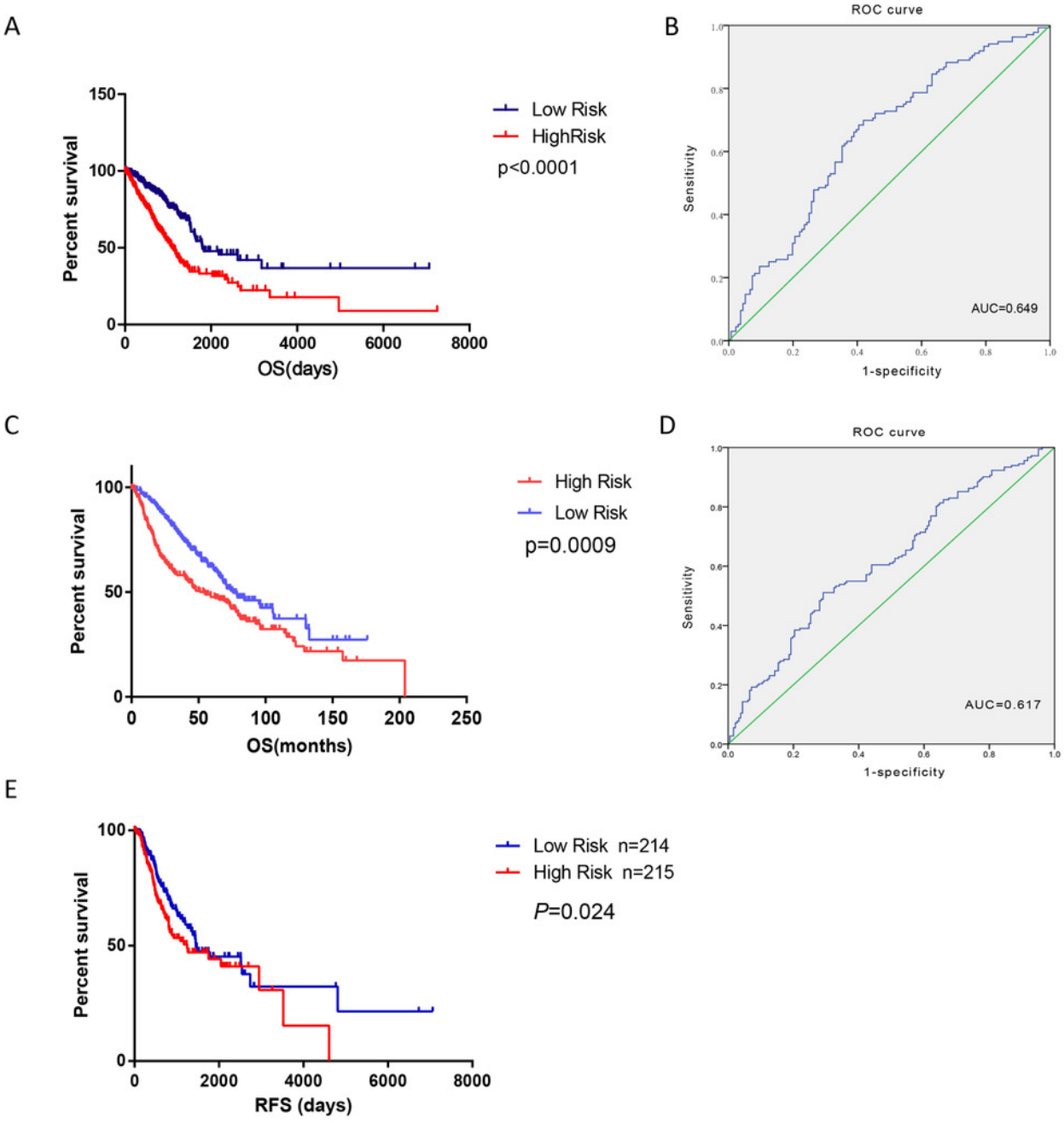

A
Percent survival
150
100
50
0
0
2000
4000
6000
8000
OS(days)
Low Risk
HighRisk
p<0.0001
B
ROC curve
Sensitivity
1-specificity
AUC=0.649
C
Percent survival
100
50
0
0
50
100
150
200
250
OS(months)
High Risk
Low Risk
p=0.0009
D
ROC curve
Sensitivity
1-specificity
AUC=0.617
E
Percent survival
100
50
0
0
2000
4000
6000
8000
RFS (days)
Low Risk n=214
High Risk n=215
P=0.024

Figure 9

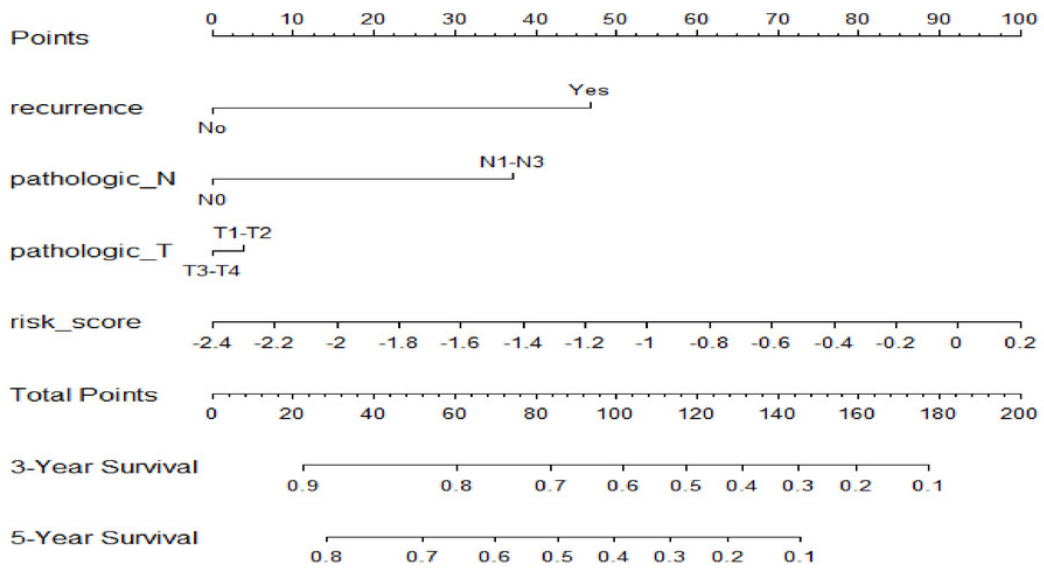
Nomogram and Calibration curve

A) Construction of a nomogram for predicting survival probability at 3, 5, and year of LUAD cases from TCGA data set.

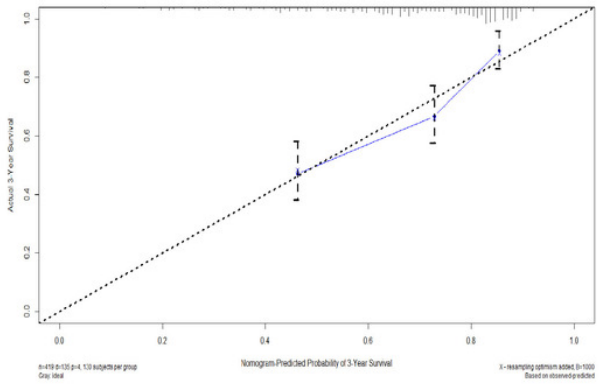
Calibration curve for the nomogram when predicting 3 (B) and 5 (C) year OS.

(D-F) ROC curve of the optimized model in 3-years and 5 years.

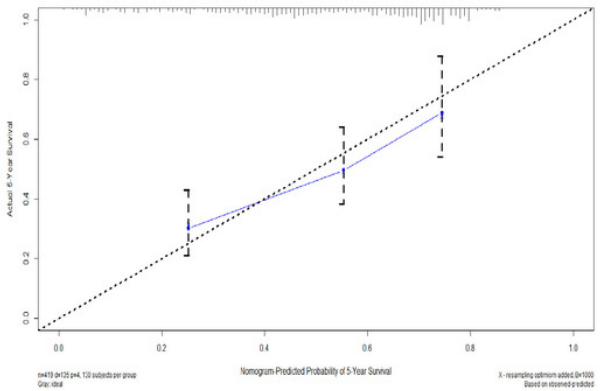
A


Points
0 10 20 30 40 50 60 70 80 90 100
recurrence
Yes
No
pathologic_N
N1-N3
N0
pathologic_T
T1-T2
T3-T4
risk_score
-2.4 -2.2 -2 -1.8 -1.6 -1.4 -1.2 -1 -0.8 -0.6 -0.4 -0.2 0 0.2
Total Points
0 20 40 60 80 100 120 140 160 180 200
3-Year Survival
0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1
5-Year Survival
0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1

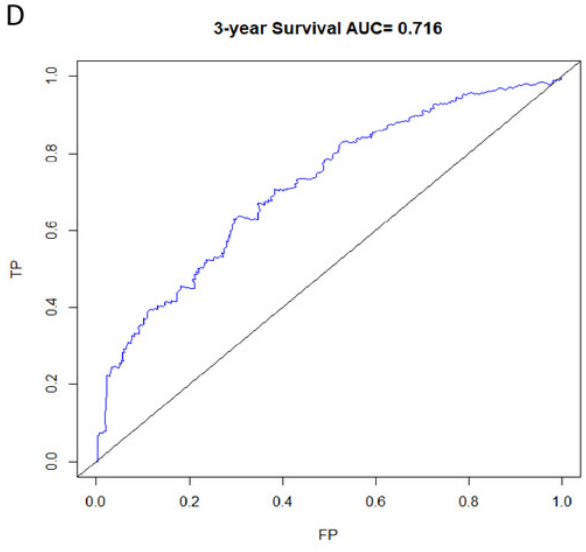
B


Actual 3-Year Survival
Nomogram-Predicted Probability of 3-Year Survival

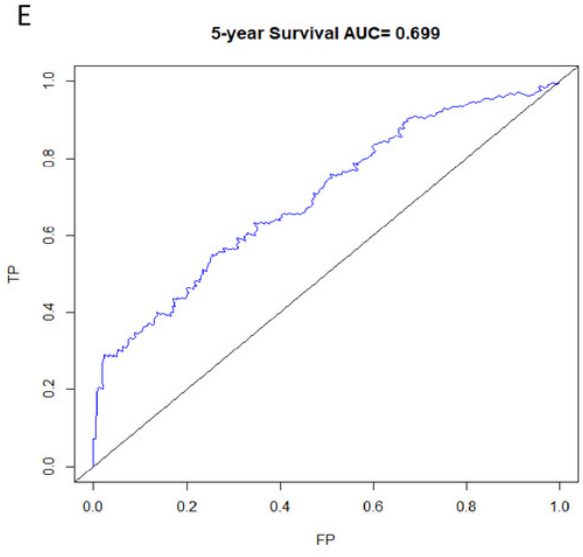
C


Actual 5-Year Survival
Nomogram-Predicted Probability of 5-Year Survival

D


3-year Survival AUC= 0.716
TP
FP

E


5-year Survival AUC= 0.699
TP
FP