A prognostic score model based on eight metabolism-associated genes predicts the survival of adult females with lung adenocarcinoma

Hongxia Wang
University of Hong Kong-Shenzhen Hospital https://orcid.org/0000-0001-8779-1201
Guangqiang Shao (shaogq@hku-szh.org)
The University of Hong Kong Shenzhen Hospital
Lei Rong
University of Hong Kong-Shenzhen Hospital
Yang Ji
University of Hong Kong-Shenzhen Hospital
Keke Zhang
University of Hong Kong-Shenzhen Hospital
Min Liu
University of Hong Kong-Shenzhen Hospital

Research

Keywords: lung adenocarcinoma, differential expression analysis, metabolism, enrichment analysis, prognostic score model, nomogram survival model

Posted Date: February 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-164370/v1

Abstract

Background

As a non-small cell lung cancer, lung adenocarcinoma (LUAD) is common in women and non-smokers. This study is aimed to construct a prognostic score (PS) model for adult females with LUAD.

Methods

The gene expression data of adult females with LUAD from The Cancer Genome Atlas database was obtained as the training set, and GSE50081 and GSE37745 from Gene Expression Omnibus database were downloaded as the validation sets. The differentially expressed genes (DEGs) between LUAD and normal samples were screened by limma package. The metabolism-associated DEGs were selected by Gene Set Enrichment Analysis, and were conducted with enrichment analysis using DAVID tool. After the independent prognosis-associated genes were identified by survival package, the optimal gene combination was screened using penalized package to build the PS model. Besides, the nomogram survival model based on the independent prognostic clinical factors was constructed by rms package. Using HPAanalyze package, the protein expression levels of the optimal genes were mined.

Results

There were 2388 DEGs between LUAD samples and normal samples. Totally, 150 metabolism-associated DEGs were screened, for which PPAR signaling pathway was enriched. The optimal gene combination (involving *CYP17A1*, *ASPG*, *DUOX1*, *CIDEC*, *TH*, *B4GALNT1*, *APOA2*, and *GCKR*) was selected, based on which the PS model was built. Combined with pathologic stage and PS model status, the nomogram survival model was constructed. Moreover, *CIDEC* was a characteristic gene in lung cancer and other cancers.

Conclusion

The PS model and the nomogram survival model might be applied for the prognostic prediction of adult females with LUAD.

Introduction

Lung adenocarcinoma (LUAD) belongs to non-small cell lung cancer (NSCLC) and makes up 40% of lung cancers (1). LUAD has no special symptoms in early stage, which only presents as common respiratory symptoms (such as cough, low fever, phlegm blood, chest pain, and chest tightness (2). Majority of LUADs originate from the mucosal epithelium of the small bronchioles, and a few of LUADs derives from the mucous glands of the large bronchioles (3). LUAD is more common in women and non-smokers, which has a lot to do with air pollution, lampblack, and working environment (4). LUAD generally grows slowly, but it sometimes occurs hematogenous metastasis or lymphatic metastasis (5, 6). Therefore, exploring the biomolecules significantly associated with LUAD in women is important for improving the outcomes of LUAD patients.

Some genes implicated in the prognosis of LUAD patients have been reported by several researches in recent years. For example, glucose-6-phosphate dehydrogenase (*G6PD*) expression has correlations with lymph node metastasis, invasion, higher TNM stage, poorer differentiation, and adverse outcome of LUAD, and thus *G6PD* may be a poor prognostic marker for the disease (7). Increased collagen type V alpha 1 (*COL5A1*) is related to LUAD metastasis, and the overexpression of *COL5A1* is detected in LUAD patients with recurrence and poor prognosis (8). Kinesin family member 18A (*KIF18A*) overexpression has independent correlation with adverse recurrence-free survival and overall survival (OS) of LUAD patients, which is a promising prognostic marker and therapeutic target for the tumor (9, 10). Up-regulated S100 calcium binding protein A14 (*S100A14*) promotes the invasion and migration of LUAD cells, and may be applied for the diagnosis, prognostic prediction, and treatment of LUAD (11). However, the prognostic mechanisms of LUAD still needed to be studied deeply.

Several gene signatures for LUAD have been built to predict the outcomes of LUAD patients. Yun-Yong et al find a 193-gene gene expression signature that can independently predict the OS of LUAD patients, helping to identify the high risk patients in stage I and the patients suitable for adjuvant chemotherapy (12). Wang et al develop a four-gene signature for the LUAD patients with lymph node metastasis, contributing to recognizing the high risk patients and improving their treatment and clinical outcomes (13). Besides, oncogenic signalling pathways have influences on metabolic processes, and glucose, lipid, and protein metabolisms play roles in tumor cell growth and survival (14, 15). Nevertheless, the risk model based on metabolism-associated genes has not been constructed for LUAD patients.

In this study, the gene expression data of LUAD samples from adult females were extracted from The Cancer Genome Atlas (TCGA) database. The differentially expressed genes (DEGs) were compared with the genes correlated with metabolism of amino acids and derivatives, metabolism of carbohydrates, and metabolism of lipids to identify the metabolism-associated DEGs. From the independent prognosis-associated genes screened from the metabolism-associated DEGs, the optimal genes were further selected for constructing the prognostic score (PS) model. Furthermore, the protein expression levels of the optimal genes in different tumors were analyzed. The PS model constructed in this study might be valuable for predicting the prognosis of LUAD patients.

Materials And Methods

Data downloading and data preprocessing

From TCGA (https://cancergenome.nih.gov/) database, the gene expression data of LUAD (platform: Illumina HiSeq 2000 RNA Sequencing; including 585 samples; downloaded in March 10, 2020) was extracted. According to the clinical information of the samples, 300 samples from females older than 18 years (including 266 LUAD samples and 34 normal samples) were included in the training set.

Using “Homo sapiens” and “lung cancer” as the searching words, eligible datasets satisfying the following criteria were searched from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database: (1) the samples were provided with histology type information; (2) the total sample size was equal to or larger than 150, and the LUAD samples was no less than 100; (3) LUAD samples had age and gender information; (4) LUAD samples had survival prognosis information, and the valid samples should be no less than 50. After searching, GSE50081 (platform: GPL570 Affymetrix Human Genome U133 Plus 2.0 Array; including 62 LUAD samples from adult females with survival prognosis information; the validation set 1) and GSE37745 (platform: GPL570

Affymetrix Human Genome U133 Plus 2.0 Array; including 60 LUAD samples from adult females with survival prognosis information; the validation set 2) were selected as the eligible datasets.

Differential expression analysis

Based on the limma package (16) (https://bioconductor.org/packages/release/bioc/html/limma.html, version 3.34.7) in R, the DEGs between the LUAD samples and normal samples in the training set were analyzed. The false discovery rate (FDR) < 0.05 and $|\log_2$ fold change (FC)| > 1 were defined as the thresholds for selecting the DEGs. Combined with the expression of the DEGs in the training set, bidirectional hierarchical clustering was conducted by the pheatmap package (17) (https://cran.r-project.org/web/packages/pheatmap/index.html, version 1.0.8) in R.

Screening of metabolism-associated DEGs

From Gene Set Enrichment Analysis (GSEA, http://software.broadinstitute.org/gsea/downloads.jsp) database (18), the genes correlated with metabolism of amino acids and derivatives, metabolism of carbohydrates, and metabolism of lipids were downloaded. The downloaded genes were compared with the identified DEGs, and the intersected genes were taken as metabolism-associated DEGs. For the metabolism-associated DEGs, Gene Ontology (GO)_biological process (BP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed using DAVID tool (19) (version 6.8, https://david.ncifcrf.gov/). The threshold of enrichment analysis was the p-value < 0.05.

Construction of PS model

Using the univariate and multivariate Cox regression analyses in the R package survival (20) (version 2.41-1, http://bioconductor.org/packages/survivalr/), the DEGs significantly correlated with OS and independent prognosis were selected from the metabolism-associated DEGs. The log-rank p-value < 0.05 was the threshold.

Combined with the independent prognosis-associated genes, the optimal gene combination was screened using the LASSO Cox regression model in the R package penalized (21)(version 0.9.50, https://cran.r-project.org/web/packages/penalized/index.html). The optimized parameter "lambda" in the model was calculated by 1000 cross-validation likelihood (cvl). Based on the prognostic coefficients and expression levels of the optimal genes in the training set, the following PS model was constructed:

Prognostic score (PS) = $\sum \beta_{DEGs} \times \text{Exp}_{DEGs}$

β_{DEGs} and Exp_{DEGs} separately represent the prognostic coefficients and expression levels of the optimal genes.

The PSs of the samples in the training set were calculated, and then their median was applied for dividing the samples into high (PS ≥ median) and low (PS < median) risk groups. Using the Kaplan-Meier (KM) curve method in survival package (20), the correlation between the actual survival prognosis and the risk grouping was evaluated. Meanwhile, the expression levels of the optimal genes were extracted from the validation sets, and the PSs of the samples in the validation sets were calculated. Subsequently, the samples in the validation sets were classified into high and low risk groups, and the correlation between the actual survival and the grouping status was assessed.

Establishment of nomogram survival model

From the training set, the independent clinical prognostic factors were screened using the univariate and multivariate Cox regression analyses in the survival package (20). The log-rank p-value < 0.05 was defined for

screening the independent clinical prognostic factors. To reveal the correlation between the risk grouping and the independent clinical prognostic factors, the independent clinical prognostic factors were performed with stratified analysis. The samples were classified into different groups based on clinical factors, and the correlation analysis of PS model was carried out in these different groups. Combined with independent prognostic clinical factors and the PS model, nomogram survival model was established by the rms package (22) (version 5.1-2, https://cran.r-project.org/web/packages/rms/index.html) in R.

Searching of the protein expression levels of the optimal genes

The Human Protein Atlas (HPA) database (23) is developed for providing the tissue and cell distribution information of all 24,000 human proteins and examining the distribution and expression of each protein in 48 normal human tissues, 20 tumor tissues, 47 cell lines and 12 blood cells. Using the HPAanalyze package (24) (version 1.4.3, http://www.bioconductor.org/packages/release/bioc/html/HPAanalyze.html) in R, “tissue atlas” and “pathology atlas” were mined for the optimal genes.

Results

Differential expression analysis

There were a total of 2388 DEGs (including 1574 up-regulated genes and 814 down-regulated genes) between the LUAD samples and normal samples in the training set (Fig. 1A). Besides, hierarchical clustering heatmap was draw based on the expression of the DEGs, indicating that the samples were clearly clustered in two directions (Fig. 1B).

Screening of metabolism-associated DEGs

Based on GSEA database, 372 genes correlated with metabolism of amino acids and derivatives, 293 genes correlated with metabolism of carbohydrates, and 738 genes correlated with metabolism of lipids were downloaded. Through comparing the downloaded genes and the DEGs, 150 intersected genes were obtained as metabolism-associated DEGs (Fig. 2). According to the results of enrichment analysis, the metabolism-associated DEGs were involved in 17 GO_BP functional terms (such as oxidation-reduction process, p-value = 1.270E-19; cellular amino acid biosynthetic process, p-value = 8.430E-08; and hormone biosynthetic process, p-value = 2.710E-06) and 20 KEGG pathways (such as Metabolic pathways, p-value = 3.630E-32; Biosynthesis of amino acids, p-value = 1.050E-08; PPAR signaling pathway, p-value = 6.420E-08) (Table 1).

Table 1
The biological processes and KEGG pathways enriched for the metabolism-associated DEGs.

Type	Term	Count	P-value	FDR
Biology Process	GO:0055114 ~ oxidation-reduction process	36	1.270E-19	1.040E-16
	GO:0008652 ~ cellular amino acid biosynthetic process	7	8.430E-08	3.450E-05
	GO:0042446 ~ hormone biosynthetic process	5	2.710E-06	5.550E-04
	GO:0006590 ~ thyroid hormone generation	5	2.710E-06	5.550E-04
	GO:0006656 ~ phosphatidylcholine biosynthetic process	6	2.290E-06	6.240E-04
	GO:0019433 ~ triglyceride catabolic process	6	2.290E-06	6.240E-04
	GO:0019372 ~ lipoxygenase pathway	5	3.890E-06	6.370E-04
	GO:0030203 ~ glycosaminoglycan metabolic process	6	4.980E-06	6.780E-04
	GO:0016125 ~ sterol metabolic process	5	3.090E-05	3.600E-03
	GO:0006486 ~ protein glycosylation	8	6.350E-05	5.179E-03
	GO:0019369 ~ arachidonic acid metabolic process	5	6.350E-05	5.751E-03
	GO:0030148 ~ sphingolipid biosynthetic process	6	6.270E-05	6.393E-03
	GO:0015721 ~ bile acid and bile salt transport	5	8.680E-05	6.437E-03
	GO:0010043 ~ response to zinc ion	5	2.740E-04	1.710E-02
	GO:0008202 ~ steroid metabolic process	5	5.470E-04	2.761E-02
	GO:0016042 ~ lipid catabolic process	6	9.260E-04	3.716E-02
	GO:0005975 ~ carbohydrate metabolic process	8	8.970E-04	3.789E-02
KEGG Pathway	hsa01100:Metabolic pathways	75	3.630E-32	4.610E-30

Note: GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes; FDR, false discovery rate.

Type	Term	Count	P-value	FDR
	hsa01230:Biosynthesis of amino acids	12	1.050E-08	6.670E-07
	hsa00590:Arachidonic acid metabolism	11	2.520E-08	1.070E-06
	hsa03320:PPAR signaling pathway	11	6.420E-08	2.040E-06
	hsa00591:Linoleic acid metabolism	7	4.370E-06	1.110E-04
	hsa00564:Glycerophospholipid metabolism	9	1.040E-04	2.208E-03
	hsa00220:Arginine biosynthesis	5	2.210E-04	4.007E-03
	hsa00380:Tryptophan metabolism	6	3.540E-04	4.076E-03
	hsa00140:Steroid hormone biosynthesis	7	2.580E-04	4.093E-03
	hsa00340:Histidine metabolism	5	3.260E-04	4.135E-03
	hsa00260:Glycine, serine and threonine metabolism	6	3.130E-04	4.413E-03
	hsa04975:Fat digestion and absorption	6	3.130E-04	4.413E-03
	hsa00565:Ether lipid metabolism	6	6.180E-04	6.516E-03
	hsa00330:Arginine and proline metabolism	6	1.007E-03	9.098E-03
	hsa01130:Biosynthesis of antibiotics	11	1.618E-03	1.361E-02
	hsa01200:Carbon metabolism	8	1.808E-03	1.426E-02
	hsa00250:Alanine, aspartate and glutamate metabolism	5	1.999E-03	1.484E-02
	hsa00350:Tyrosine metabolism	5	1.999E-03	1.484E-02
	hsa04976:Bile secretion	6	4.231E-03	2.947E-02
	hsa04913:Ovarian steroidogenesis	5	6.851E-03	4.271E-02

Note: GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes; FDR, false discovery rate.

Construction of PS model

Combined with univariate Cox regression analysis, 32 DEGs significantly correlated with the OS of the 266 LUAD samples in the training set were screened from the metabolism-associated DEGs. Afterwards, nine genes correlated with independent prognosis were further selected by multivariate Cox regression analysis. Based on the independent prognosis-associated genes, the optimal gene combination involving eight genes (cytochrome P450 family 17 subfamily A member 1, *CYP17A1*; asparaginase, *ASPG*; dual oxidase 1, *DUOX1*; cell death inducing DFFA like effector c, *CIDEC*; tyrosine hydroxylase, *TH*; beta-1,4-N-acetyl-galactosaminyl transferase 1, *B4GALNT1*; apolipoprotein A2, *APOA2*; glucokinase regulator, *GCKR*) was identified (Table 2). Based on the prognostic coefficients of the optimal genes and their expression levels in the training set, the PS model was built.

Table 2
The eight genes involved in the optimal gene combination.

Symbol	Multi-variate Cox regression analysis			LASSO coefficient
	HR	95%CI	P-value	
CYP17A1	0.2250	0.056–0.912	3.664E-02	-1.01474
ASPG	0.5120	0.248–0.954	4.693E-02	-0.50098
DUOX1	0.9630	0.758–0.994	4.759E-02	-0.06046
CIDEC	1.4240	1.086–1.866	1.054E-02	0.32862
TH	3.5030	1.665–7.371	9.590E-04	1.05176
B4GALNT1	1.2120	1.022–1.593	1.683E-02	0.17280
APOA2	2.4390	1.536–3.875	1.580E-04	0.81560
GCKR	1.5440	1.057–2.490	4.748E-02	0.36242
Note: HR, hazard ratio; CI, confidence interval.				

After the median of the PSs of the samples in the training set were calculated, the samples were classified into high risk and low risk groups. Similarly, the samples in the validation sets were divided into high and low risk groups. Subsequently, the KM curves were draw to assess the correlation between the actual survival prognosis and the risk grouping. In the training set (p = 8.442E-06; hazard ratio (HR): 2.635 (1.695–4.097); area under the receiver operating characteristic curve (AUC): 0.904) and the validation sets GSE50081 (p = 6.252E-03; HR: 2.312 (1.929–5.751); AUC: 0.773) and GSE37745 (p = 1.354E-02; HR: 1.755 (1.117–2.758); AUC: 0.904), there were significant correlations between the different risk groups divided by the PS model and the actual prognosis (Fig. 3).

Establishment of nomogram survival model

Based on the univariate and multivariate Cox regression analyses, two independent clinical prognostic factors (pathologic stage and PS model status) were screened from the training set (Table 3). The KM curves for pathologic stage showed that the LUAD patients with lower pathologic stage had better prognosis (p = 1.628e-08, HR: 1.784 (1.448–2.198)) (Fig. 4A). The samples in the training set were divided according to their pathologic stage, and then the correlation between the prediction results of PS model and actual prognosis was analyzed in the samples in stage I (p = 3.806e-05, HR: 4.016 (1.973–8.172)), stage II (p = 2.569e-01, HR: 1.692 (0.675–4.242)),

stage III (p = 5.507e-01, HR: 1.294 (0.554–3.020)), and stage IV (p = 7.495e-01, HR: 1.411 (0.169–11.80)) (Fig. 4B). Moreover, the samples were divided into high and low risk groups based on PS model, and KM curve analysis of pathologic stage was conducted in the low risk (p = 3.076e-06, HR: 2.401 (1.615–3.567)) (Fig. 5A) and high risk (p = 2.062e-03, HR: 1.460 (1.143–1.867)) (Fig. 5B) groups.

Table 3
Univariate and multivariate Cox regression analyses for selecting the independent clinical prognostic factors.

Clinical characteristics	TCGA (N = 266)	Uni-variable cox		Multi-variable cox	
		HR (95% CI)	P-value	HR (95% CI)	P-value
Age(years, mean ± sd)	65.14 ± 10.26	1.012(0.992–1.032)	2.466E-01	-	-
Pathologic M(M0/M1/-)	167/11/88	4.722(2.201–10.13)	1.101E-01	-	-
Pathologic N(N0/N1/N2/-)	178/41/38/9	1.650(1.289–2.112)	4.339E-05	1.116(0.756–1.648)	5.817E-01
Pathologic T(T1/T2/T3/T4/-)	108/128/19/9/2	1.448(1.096–1.913)	9.133E-03	1.616(0.876–1.538)	2.984E-01
Pathologic stage(I / II / III / IV /-)	156/52/43/12/3	1.784(1.448–2.198)	1.628E-08	1.430(1.192–2.064)	5.560E-03
Tobacco history(Reformed/Current/Never/-)	69/21/18/158	0.980(0.658–1.461)	9.219E-01	-	-
PS model status(High/Low)	133/133	2.635(1.695–4.097)	8.442E-06	2.341(1.489–3.696)	2.340E-04
Vital status(Dead/Alive)	94/172	-	-	-	-
Overall survival free survival time(months,mean ± sd)	29.28 ± 26.42	-	-	-	-

Note: TCGA, The Cancer Genome Atlas; HR, hazard ratio; CI, confidence interval; PS, prognostic score.

In the training set, the nomogram survival model was established to analyze the correlations between the independent clinical prognostic factors and survival prognosis (Fig. 6A). The "Total points" axis of the nomogram was used to predict the survival of the samples by integrating pathologic stage and PS model status. Through comparing the survival probabilities predicted by the nomogram with the actual survival probabilities, the C-indexes for three-year survival probability and five-year survival probability separately were found to be 0.731 and 0.702 (Fig. 6B).

Searching of the protein expression levels of the optimal genes

Combined with HPA database, the protein expression levels of the optimal genes in different tissues, different cell components, and different kinds of tumors were analyzed. Since the subjects of this study were adult females, several tumors with female characteristics (including lung cancer, breast cancer, cervical cancer and ovarian cancer) were selected. According to the results, *CIDEC* was found to be a characteristic gene in lung tissue. In addition, the expression of *CIDEC* was also characteristic in other cancers, especially in breast cancer (Fig. 7).

Discussion

In this study, 2388 DEGs (including 1574 up-regulated genes and 814 down-regulated genes) between the LUAD samples and normal samples in the training set were identified. For the 150 metabolism-associated DEGs, 17 GO_BP functional terms and 20 KEGG pathways (such as PPAR signaling pathway) were enriched. Based on the optimal gene combination involving eight genes (*CYP17A1*, *ASPG*, *DUOX1*, *CIDEC*, *TH*, *B4GALNT1*, *APOA2*, and *GCKR*), the PS model was constructed. Combined with pathologic stage and PS model status, the nomogram survival model was established.

Through mediating the p(38)/β-catenin/PPARγ pathway, transforming growth factor β (*TGFβ*) enhances the invasion, migration, and epithelial mesenchymal transition of NSCL CH460 cells (25, 26). Therefore, the PPAR signaling pathway might function in the pathogenesis of LUAD. The silver nanobiocomposite of *ASPG* has lower cytotoxicity against lung cancer cells, and thus it can be applied as an promising anticancer agent for treating the tumor (27). *DUOX1* affect epithelial homeostasis and innate airway host defense, and its silencing accelerates epithelial-to-mesenchymal transition in lung epithelial cancer and may be related to invasive and metastatic lung cancer (28, 29). *DUOX1* expression is inhibited by epigenetic mechanisms in lung cancer, which may be implicated in the diagnosis and therapy of the tumor (30). Thus, *ASPG* and *DUOX1* might be implicated in the prognosis of LUAD patients.

CIDEC, a member of the cell death-inducing DNA fragmentation factor-like effector family, could promote lipid droplet formation in adipocytes and mediate adipocyte apoptosis(31). It was reported that the CIDE family regulated lipid metabolism and played an important role in the development of metabolic disorders such as obesity(32), insulin resistance(33) and hepatic steatosis(34)Ming Yu et al.According to HPA database, *CIDEC* was a characteristic gene in lung cancer and other cancers. A four-gene signature (involving *CIDEC*; dickkopf WNT signaling pathway inhibitor 1, *DKK1*; ubiquitin specific peptidase 4, *USP4*; and ZFP3 zinc finger protein, *ZFP3*) is an independent prognostic marker, which can be used to predict the outcomes of LUAD patients (35). Through CIDEC/extracellular regulated MAP kinase (ERK)/p38 pathway, ADP ribosylation factor like GTPase 14 (*ARL14*) is involved in the mechanisms of LUAD and can be novel prognostic marker and therapeutic target of the disease (36). *B3GALNT1* and tubulointerstitial nephritis antigen-like 1 (*TINAGL1*) are found to be key genes mediating the metastasis of NSCLC, which are candidate target genes for the treatment of the disease (37, 38). A model involving six biomarkers (including *APOA2*) and age has high sensitivity and specificity, which is effective for the diagnosis of the patients with lung cancer (39). These suggested that *CIDEC*, *B4GALNT1*, and *APOA2* might also be correlated with the prognosis of LUAD patients.

Although the roles of *CYP17A1*, *TH*, and *GCKR* in LUAD have not been studied, they have influences on other tumors. For example, *CYP17A1* and *CYP19A1* are important for estrogen biosynthesis, and their expression may contribute to improving the accuracy of diagnosis and selecting the proper treatment for invasive ductal breast cancer (40).Recently,one research reported that *CYP17A1* significantly distinguished between non-smoking and smoking-associated adenocarcinomas.(41)The CT and CC genotypes of *TH* have higher frequencies in gastric cancer patients compared with the controls, therefore, they are correlated with a significantly higher risk of the tumor (42). The rs780093 and rs780094 polymorphisms in *GCKR* have significant correlations with OS and progression-free survival, and thus *GCKR* polymorphisms may be independent prognostic marker in metastatic gastric cancer patients receiving EOF chemotherapy (43). Therefore, *CYP17A1*, *TH*, and *GCKR* might play roles in the prognosis of LUAD patients.

This study constructed a PS model based on metabolism-associated genes and a nomogram survival model combined with multiple bioinformatics methods. However, no experiments were conducted to validate these findings. Therefore, these results still needed to be confirmed by experimental studies.

In conclusion, 2388 DEGs between the LUAD and normal samples were obtained. Besides, the PPAR signaling pathway might affect the pathogenesis of LUAD. Furthermore, the PS model (involving *CYP17A1*, *ASPG*, *DUOX1*, *CIDEC*, *TH*, *B4GALNT1*, *APOA2*, and *GCKR*) and the nomogram survival model might be effective for the prognostic prediction of LUAD patients.

Declarations

Disclosures

No conflicts of interest, financial or otherwise are declared by the authors.

Author Contributions

W.HX and S.GQ. designed the research; W.HX and S.GQ. and Z.KK. performed computation and data analysis. W.H wrote the main manuscript text and prepared all the figures. W.HXand S.GQ.discussed the results and revised the manuscript. All authors contributed to discussions about the results and the manuscript.

Acknowledgments

This study was financially supported by Sanming Project of Medicine in Shenzhen “the Integrated Airways Disease team led by Professor Kian Fan Chung from Imperial College London” (SZSM201612096).

References

1. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018;553(7689):446–54.
2. Latimer KM. Lung Cancer: Clinical Presentation and Diagnosis. Fp Essent. 2018;464:23–6.
3. Shim HS, et al. Histopathologic Characteristics of Lung Adenocarcinomas With Epidermal. Archives of Pathology Laboratory Medicine. 2011;135(10):1329–34.
4. Wood DE, et al. Lung Cancer Screening, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. Journal of the National Comprehensive Cancer Network Jnccn. 2018;16(4):412–41.
5. Rami-Porta R, et al. Lung cancer staging: a concise update. Eur Respir J. 2018;51(5):1800190.
6. Dos Santos VM, Lam DS. Cardiac and lymphatic metastases from lung cancer. Archives of Iranian Medicine. 2018;21(2):82.
7. Nagashio R, et al. Prognostic significance of G6PD expression and localization in lung adenocarcinoma. Biochim Biophys Acta. 2018;1867(1):38–46.
8. Liu W, et al. COL5A1 may contribute the metastasis of lung adenocarcinoma. Gene. 2018;665:57–66.
9. Zhong Y, Lin JL. H, et al., Overexpression of KIF18A promotes cell proliferation, inhibits apoptosis, and independently predicts unfavorable prognosis in lung adenocarcinoma. IUBMB Life. 2019;71(7):942–55.

10. Li X, et al. High kinesin family member 18A expression correlates with poor prognosis in primary lung adenocarcinoma. Thoracic Cancer. 2019;10(5):1103–10.
11. Ding F, et al. Overexpression of S100A14 contributes to malignant progression and predicts poor prognosis of lung adenocarcinoma. Thoracic Cancer. 2018;9(7):827–35.
12. Yun-Yong P, et al. Development and Validation of a Prognostic Gene-Expression Signature for Lung Adenocarcinoma. Plos One. 2012;7(9):e44225.
13. Wang Y, et al. A novel 4-gene signature for overall survival prediction in lung adenocarcinoma patients with lymph node metastasis. Cancer Cell Int. 2019;19(1):100.
14. Brault C, Schulze A. The Role of Glucose and Lipid Metabolism in Growth and Survival of Cancer Cells. Metabolism in Cancer. 2016;207:1–22.
15. Dodesini AR, et al. Protein, glucose and lipid metabolism in the cancer cachexia: A preliminary report. Acta Oncol. 2007;46(1):118–20.
16. Ritchie ME, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
17. Diao X, Xiao. Identification and analysis of key genes in osteosarcoma using bioinformatics. Oncology Letters. 2018;15(3):2789–94.
18. Tilford CA. S.N.O., Gene Set Enrichment Analysis. Methods Mol Biol. 2009;563:99–121.
19. Da WH, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008;4(1):44–57.
20. Wang P, et al. A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. Oncotarget. 2016;7(34):55343–51.
21. Goeman JJ. L1 Penalized Estimation in the Cox Proportional Hazards Model. Biom J. 2010;52(1):70–84.
22. Eng KH, Schiller E, Morrel K. On representing the prognostic value of continuous gene expression biomarkers with the restricted mean survival curve. Oncotarget. 2015;6(34):36308–18.
23. Pontén F, Jirström K, Uhlen M. The Human Protein Atlas - A tool for pathology. J Pathol. 2008;216(4):387–93.
24. Anh N, Tran, et al. HPAanalyze: an R package that facilitates the retrieval and analysis of the Human Protein Atlas data. BMC Bioinformatics. 2019;20:463.
25. Lin LC, et al., *TGFβ can stimulate the p38/β-catenin/PPARγ signaling pathway to promote the EMT, invasion and migration of non-small cell lung cancer (H460 cells)*. Clinical & Experimental Metastasis. 31(8): p. 881–895.
26. Reka AK, et al. Molecular cross-regulation between PPAR-γ and other signaling pathways: Implications for lung cancer therapy. Lung Cancer. 2011;72(2):154–9.
27. Baskar G, George GB, Chamundeeswari M. Synthesis and Characterization of Asparaginase Bound Silver Nanocomposite Against Ovarian Cancer Cell Line A2780 and Lung Cancer Cell Line A549. Journal of Inorganic Organometallic Polymers Materials. 2016;27(1):1–8.
28. Little AC, et al. DUOX1 silencing in lung cancer promotes EMT, cancer stem cell characteristics and invasive properties. Oncogenesis. 2016;5(10):e261.
29. Kolářová H, et al., *The Expression of NADPH Oxidases and Production of Reactive Oxygen Species by Human Lung Adenocarcinoma Epithelial Cell Line A549*. 2010. 56(5): p. 211.
30. Little AC, et al. Paradoxical roles of dual oxidases in cancer biology. Free Radical Biol Med. 2017;110:117–32.

31. Xu Y, Gu Y, Liu G, Zhang F, Li J, Liu F, et al. Cidec promotes the differentiation of human adipocytes by degradation of AMPKalpha through ubiquitin-proteasome pathway. Biochim Biophys Acta. 2015;1850:2552–62.
32. Zhou L, Yu M, Arshad M, Wang W, Lu Y, Gong J, et al. Coordination Among Lipid Droplets,Peroxisomes, and Mitochondria Regulates Energy Expenditure Through the CIDE-ATGL-PPARalpha Pathway in Adipocytes. Diabetes. 2018;67:1935–48.
33. Wang H, Ti Y, Zhang JB, Peng J, Zhou HM, Zhong M, et al. Single nucleotide polymorphisms in CIDEC gene are associated with metabolic syndrome components risks and antihypertensive drug efficacy. Oncotarget. 2017;8:27481–8.
34. Andersen E, Ingerslev LR, Fabre O, Donkin I, Altintas A, Versteyhe S, et al. Preadipocytes from obese humans with type 2 diabetes are epigenetically reprogrammed at genes controlling adipose tissue function. Int J Obes (Lond). 2019;43:306–18.
35. Yin XH, et al. Development and validation of a 4-gene combination for the prognostication in lung adenocarcinoma patients. J Cancer. 2020;11(7):1940–8.
36. Guo F, et al. Silencing of ARL14 Gene Induces Lung Adenocarcinoma Cells to a Dormant State. Front Cell Dev Biol. 2019;7:238.
37. Umeyama H, Iwadate M, Taguchi Y-h. TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer. Bmc Genomics. 2014;15(9 Supplement):S2.
38. Yoshida H, et al. B4GALNT1 induces angiogenesis, anchorage independence growth and motility, and promotes tumorigenesis in melanoma by induction of ganglioside GM2/GD2. Sci Rep. 2020;10(1):1199.
39. Yoon HI, et al. Diagnostic Value of Combining Tumor and Inflammatory Markers in Lung Cancer. J Cancer Prev. 2016;21(3):187–93.
40. Tüzüner MB, et al. Evaluation of Local CYP17A1 and CYP19A1 Expression Levels as Prognostic Factors in Postmenopausal Invasive. Ductal Breast Cancer Cases. 2016;54(6):1–19.
41. Zhou D, Sun Y, Jia Y, et al. Bioinformatics and functional analyses of key genes in smoking-associated lung adenocarcinoma. Oncol Lett. 2019;18(4):3613–22. doi:10.3892/ol.2019.10733.
42. Li ZQ, et al. Association of gastric cancer with tyrosine hydroxylase gene polymorphism in a northwestern Chinese population. Clinical Experimental Medicine. 2007;7(3):98–101.
43. Liu X, et al. Effects of IGF2BP2, KCNQ1 and GCKR polymorphisms on clinical outcome in metastatic gastric cancer treated with EOF regimen. Pharmacogenomics. 2015;16(9):1–12.

Figures

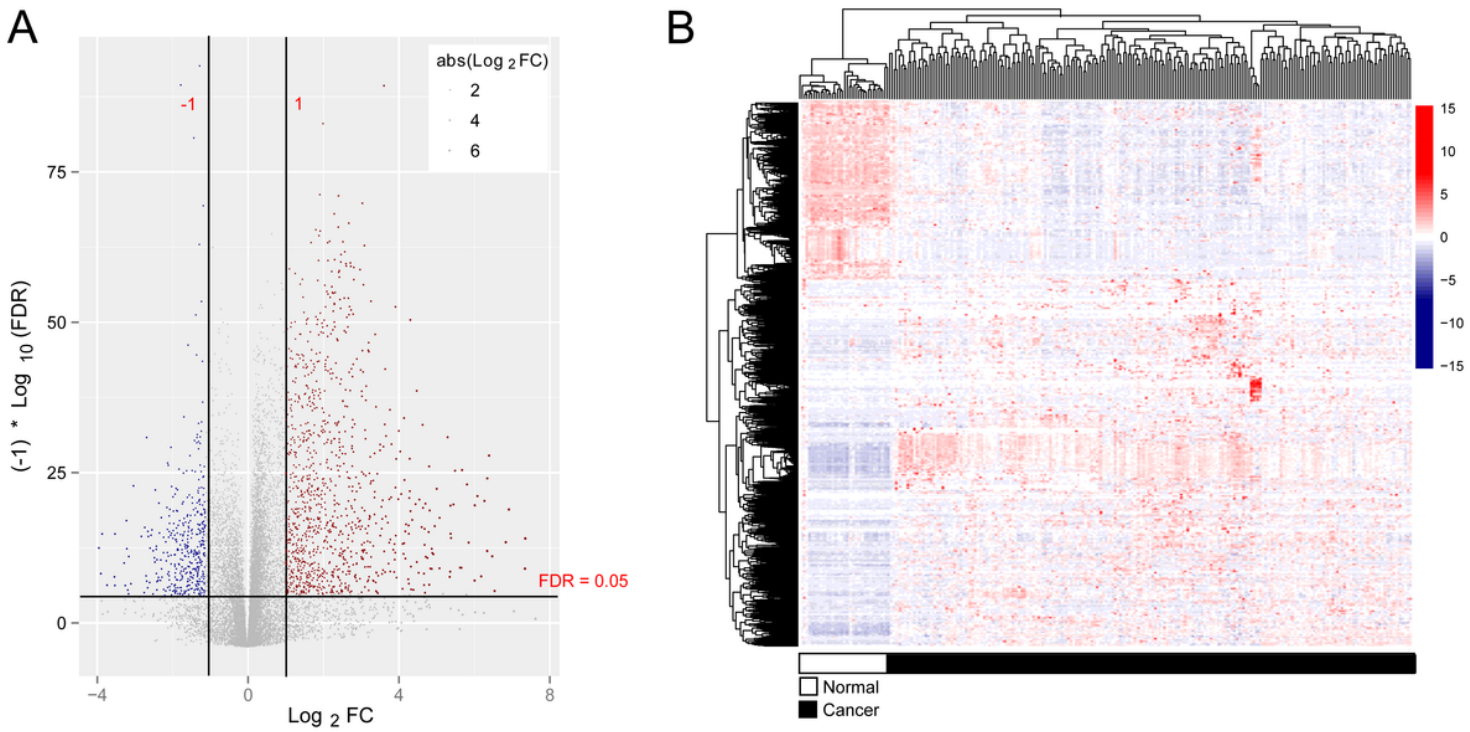

Figure 1

The results of differential expression analysis. (A) The volcano plot for the differentially expressed genes (DEGs) (red and blue dots represent DEGs; the horizontal line and vertical lines separately represent FDR < 0.05 and |log2FC| > 1); (B) The bidirectional hierarchical clustering heatmap for the DEGs. White sample strip and black sample strip represent normal and cancer samples, respectively. FC, fold change; FDR, false discovery rate.

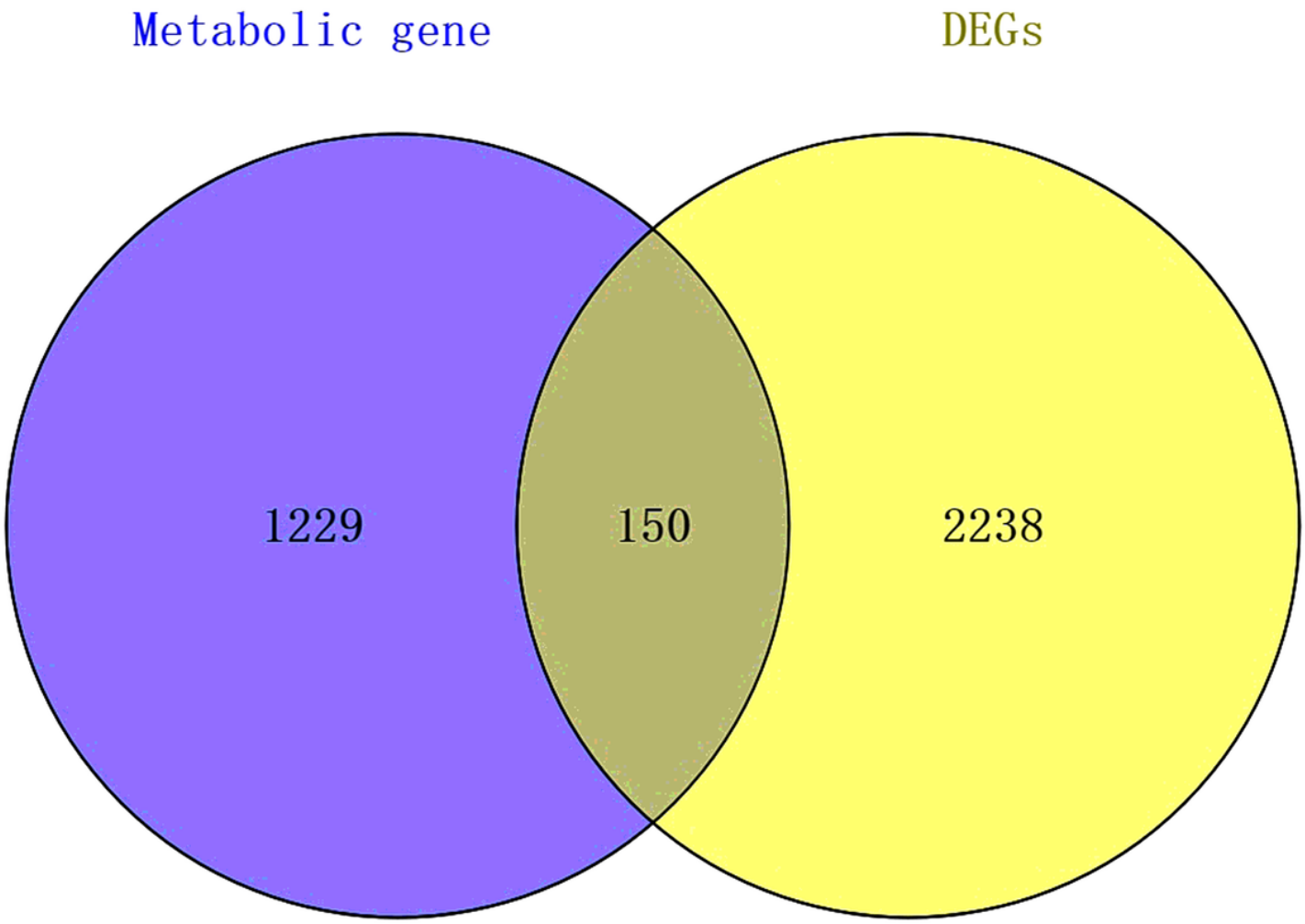

Figure 2

The Venn diagram showing the intersected genes of the metabolic genes and the differentially expressed genes (DEGs).

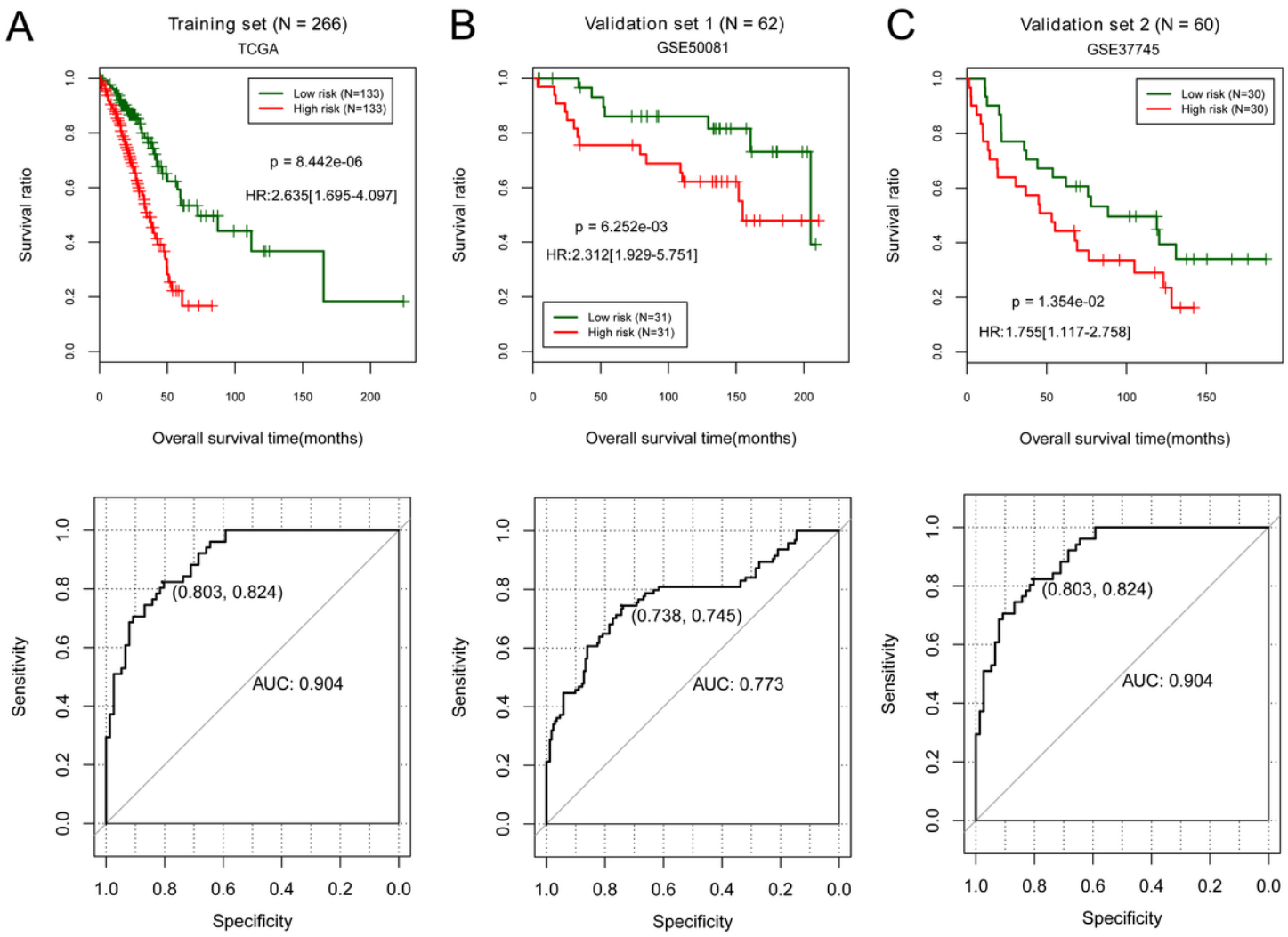

Figure 3

The Kaplan-Meier (KM) curves and receiver operating characteristic (ROC) curves showing the correlation between the actual survival prognosis and the risk grouping. (A) The KM curves (above) and ROC curve (below) for the training set; (B) The KM curves (above) and ROC curve (below) for the validation set GSE50081; (C) The KM curves (above) and ROC curve (below) for the validation set GSE37745. In KM curves, green and red curves separately represent low risk and high risk groups. In ROC curves, the numbers in brackets denote the specificity and sensitivity of the ROC curves. TCGA, The Cancer Genome Atlas; HR, hazard ratio; AUC, area under the receiver operating characteristic curve.

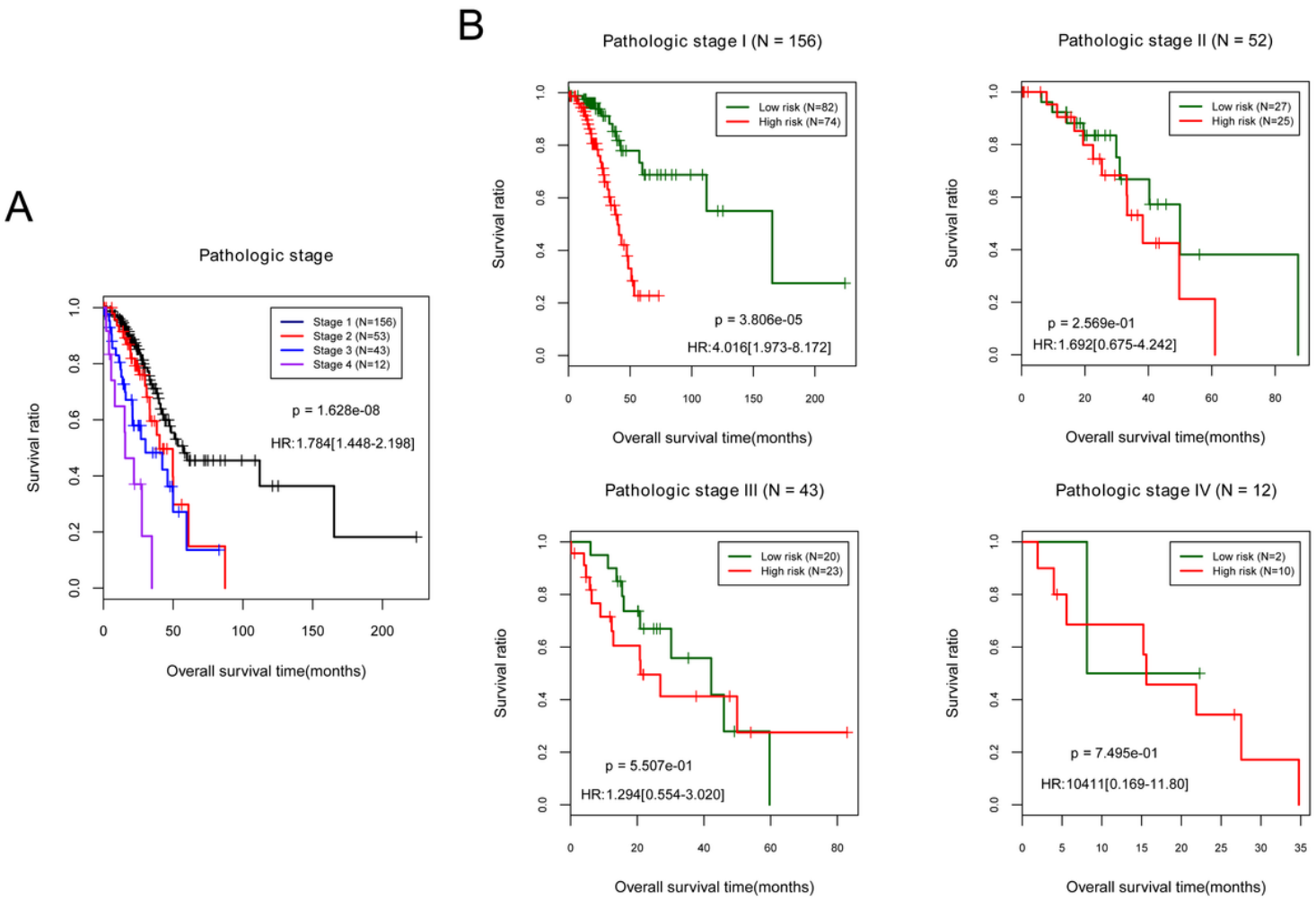

Figure 4

The Kaplan-Meier (KM) curves for pathologic stage in the training set. (A) The KM curves showing the correlation between pathologic stage and the actual prognosis (black, red, blue, and purple curves separately represent stage I, stage II, stage III, and stage IV groups); (B) The KM curves showing the correlations between the prediction results of prognostic score model and the actual prognosis in the samples in pathologic stage I, stage II, stage III, and stage IV (green and red curves separately represent low risk and high risk groups). HR, hazard ratio.

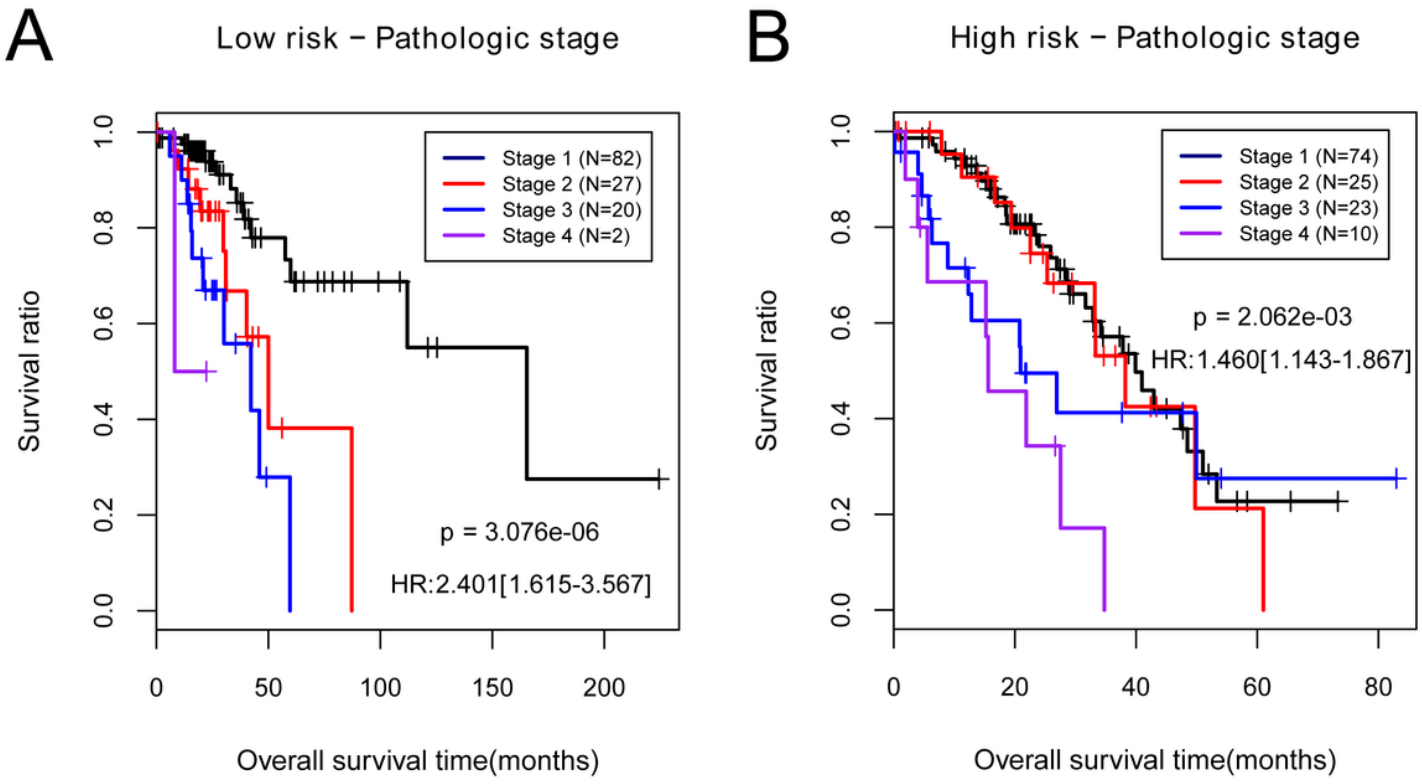

Figure 5

The Kaplan-Meier (KM) curves of pathologic stage in the low risk and high risk groups. (A) The KM curves in low risk group; (B) The KM curves in high risk group. Black, red, blue, and purple curves represent stage I, stage II, stage III, and stage IV groups, respectively.

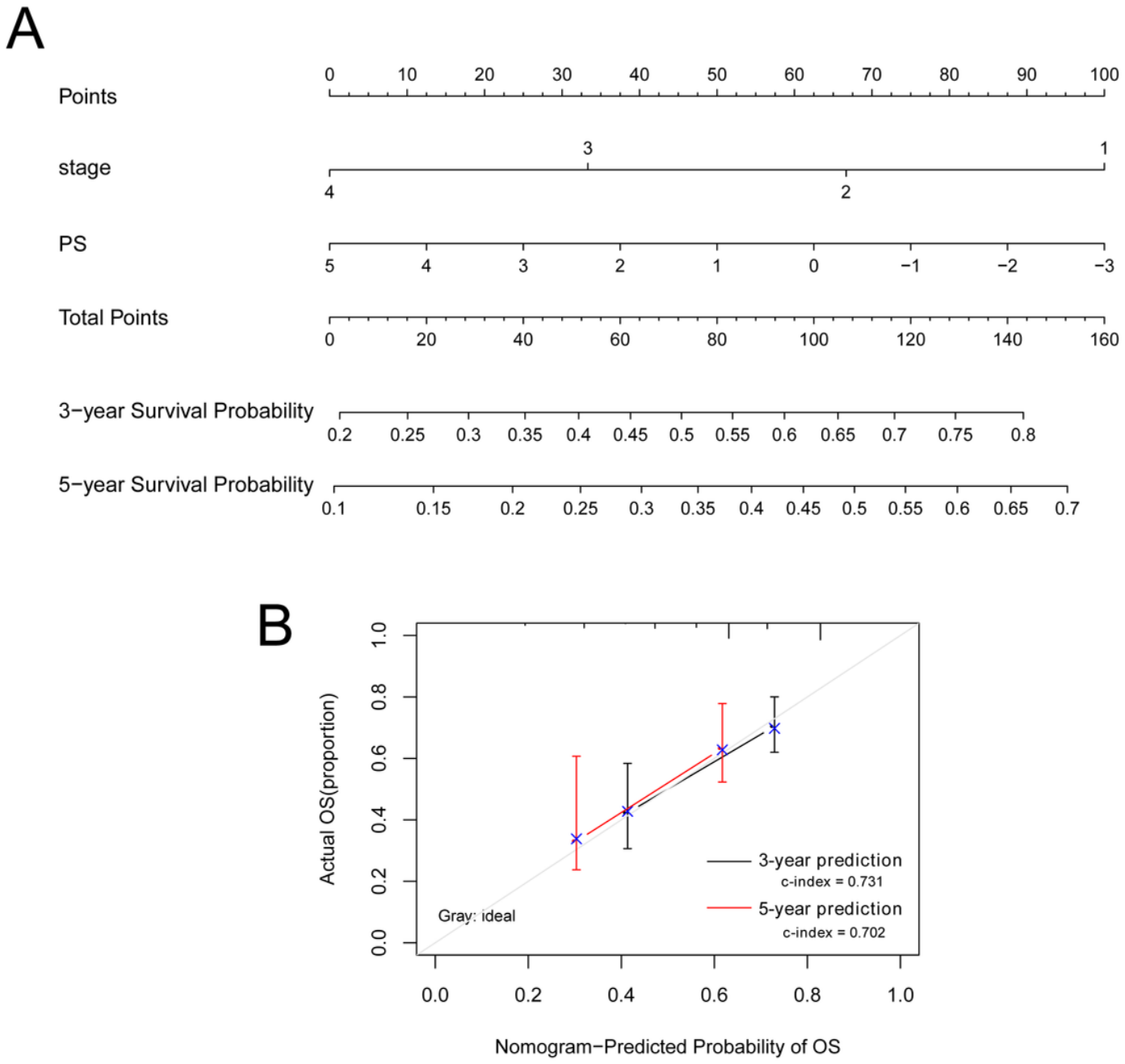

Figure 6

The nomogram survival model and the consistency chart of the nomogram-predicted probability of overall survival (OS) and the actual OS. (A) The nomogram survival model involving pathologic stage and prognostic score (PS) model status. (B) The consistency chart of the nomogram-predicted probability of OS and the actual OS (the horizontal axis and vertical axis separately represent nomogram-predicted probability of OS and the actual OS; black and red represent 3-year prediction and 5-year prediction, respectively).

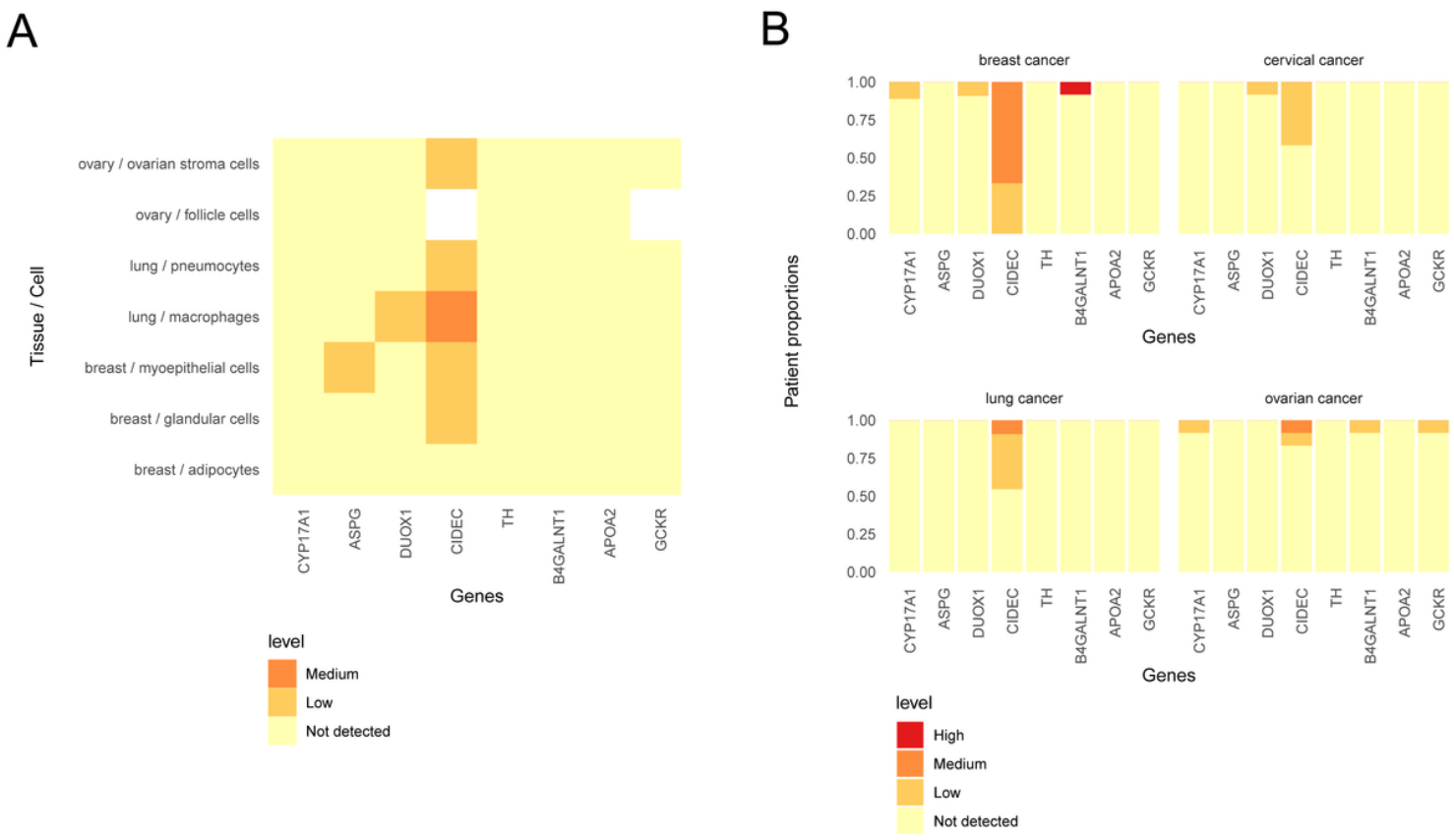

Figure 7

The protein expression levels of the optimal genes in Human Protein Atlas (HPA) database. (A) The heatmap for tissue atlas; (B) The heatmap for pathology atlas.