A Glycolysis-Based Three-Gene Signature Predicts Survival in Patients with Lung Squamous Cell Carcinoma

Guichuan Huang (drhuangguichuan@126.com)
The third affiliated hospital of Zunyi Medical University
Jing Zhang
The affiliated hospital of Zunyi Medical University
Ling Gong
the third affiliated hospital of Zunyi Medical University
Yi Huang
the third affiliated hospital of Zunyi Medical University
Daishun Liu
Zunyi Medical University

Research article

Keywords: lung cancer, glycolysis, prognosis, gene signature

Posted Date: July 16th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-42550/v1

Version of Record: A version of this preprint was published at BMC Cancer on May 27th, 2021. See the published version at https://doi.org/10.1186/s12885-021-08360-z.

Abstract

Purpose: Lung cancer is one of the most lethal and most prevalent malignant tumors worldwide, and lung squamous cell carcinoma (LUSC) is one of major histological subtypes. Although, numerous biomarkers were found to be associated with prognosis in LUSC, the prediction effect of a single gene biomarker is not sufficient, especially for glycolysis-related genes. Therefore, we aimed to develop a novel glycolysis-related gene signature to predict survival of patients with LUSC.

Material and Methods:

The mRNA expression files and clinical information of LUSC were obtained from The Cancer Genome Atlas (TCGA) dataset.

Results: Based on Gene set enrichment analysis (GSEA), we found 5 glycolysis-related gene sets were significantly enriched in LUSC tissues. Univariate and multivariate Cox proportional regression models were conducted to choose prognostic-related gene signature. Based on Cox proportional regression model, a risk score of three-gene signature (including HKDC1, ALDH7A1, and MDH1) was established to divide patients into high-risk and low-risk subgroups. We found that a risk score of three-gene signature was an independent of prognostic indicator in LUSC using multivariate Cox regression analysis.

Conclusion: In conclusion, a glycolysis-based three-gene signature could serve as a novel biomarker in predicting prognosis of patients with LUSC, which provided more gene targets to cure LUSC patients.

Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide. There are two clinical subtypes for lung cancer: non-small cell lung cancer (NSCLC) (approximately 85%), and small cell lung cancer (SCLC) (approximately 15%).[1] Based on pathological and molecular features, NSCLC is divided into the following major subtypes: lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), and large cell lung cancer.[2] Recently, advances in the targeted treatments have obviously improved the overall survival (OS) of patients with LUAD, such as epidermal growth factor receptor (EGFR) kinase inhibitors.[3] However, no specific biomarker and relatively optimal targeted therapies were identified for LUSC patients, and the 5-year survival rate of LUSC is less than 20%.[4] Therefore, it is necessary to explore specific diagnostic and prognostic biomarkers for LUSC.

Glycolysis, also known as Warburg effect, often observed in human cancer cells, in which the cancer cells favor glucose metabolism via glycolysis even in the presence of oxygen.[5] This phenomenon is a unique energy metabolism exist in cancer cells. In recent years, many biomarkers, including glycolysis-associated genes, for LUSC have been determined, such as kininogen 1 (KNG1) [6] and tripartite motif-containing protein 59 (TRIM59). [7] With the development of high-throughput sequencing, various patient genome databases were constructed, which makes us have a deep understanding of genomic changes. Based on database mining, an increasing number of biomarkers were identified that were related to

survival of patients with cancer. [8,9] However, a single gene cannot have good predictive effects. Multigene prognostic model from original tumor biopsy can guided clinicians to choose more effective treatment strategies. Thus, a signature based on multigene expression associated with glycolysis should be established to predict the prognosis of LUSC patients.

In the present study, we used a genome-wide analysis of mRNA expression profiles in LUSC patients from The Cancer Genome Atlas (TCGA) to construct a glycolysis-related gene signature, which could effectively predict the prognosis in LUSC patients.

Materials And Methods

Patient dataset

The mRNA expression profiles of LUSC patients and their corresponding clinical information were obtained from the TCGA database (https://portal.gdc.cancer.gov/).

A total of 501 patients with LUSC and 49 adjacent normal samples were included for the following study. Clinical information, including age, gender, TNM stage, T stage, M stage, and N stage was included in present study (Table 1).

Table 1
Clinical characteristic of LUSC (n = 501) from TCGA database

Clinical characteristic	N	%
Age		
≤ 65	190	37.9
> 65	302	60.3
NA	9	1.8
Gender		
Female	130	25.9
Male	371	74.1
TNM Stage		
I-II	406	81.0
III-IV	91	18.2
NA	4	0.8
T stage		
T1-T2	407	81.2
T3-T4	94	18.8
N stage		
N0	319	63.7
N1–3	176	35.1
NA	6	1.2
M stage		
M0	411	82.0
M1	7	1.4
NA	83	16.6

Gene Set Enrichment Analysis (gsea)

GSEA was performed to determine whether the identified gene sets were significant difference between the LUSC and normal groups. The expression levels of 443 glycolysis-related genes were analyzed in

LUSC samples and in adjacent non-cancerous tissues. Normalized *p* value less than 0.05 was considered statistically significant.

Prognostic Analysis

We conducted univariate Cox proportional hazard regression analysis to determine the relationship between glycolysis-related genes and overall survival in LUSC patients.

If the $p < 0.01$, the corresponding glycolysis-related genes were retained and regarded as the candidate prognostic genes of LUSC. Then, the multivariate Cox proportional hazards regression analysis was performed among the pool candidate prognostic glycolysis-related genes to establish the prognostic model. These analyses were performed with the use of R package of survival.

Statistical analysis

The selected mRNAs were divided into the risky (HR > 1) and protective (0 < HR < 1) types. Based on a linear combination of the expression level of filtered mRNAs weighted by the regression coefficient (β), the formula of risk score was illustrated as follows: Risk score = expression of gene 1 × β1 + expression gene 2 × β2 +…+expression of gene n × βn. The β represents the regression coefficient of the corresponding gene obtained from the multivariate cox regression model. According the median value of risk score, patients were divided into high-risk and low-risk groups. Kaplan-Meier curves and log-rank test were utilized to validate the prognostic significance of the risk score. The Student's t test was conducted to explore the differential expression of selected genes in LUSC tissues and adjacent normal tissue. Filtered gene alterations in LUSC were explored using cBioPortal database (http://www.cbioportal.org/). All statistical analyses were conducted with the use of SPSS 23.0 and GraphPad Prism 8.0 software.

Results

Initial Screening Of Genes By Gsea

The mRNA expression data set and clinical information of 501 patients with LUSC were obtained from the TCGA database. We found five glycolysis-related gene sets on the Molecular Signatures Database v7.0, including (1)BIOCARTA_GLYCOLYSIS_ PATHWAY, (2) GO_GLYCOLYTIC_PROCESS, (3) HALLMARK_GLYCOLYSIS, (4) KEGG_GLYCOLYSIS_GLUCONEOGENESIS, (5) REACTOME_ GLYCOLYSIS. We performed GSEA to explore whether the identified gene sets were significant difference between LUSC and normal tissues. We found these 5 gene sets were significantly enriched (Fig. 1 and Table 2). Then, we collected 443 genes from 5 gene sets for further analysis.

Table 2
Gene set enriched in LUSC

Gene sets follow link to MSigDB	Size	NES	NOM p-val	FDR q-val
BIOCARTA_GLYCOLYSIS_ PATHWAY	3	1.43	0.029	0.029
GO_GLYCOLYTIC_PROCESS	106	2.01	0	0
HALLMARK_GLYCOLYSIS	200	2.29	0	0
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	62	1.56	0.028	0.028
REACTOME_ GLYCOLYSIS	72	2.28	0	0

Identification Of Glycolysis-related Genes Associated With Patient Survival

First, univariate Cox proportional hazard regression analysis was conducted to 443 genes that were significantly enriched in LUSC samples from GSEA. A total of 4 genes were obtained which were significantly correlated to the survival of patients ($p < 0.01$). Next, we performed multivariate Cox regression analysis to further explore the association between the 4 mRNA expression profiles and the overall survival of patients. Finally, 3 genes, including HKDC1, ALDH7A1, and MDH1, were included to construct prognostic model. As shown in Table 3, two of the three genes were verified as independent prognostic markers in LUSC. Among three genes, one gene (MDH1) was considered as protective factor owing to $0 < HR < 1$, whereas the remaining two genes (HKDC1 and ALDH7A1) might be prognostic risky factors with their $HR > 1$.

Table 3
Details of three genes for constructing the prognostic model

Gene	Ensemble ID	Location	HR (95%CI)	Coefficient	*p* value
HKDC1	ENSG00000156510	chr10: 69,220,303 – 69,267,559	1.1733	0.1598	0.0446
ALDH7A1	ENSG00000164904	chr5: 126,531,200 – 126,595,390	1.1701	0.1571	0.0097
MDH1	ENSG00000014641	chr2: 63,588,609 – 63,607,197	0.7682	-0.2636	0.0559

Subsequently, we explored the alterations of three selected genes in 501 LUSC samples using cBioPotral database. The results showed that the rate of alterations in HKDC1, ALDH7A1, and MDH1 genes were 1.9%, 1.1%, and 5%, respectively (Supplementary Fig. 1).

The expression level of three genes was conducted between adjacent normal tissues and LUSC tissues. We found that all the three genes were upregulated in LUSC tissues compared with in normal tissues (Fig. 2).

Construction Of The Three-gene Signature To Predict Patient Prognosis

To predict patient's prognosis using glycolysis-related genes expression, a prognostic risk model was developed based on the regression coefficients of multivariate Cox

regression model to weight the expression level of each gene in the three-gene signature:

risk score = 0.1598 × expression value of HKDC1 + 0.1571 × expression value of ALDH7A1 + (-0.2636) × expression value of MDH1. Owing to 6 of 501 patients lacked the data of survival time, a total of 495 patients were included in the survival analysis. According to the risk score formula, patients were classified into the high-risk (n = 247) and the low-risk group (n = 248) with a medial value of risk score as a cut-off (Fig. 3A). The survival time of each patient was shown in Fig. 3B. As shown in Fig. 3D, patients in high-risk group had a shorter survival than in low-risk group ($p < 0.001$). The 3-year and 5-year survival rates of patient in high-risk group were 45.4% and 35.0%, respectively. However, the 3-year and 5-year survival rates of low-risk group were 71.9% and 58.1%, respectively. Additionally, a heatmap presented the expression profiles of three mRNAs (Fig. 3C). As the risk score increased in the patients with LUSC, the mRNA expression of HKDC1 and ALDH7A1 was obviously upregulated; in contrast, the mRNA expression of MDH1 was downregulated. The area under the ROC curve (AUC) for the risk score on 3-year overall survival was 0.665 (Fig. 4).

Risk score from three-gene signature is an independent prognostic indicator

Univariate and multivariate Cox regression analysis were performed to evaluate the independent risk factors in patients with LUSC. Several clinicopathological parameters, including age, gender, TNM stage, T stage, N stage, and M stage, as well as risk score were included. The results showed that only risk score was associated with prognosis in the univariate Cox analysis (HR = 2.553, 95%CI: 1.710–3.811, $p < 0.0001$) (Table 4). In the following multivariate Cox analysis, we found age and risk score as independent prognostic indicators (Table 4). These results indicated that the risk score was reliable in predicting the prognosis of patients with LUSC.

Table 4
Univariate and multivariate Cox regression analysis of clinicopathologic factors and glycolysis-related genes signature for OS

Clinical features	Univariate analysis			Multivariate analysis		
	HR	95%CI of HR	P value	HR	95%CI of HR	P value
Age (> 65 vs. <=65)	1.376	1.000-1.892	0.050	1.489	1.078–2.055	0.016
Gender (Male vs. Female)	1.360	0.944–1.958	0.099	1.300	0.901–1.877	0.161
TNM stage (III-IV vs. I-II)	1.392	0.974–1.989	0.070	1.185	0.716–1.962	0.510
T stage (T3 + T4 vs. T1 + T2)	1.412	0.973–2.048	0.069	1.335	0.845–2.110	0.216
M stage (M1 vs. M0)	2.432	0.898–6.591	0.080	2.216	0.768–6.399	0.141
N stage (N1 + N2 + N3 vs. N0)	1.062	0.780–1.444	0.704	1.087	0.765–1.545	0.642
Risk score (high risk vs. low-risk)	2.553	1.710–3.811	< 0.0001	2.663	1.790–3.962	< 0.0001

Validation of three-gene signature for survival prediction by Kaplan-Meier curve analysis

To further verify the prognostic value of the risk score of the three-gene signature associated with glycolysis, patients with LUSC were stratified by age (≤ 65 or > 65), gender (Female or Male), TNM stage (I + II or III + IV), T stage (T1 + T2 or T3 + T4), N stage (N0 or N1 + N2 + N3), and M stage (M0 or M1) (Fig. 5). We found no significant difference between high-risk and low-risk in patients with tumor remote metastasis (Fig. 5). However, in the subgroup of patients without tumor remote metastasis, the risk score of the three-gene signature was still an independent prognostic indicator. Additionally, regardless of the age, gender, TNM stage, T stage, N stage, patients in the high-risk group based on the risk score had a poor prognosis than patients in the low-risk group. These findings demonstrated that the three-gene signature predicts effectively the survival of LUSC patients.

Discussion

Recently, numerous genes were considered as biomarkers for cancer prognosis and the clinical significance of the biomarkers has been explored. For example, a study made by Tang and his colleagues found that the overexpression of dipeptidyl peptidase 9 (DPP9) was a significant independent factor for poor prognosis in patients with NSCLC.[10] Similarly, Feng *et al.* [11] reported that high expression of forkhead box Q1 (FoxQ1) was associated with the poor prognosis in patients with NSCLC. However, the expression level of single gene can be influenced by multiple factors, making these biomarkers hard to be

reliable and independent prognosis indications in clinical. Therefore, a statistical model based on the combination of multiple genes was used to improve the prediction of prognosis in cancer patients. Studies have shown that the pool of multiple genes was more accurate than single gene in predicting the prognosis of patients with cancer.[12,13]

In the present study, we obtained mRNA expression profiles in 501 LUSC patients from TCGA database. We found that 5 glycolysis-related gene sets were significantly enriched in LUSC samples using GSEA. Univariate and multivariate Cox regression analysis were carried out to identify the risk score of the three-gene signature with prognostic value for patients with LUSC. Kaplan-Meier curve analysis indicated that

patients with high risk score had a poor prognosis than patients with low risk score. In additional, in stratified analysis, the risk score of three-gene signature could effectively predict the prognosis of LUSC patients in all subgroups except for the subgroup of patients with tumor remote metastasis. The reason for this discrepancy might be that the number of patients with tumor remote metastasis was too small (n = 7). These results demonstrated that the risk score of three-gene signature could be as an independent prognostic indicator for LUSC patients. Moreover, measuring the risk score of patients might help clinicians choose optimal therapy methods.

The metabolism of tumor cells is more active than the normal cells, thus tumor cells need more energy to keep their higher proliferation.[14] Glycolysis and oxidative phosphorylation are the two important metabolic pathways related to energy supply. Glycolysis is a relatively low-energy-providing pathway compared with oxidative phosphorylation. In 1920s, Warburg have found that cancer cells are very active in glycolysis and require a large amount of glucose to obtain ATP for metabolic activities.[15] This aberrant phenomenon of glucose metabolism was called aerobic glycolysis, also known as the Warburg effect.[15,16] After that, the main genes and enzymes related to glycolysis were begun to explore and further researched for understanding their functions in metabolism of tumor cells. In recent years, studies have shown that aerobic glycolysis plays a significant role in tumorigenesis, tumor progression and metastasis. For example, Enolase1 (ENO1) was proved to promote cell glycolysis, growth, migration, and invasion in NSCLC.[17] Glucose transporter 1 (GLUT1) facilitated increased transport of glucose into cancer cells to maintain an elevated rate of glycolysis under aerobic conditions.[18] A high expression of GLUT1 significantly associated with a poor prognosis in lung cancer patients.[19] However, no set of glycolysis-related genes for predicting LUSC prognosis has been established.

This study was the first to report that a glycolysis-based three-gene signature could serve as a prognostic indicator for patients with LUSC. A higher risk score indicates a worse prognosis. Of course, there still exist some limitations: First, the risk score model was constructed using TCGA database and should be verified in other cohorts in future researches. Second, studies on the three predicted genes should be made to explore the concreate mechanism in the occurrence and development of LUSC.

In conclusion, our study suggested that the three-gene signature associated with glycolysis might not only help to predict prognosis of LUSC patients, but also cloud provide more gene targets to cure LUSC

patients.

Abbreviations

LUSC, lung squamous cell carcinoma; TCGA, The Cancer Genome Atlas; NA, not available.

LUSC, lung squamous cell carcinoma; MSigDB, molecular signatures database; NES, normalized enrichment score; NOM p-val, nominal p-value; FDR q-val, false discovery rate q-value.

HR, hazard ratio.

OS, overall survival; HR, hazard ratio

Declarations

Disclosure Statement

The authors declare that they have no conflicts of interest.

Acknowledgements

This research was supported by the program of “Workstation of Academician in the First People's Hospital of Zunyi” funded by Zunyi Municipal Science and Technology Bureau (Zunshi Kehe (2015) No.17).

References

1. Osmani L, Askin F, Gabrielson E, Li QK. Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy. Semin Cancer Biol. 2018;52:103–9.
2. Molinier O, Goupil F, Debieuvre D, Auliac J-B, Jeandeau S, Lacroix S,et al. Five-year survival and prognostic factors according to histology in 6101 non-small-cell lung cancer patients. Respir Med Res. 2019;77:46–54.
3. Pirker R. What is the best strategy for targeting EGF receptors in non-small-cell lung cancer? Future oncology. 2015;11:153–67.
4. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69:7–34.
5. Sophia Y, Lunt MG, Vander Heiden. Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. Annu Rev Cell Dev Biol. 2011;27:441–64.
6. Wang W, Wang S, Zhang M. Evaluation of kininogen 1, osteopontin and α-1-antitrypsin in plasma, bronchoalveolar lavage fluid and urine for lung squamous cell carcinoma diagnosis.
7. Oncol Lett. 2020;19:2785–2792.

8. Lou M, Gao Z, Zhu T, Mao X, Wang Y, Yuan K, et al. TRIM59 as a novel molecular biomarker to predict the prognosis of patients with NSCLC. Oncol Lett. 2020;19:1400–8.
9. Wang X, Li G, Luo Q, Xie J, Gan C. Integrated TCGA analysis implicates lncRNA CTB-193M12.5 as a prognostic factor in lung adenocarcinoma. Cancer Cell Int. 2018;18:27.
10. Ge H, Yan Y, Wu D, Huang Y, Tian F. Potential role of LINC00996 in colorectal cancer: a study based on data mining and bioinformatics. Onco Targets Ther. 2018;11:4845–55.
11. Tang Z, Li J, Shen Q, Feng J, Liu H, Wang W, et al. Contribution of upregulated dipeptidyl peptidase 9 (DPP9) in promoting tumoregenicity, metastasis and the prediction of poor prognosis in non-small cell lung cancer (NSCLC). Int J Cancer. 2017;140:1620–32.
12. Feng J, Zhang X, Zhu H, Wang X, Ni S, Huang J. FoxQ1 overexpression influences poor prognosis in non-small cell lung cancer, associates with the phenomenon of EMT. PLoS One. 2012;7:e39937.
13. Zhang L, Zhang Z, Yu Z. Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. J Transl Med. 2019;17:423.
14. Liu C, Li Y, Wei M, Zhao L, Yu Y, Li G. Identification of a novel glycolysis-related gene signature that can predict the survival of patients with lung adenocarcinoma. Cell Cycle. 2019;18:568–79.
15. Abbaszadeh Z, Cesmeli S, Biray Avci C. Crucial players in glycolysis: Cancer progress. Gene 2020;726:144158.
16. Warburg O. On respiratory impairment in cancer cells. Science. 1956;124:269–70.
17. Koppenol WH, Bounds PL, Dang CV. Otto Warburg's contributions to current concepts of cancer metabolism. Nat Rev Cancer. 2011;11:325–37.
18. Fu Q, Liu Y, Fan Y, Hua S, Qu H, Dong S, et al. Alpha-enolase promotes cell glycolysis, growth, migration, and invasion in non-small cell lung cancer through FAK-mediated PI3K/AKT pathway. J Hematol Oncol. 2015;8:22.
19. Osugi J, Yamaura T, Muto S, Okabe N, Matsumura Y, Hoshino M, et al. Prognostic impact of the combination of glucose transporter 1 and ATP citrate lyase in node-negative patients with non-small lung cancer. Lung Cancer. 2015;88:310–8.
20. Zhang B, Xie Z, Li B. The clinicopathologic impacts and prognostic significance of GLUT1 expression in patients with lung cancer: A meta-analysis. Gene. 2019;689:76–83.

Figures

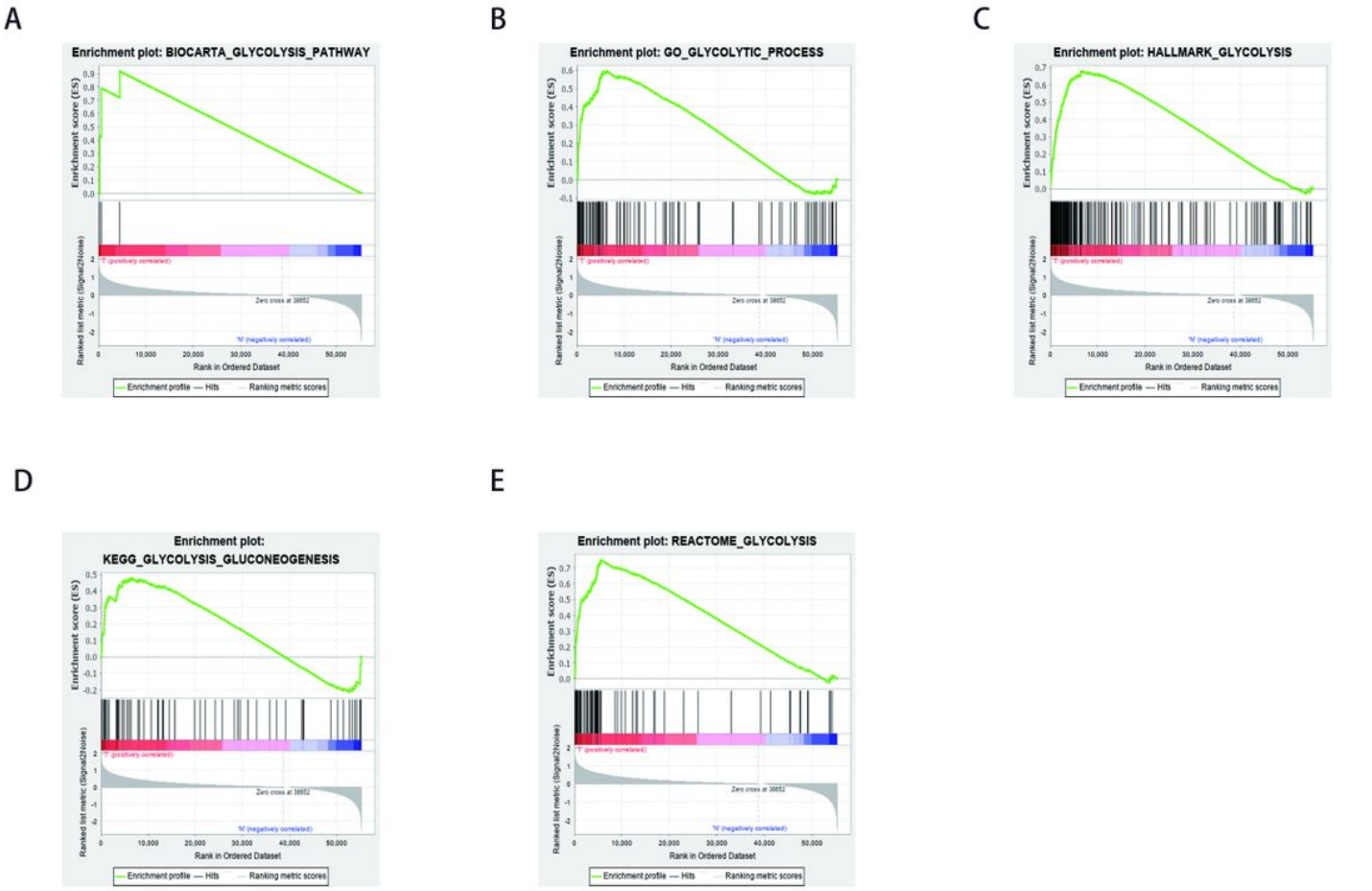

Figure 1

GSEA results for enrichment plots of five gene sets which were significantly differentiated between in LUSC and normal tissues. (A), BIOCARTA_GLYCOLYSIS_ PATHWAY; (B), GO_GLYCOLYTIC_PROCESS; (C), HALLMARK_GLYCOLYSIS; (D), KEGG_GLYCOLYSIS_GLUCONEOGENESIS; (E), REACTOME_ GLYCOLYSIS Abbreviations: GSEA, gene set enrichment analysis; LUSC, lung squamous cell carcinoma.

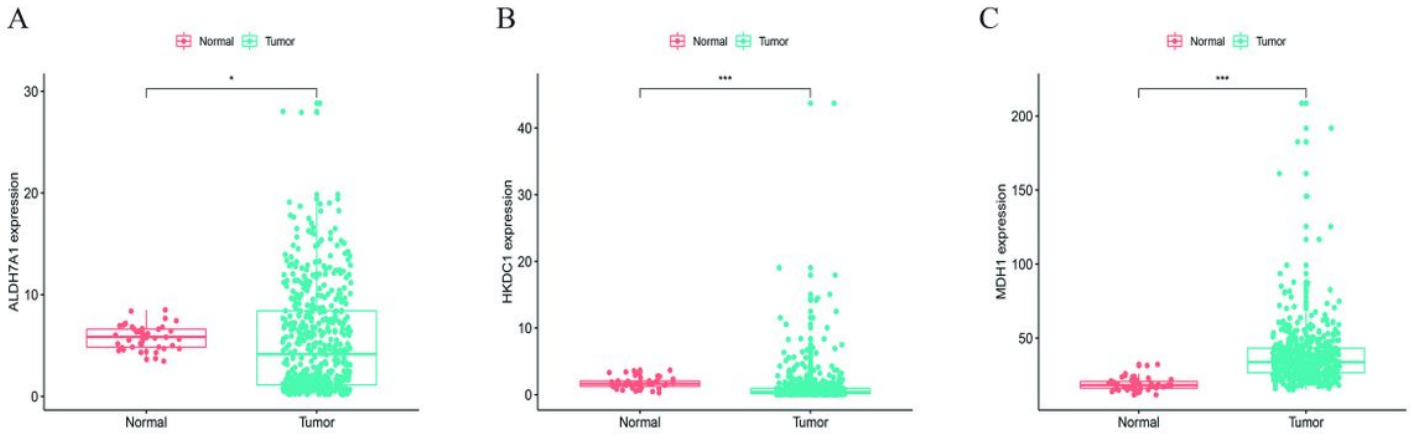

Figure 2

Differential expression of three genes in the normal tissues (n=49) and tumor tissues (n=501). (* p<0.05, **p<0.01, ***p<0.001) (A), ALDH7A1; (B), HKDC1; (C), MDH1.

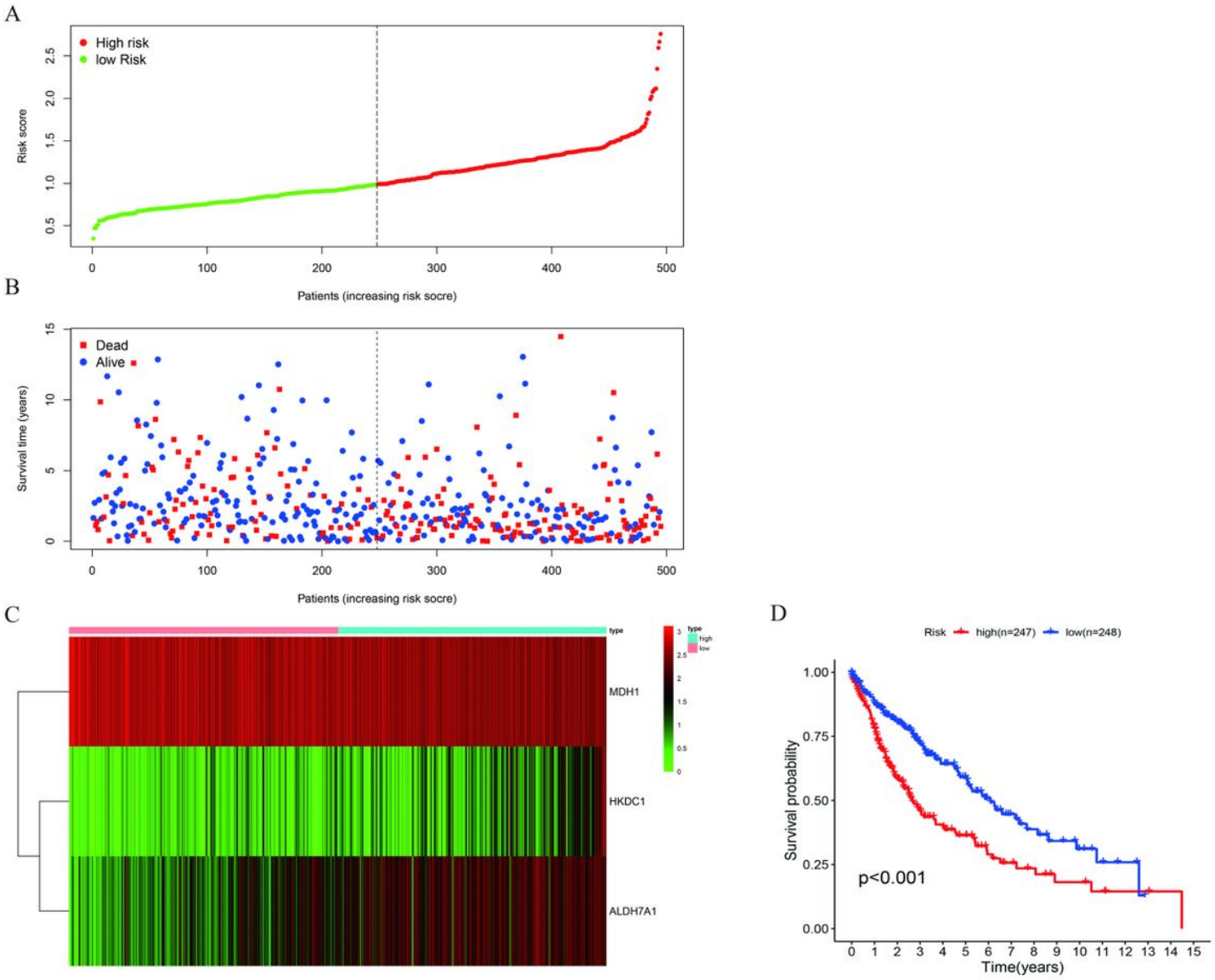

Figure 3

A risk of three-gene signature predicted the overall survival in patients with LUSC. (A), Distribution of risk score per patient; (B), Survival status of each patients; (C), A heatmap of three genes expression profile; (D), Kaplan-Meier survival curve analysis for LUSC patients divided into the high-risk and low-risk groups. Abbreviation: LUSC, lung squamous cell carcinoma.

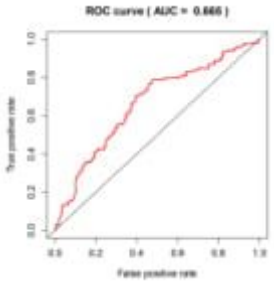

Figure 4

The time-independent ROC curve of the risk score for prediction the 3-year overall survival.

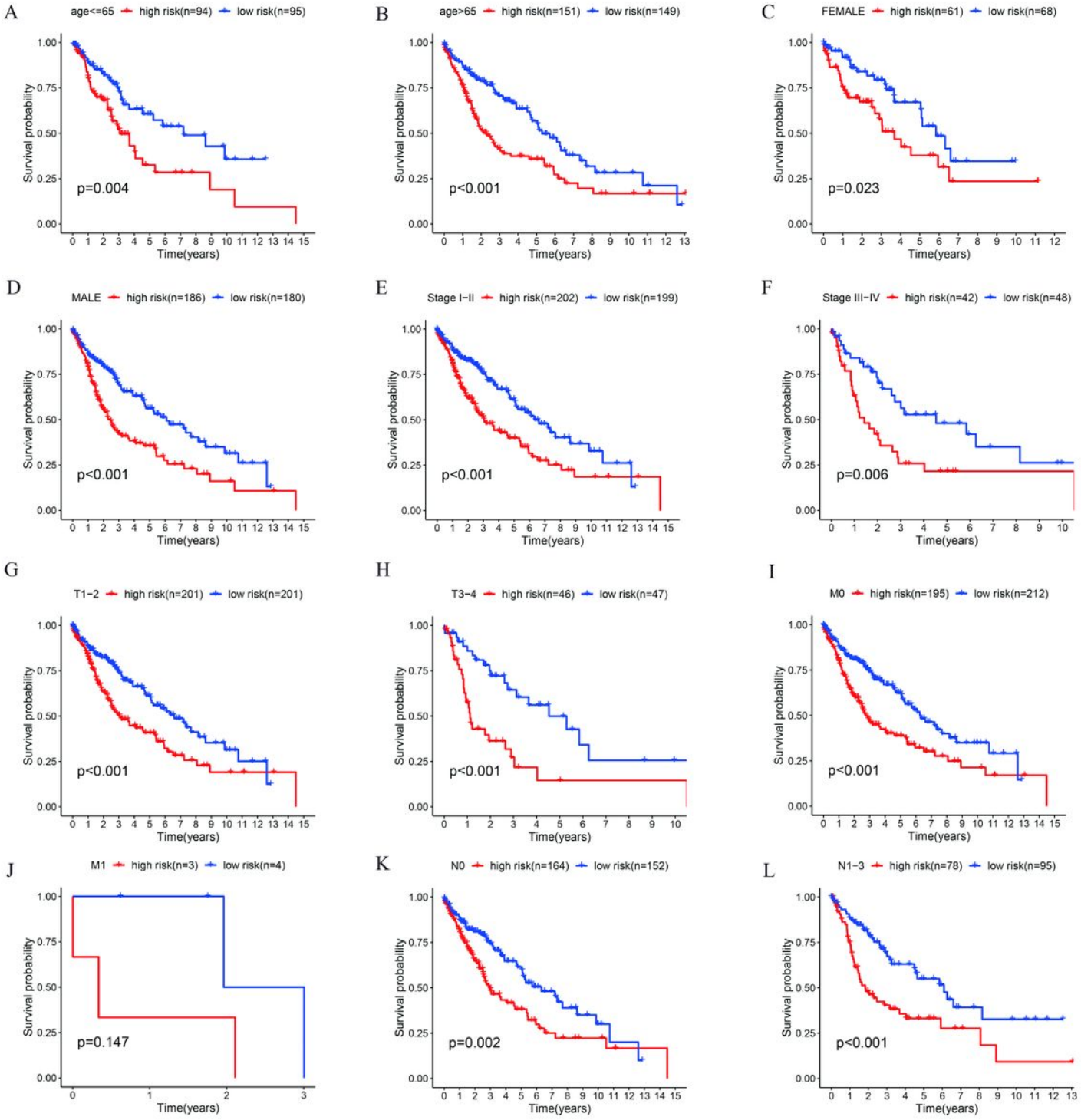

Figure 5

Supplementary Files

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

- SupplementaryFigure1.tif