



Research article

Development of metastasis-associated seven gene signature for predicting lung adenocarcinoma prognosis using single-cell RNA sequencing data

Jinqi He¹, Wenjing Zhang², Faxiang Li^{2,*} and Yan Yu^{1,*}

¹ Department of Medical Oncology, Harbin Medical University Cancer Hospital, Harbin 150081, China

² Department of Hematology Oncology, The Central Hospital of Shaoyang, Shaoyang 422000, China

* **Correspondence:** Email: yuyan@hrbmu.edu.cn; lifaxiang2011@163.com.

Abstract: Metastasis is the primary cause of lung adenocarcinoma (LUAD)-related death. This study evaluated the metastasis-associated genes (MAGs) in single-cell RNA sequencing (scRNA-seq) data from LUAD tissues and developed a MAG signature to predict overall survival (OS) of LUAD patients. The LUAD scRNA-seq data was downloaded from the Gene Expression Omnibus (GEO) Database and MAGs were identified from LUAD scRNA-seq data. The LUAD transcriptomic and clinical data were obtained from The Cancer Genome Atlas (TCGA). Cox and LASSO regression analyses were performed to identify differentially expressed MAGs (DEMAGs) with prognostic value that were then used to construct a MAG signature and MAG-nomogram model. Finally, a functional enrichment analysis was performed via Gene Set Enrichment Analysis (GSEA). 414 MAGs and 22 prognostic DEMAGs were revealed in the study. Multivariate Cox proportional hazards regression analysis was utilized to construct a 7-MAG signature for predicting the OS of LUAD patients. Patients with high risk scores had a significantly worse OS than those with low risk scores in the training group ($n = 236$), and the 7-MAG signature was successfully confirmed in the testing group ($n = 232$) and the entire TCGA-LUAD cohort ($n = 468$). Furthermore, univariate and multivariate Cox regression suggested that the 7-MAG signature was an independent prognostic indicator. Additionally, based on the 7-MAG signature, a nomogram was established that could more intuitively help to predict the OS of LUAD patients. The GSEA revealed the underlying molecular mechanisms of the 7-MAG signature in LUAD metastasis. In conclusion, a 7-MAG signature was developed based on LUAD scRNA-seq data that could effectively predict LUAD patient prognosis and provide novel insights for therapeutic targets and the potential molecular mechanism of metastatic LUAD.

Keywords: lung cancer; metastasis; single-cell RNA sequencing; prognosis; gene signature

Abbreviations: LUAD: Lung adenocarcinoma; MAG: Metastasis-associated gene

1. Introduction

Lung cancer is one of the most commonly diagnosed malignancies in the world and accounts for the number one incidence and mortality rates among all human cancers [1,2]. Approximately 40–60% of all lung cancer cases show metastasis at their initial diagnosis [3,4] and the secondary tumors are frequently found in the brain, bones, liver, and adrenal glands [5–10]. A large portion (70–90%) of lung cancer patients succumb to the disease as a result of distant tumor metastasis rather than uncontrolled primary tumor growth [11].

Lung adenocarcinoma (LUAD) is the most common subtype of lung cancer and accounts for approximately 40% of all lung cancers. LUAD is a heterogeneous malignant disease and the existence of heterogeneity makes LUAD therapy challenging. Previous studies have widely employed Bulk RNA sequencing (RNA-seq) to clarify the transcriptome of LUAD. However, Bulk RNA sequencing (RNA-seq) only shows the average expression across all cells, but does not reveal the gene expression patterns of individual cells, resulting in the neglect of heterogeneity among individual cells. The use of scRNA-seq could provide a better insight into the heterogeneity of different subgroups of cells. Importantly, this technology has been successfully used to evaluate tumor heterogeneity [12] and has revealed the complexity of the tumor microenvironment [13]. Moreover, recent data using scRNA-seq has shown better identification of tumor heterogeneity, including better illustration of tumor growth, resistance to treatment, and tumor metastasis, as well as understanding of tumor biology [14–17]. In addition, scRNA-seq could help to identify an association between tumor transcriptional heterogeneity and patient prognosis [18,19]. Therefore, the aim of this study was to reveal the transcriptome and heterogeneity between primary LUAD and metastatic LUAD through scRNA-seq and to identify metastasis-associated genes (MAGs) to provide new insights for metastatic LUAD treatment and prognosis judgment.

In this study, a 7-MAG signature was established based on LUAD scRNA-seq data, and this model performed well in predicting the OS of LUAD patients. Impactful information has been presented for prognostic indicators and novel therapeutic targets of metastatic LUAD.

2. Material and methods

2.1. Database search and data acquisition

The LUAD scRNA-seq data on 126 LUAD cell samples (datasets of PDX-LC-PT-45 and PDX-LC-MBT-15) were downloaded from the Gene Expression Omnibus Database (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>) with the accession number GSE69405. The LC-PT-45 tumor was derived from a 60-year-old male, treatment-naïve LUAD patient [20]. The LC-MBT-15 tumor originated from a 57-year-old woman with LUAD heterochronous brain metastasis after standard chemotherapy and erlotinib treatment [20]. Furthermore, the transcriptome profile together with clinicopathological data (including gender, age, tumor grade, TNM stage, pathological stage, and

follow-up data) were obtained from The Cancer Genome Atlas (TCGA) database (<https://portal.gdc.cancer.gov/>). Clinical data for the LUAD patients included in this study are shown in Table 1.

Table 1. Clinicopathological characteristics of LUAD patients.

Characteristics	Training group (n = 236)	Testing group (n = 232)
<i>Age</i>		
≤ 65	111	113
> 65	118	116
<i>Gender</i>		
Male	102	120
Female	134	112
<i>Tumor stage</i>		
T1+T2	201	206
T3+T4	33	25
<i>Lymph node stage</i>		
N0+N1	189	199
N2+N3	40	28
<i>Metastasis stage</i>		
M0	161	154
M1	14	10
<i>Pathological stage</i>		
I+II	175	185
III+IV	57	43
<i>Survival status</i>		
Alive	147	144
Deceased	89	88
<i>MAG risk score</i>		
Low	118	116
High	118	116

2.2. Analysis of scRNA-seq and TCGA data

The LUAD scRNA-seq data of the raw reads from 126 tumor cell samples were generated using the Illumina HiSeq2500 System and mapped to the GRCH36 human genome sequences. A series of data preprocessing, including readings of the LUAD scRNA-seq data matrix and taking the average of duplicate genes, was carried out. The LUAD scRNA-seq data were then converted into a Seurat object and subjected to data filtering. The exclusion criteria were set as: 1) data on cells with fewer than 200 detected genes; 2) genes identified in fewer than three cells; and 3) mitochondrial DNA reads greater than 11%. Log normalization was used to normalize single-cell transcriptome profiling and FindVariableGene was used to assess the variable genes across single cells for downstream analysis. Prior to this, ScaleData was used to scale the data and remove unnecessary variable sources, such as technical noise. Subsequently, a principal component analysis (PCA) was conducted to evaluate the

most significant principal components for cluster analysis. Afterwards, t-Distributed Stochastic Neighbor Embedding (t-NSE) was performed to assess the cluster classification, and FindAllMarkers was used to screen the co-expressed variable genes in the clusters, which were defined as MAGs. An absolute value of logFC ($|\text{Log FC}| > 0.5$) and an adjusted P value ($\text{adj P} < 0.05$) were used as the cut-off values for MAGs. Next, the TCGA-LUAD transcriptome profile normalization was performed using the edgeR package.

2.3. Identification of differentially expressed MAGs (DEMAGs) associated with OS

The MAG expression profile was first extracted from the TCGA-LUAD transcriptome profile and the R package Limma was performed to identify the DEMAGs using $\log\text{FC} > 1.0$ and false discovery rate ($\text{FDR} < 0.01$) as the cut-off criteria. Next, we matched the MAG expression profile with OS of LUAD patients in the entire TCGA-LUAD cohort. The prognostic MAGs were identified through univariate Cox regression analysis when $P \text{ value} < 0.01$ in the entire TCGA-LUAD cohort. The prognostic DEMAGs identified by intersecting the results of (a) the prognostic MAGs and (b) the DEMAGs were used for downstream analysis. MAGs with a hazard ratio ($\text{HR} > 1$) were considered risk factors, whereas MAGs with an $\text{HR} < 1$ were considered protective.

2.4. Construction of a MAG signature and a MAG nomogram model for the prediction of LUAD prognosis

The entire TCGA-LUAD cohort was randomly divided into either the training ($n = 236$) or the testing group ($n = 232$), and the training group was utilized to construct a MAG signature for the prediction of LUAD patient OS. Before constructing a MAG signature, these prognostic DEMAGs were initially visualized through univariate Cox analysis, and a LASSO regression analysis was then implemented to further screen and narrow down these prognostic DEMAGs. Multivariate Cox proportional hazards regression analysis was used to construct a prognostic model in the training group, and the HR and regression coefficient for each prognostic DEMAG were calculated. Eventually, seven DEMAGs were identified to establish the MAG signature for the prediction of OS in LUAD patients. The calculation of the MAG signature was as follows: Risk score $= \sum (\beta_i \times \text{Exp}_i)$, where β_i represented the coefficient of gene i , standing for the weight of gene i , and Exp_i represented the expression level of gene i . Kaplan-Meier curves and log-rank tests were performed to associate this 7-MAG signature with OS of LUAD patients, then a receiver operating characteristic (ROC) curve was used to identify the performance of the MAG signature in predicting OS of LUAD patients in both the training and testing groups, as well as the entire TCGA-LUAD cohort.

Finally, the rms R package was used to construct a nomogram to predict OS of LUAD patients, which incorporated these seven DEMAGs. This nomogram model was a prognostic statistical model made using simple graphs according to previous studies [21,22].

2.5. Functional pathway analysis of the 7-MAG signature

To search for and identify potential molecular mechanisms, Gene Set Enrichment Analysis (GSEA) was performed. The risk scores was defined as the phenotype, and then the entire TCGA-LUAD cohort was divided into either a high- or low-risk group using the median of the risk score as

the cutoff value. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was then used to enrich the functional pathways of these high- and low-risk groups with the $FDR < 0.05$ as the cut-off value for significance.

2.6. Statistical analysis

All statistical analysis were performed using the R package, and glmnet was used for LASSO coefficient regression to screen and narrow down DEMAGs. The Kaplan-Meier with log-rank analysis was performed to evaluate OS of LUAD patients. The generalized linear model (GLM) was established using the rms software package to develop a quantitative model for the prediction of OS in LUAD patients. A P value < 0.05 was considered statistically significant.

3. Results

3.1. Identification of MAGs in LUAD samples using scRNA-seq data

A total of 77 and 49 high-quality tumor cell samples were obtained from PDX-LC-PT-45 and PDX-LC-MBT-15, respectively. The quality control chart was shown in Figure 1A, illustrating the range of single cell RNA numbers, the sequencing count, and the proportion of mitochondrial sequencing counts of each cell. In total, 1500 variable genes were found in the single cell samples (Figure 1B). The PCA method was then used to divide these single cell samples into 20 different components (Figure 1C), in which the statistically significant components were used for further analysis. In addition, the single cell samples were mapped into two independent dimensions based on the PC1 and PC2 (Figure 1D). Apart from performing PCA, the t-NSE algorithm was also completed to successfully divide the single-cell sample data into primary- and metastatic- tumor cell subpopulations, defined as cluster 1 and cluster 0, respectively (Figure 1E). Subsequently, 414 co-expressed genes were identified from these two clusters as MAGs using limma software with $|\text{Log FC}| > 0.5$ and an adj P < 0.05 , and the heatmap data revealed the top 10 genes between these two clusters (Figure 1F).

3.2. Verification of DEMAGs associated with OS in the entire TCGA-LUAD cohort

The 414 MAGs expression profile was obtained from TCGA-LUAD transcriptome profile consisting of 551 samples (497 LUAD samples and 54 normal lung tissue samples). The 414 MAGs expression profiling was uploaded into R software packages to identify DEMAGs, which revealed a total of 114 DEMAGs using the criteria of $\log FC > 1.0$ and $FDR < 0.01$, including 95 upregulated and 19 downregulated MAGs. The heatmap and volcano plot of these DEMAGs were shown in Figure 2A,B.

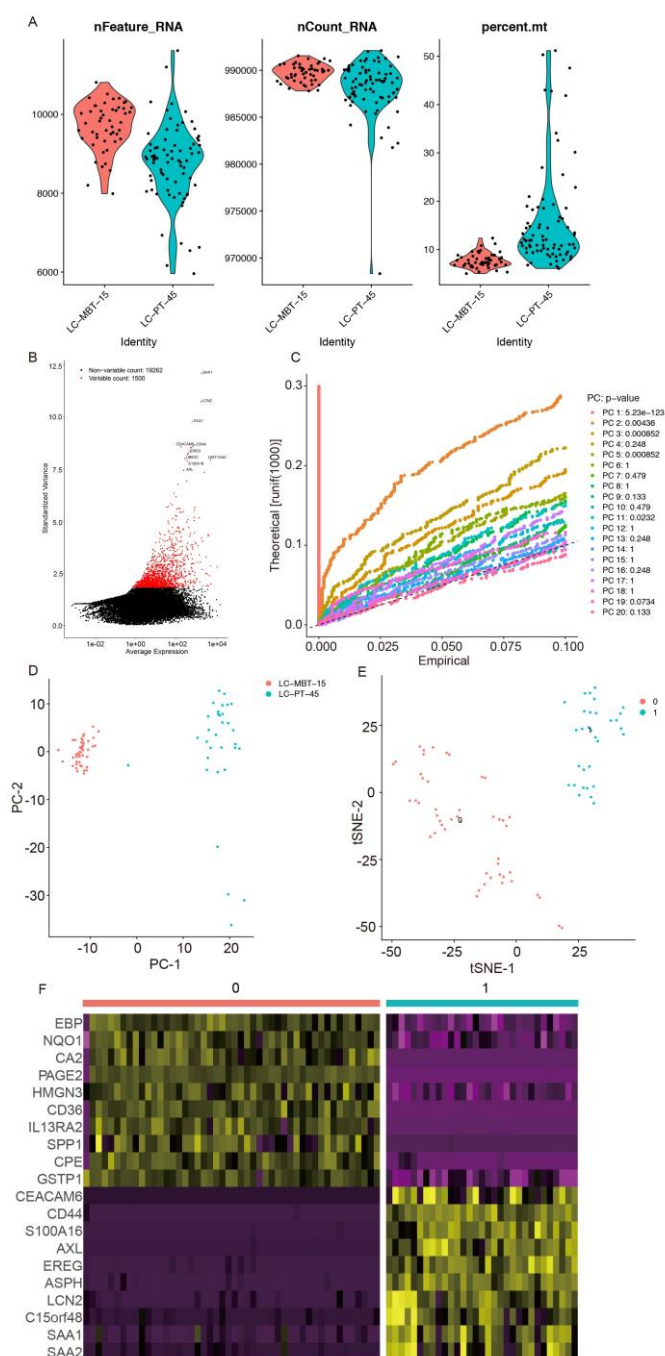


Figure 1. Identification of MAGs using LUAD scRNA-seq data. (A) Quality control of LUAD scRNA-seq data between the two cell sub-populations. (B) The variance diagram. The result revealed positive gene symbols with significant differences across cells. Note: black dot, non-variable counts (19,262 in total); red dot, variable counts (1500 in total). (C) Twenty PCs with estimated P values were identified based on LUAD scRNA-seq data. (D) The cell groups were classified into two categories termed PC1 and PC2 via PCA. (E) Based on the available significant components, the cells were classified into two clusters using the t-SNE algorithm. Cluster 0 is a metastatic tumor cell subpopulation, while Cluster 1 a primary tumor cell subpopulation. (F) Heatmap. The top 10 of 414 MAGs between primary and metastatic LUAD tissues.

To identify prognostic DEMAGs, we initially defined a cohort of 468 LUAD patients with clinical information from TCGA database as the entire TCGA-LUAD cohort. Next, the 414 MAGs expression profile was incorporated with the clinical information from the entire TCGA-LUAD cohort and univariate Cox regression was performed (Supplementary Figure S1) to identify 51 MAGs associated with OS of LUAD patients. Then, the genes that overlapped for both prognostic MAGs and DEMAGs were defined as prognostic DEMAGs. There were 22 prognostic DEMAGs (Figure 3A,C) found in the entire TCGA-LUAD cohort, and the distribution and correlation of these prognostic DEMAGs was shown in Figure 3B and Figure 3D, respectively. LASSO regression analysis was then performed, which identified 12 prognostic DEMAGs for further analysis (Figure 3E,F).

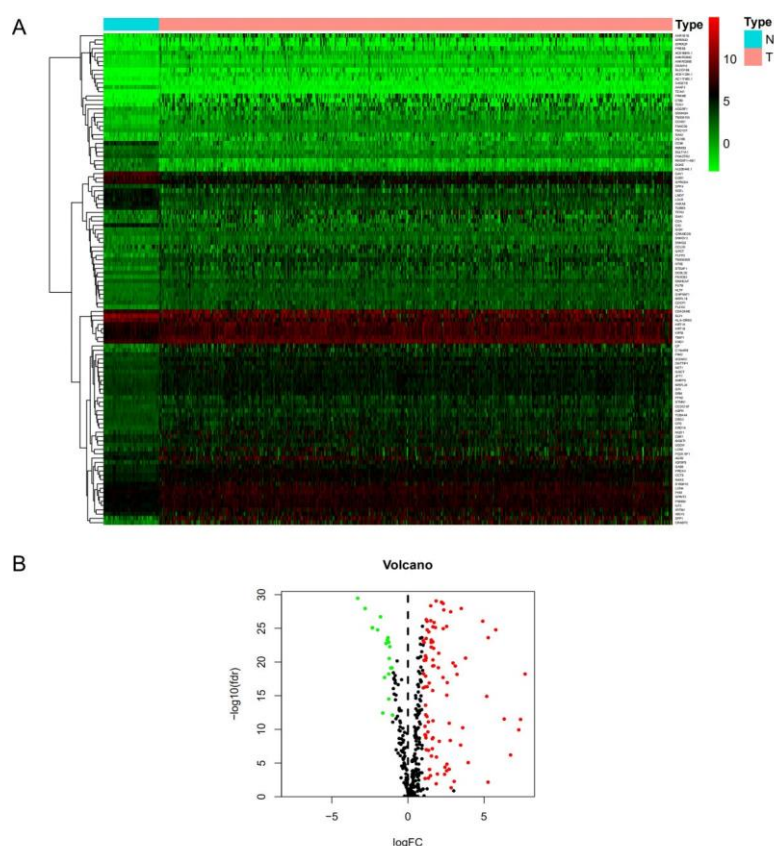


Figure 2. Identification of DEMAGs in 497 LUAD tissues vs. 54 normal lung tissues. (A) Heat map. (B) Volcano plot.

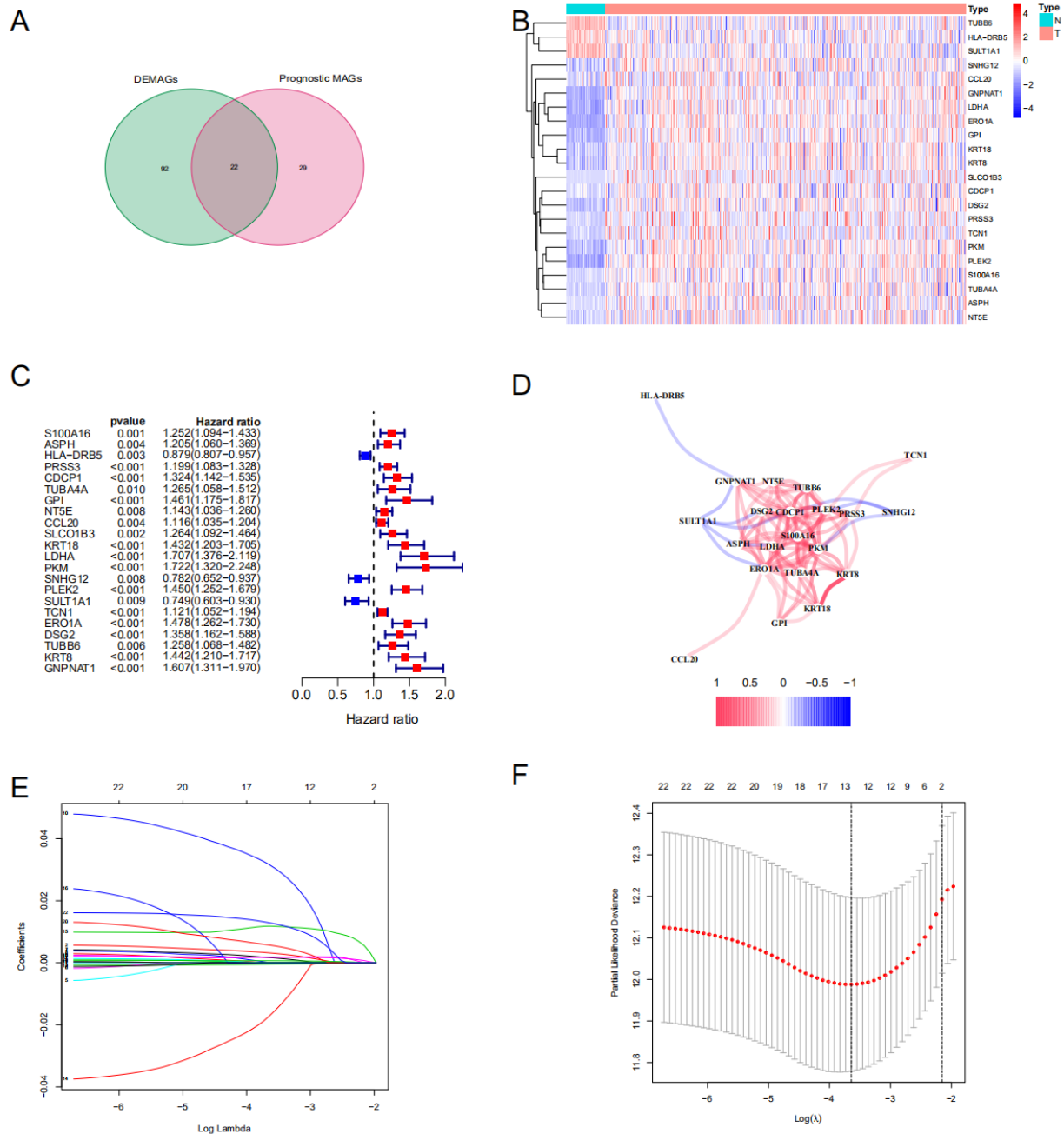


Figure 3. Identification of prognostic DEMAGs in the TCGA-LUAD cohort. (A) Venn diagram. Twenty-two DEMAGs were identified to associate with overall survival (OS) of LUAD patients. (B) Heatmap. Twenty-two DEMAGs between LUAD and matched normal lung tissues. (C) Forest plot. Univariate Cox regression analysis was performed to illustrate the prognostic effect of these 22 DEMAGs on OS of LUAD patients. (D) Correlation network of 22 DEMAGs are illustrated, and the different colors represent different correlation coefficients. (E) The LASSO coefficient values of the 22 selected DEMAGs. (F) The plot of the tuning parameter selection of the LASSO regression. The λ is the tuning parameter.

3.3. Identification of the MAG signature in prediction of LUAD prognosis

The 12 prognostic DEMAGs were further analyzed using multivariate Cox proportional hazards regression analysis and seven genes of interest were identified in the training group (n = 236): serine protease 3 (PRSS3), glucose-6-phosphate isomerase (GPI), chemokine ligand 20 (CCL20), keratin-18 (KRT18), transcobalamin I (TCN1), solute carrier organic anion transporter family member 1B3 (SLCO1B3), and glucosamine-phosphate N-acetyltransferase 1 (GNPNAT1). The MAG signature (Figure 4) to predict the OS of LUAD patients was calculated as follows: Risk score = $\text{Exp (PRSS3)} \times 0.1263 + \text{Exp (GPI)} \times 0.3664 + \text{Exp (CCL20)} \times 0.0889 + \text{Exp (KRT18)} \times 0.3787 + \text{Exp (TCN1)} \times 0.1570 + \text{Exp (SLCO1B3)} \times 0.1756 + \text{Exp (GNPNAT1)} \times 0.3249$. The training group (n = 236) was divided into a high- or low-risk group according to the median risk score as the cut-off value for each sample (Figure 5A). As shown in Figure 5A, the distribution of the vitals demonstrated that the high-risk group had more cases of death compared to the low-risk group, and LUAD patients with high risk scores tended to express a higher level of PRSS3, GPI, CCL20, KRT18, TCN1, SLCO1B3 and GNPNAT1 than those with low risk scores. The Kaplan-Meier curve analysis revealed that LUAD patients with high risk scores had significantly worse OS than those with low risk scores ($P = 7.947\text{e-}05$; Figure 5B). The ROC curve also illustrated that the area under curve (AUC) of the MAG signature was 0.767 (Figure 5C), which indicated a moderate prediction value.

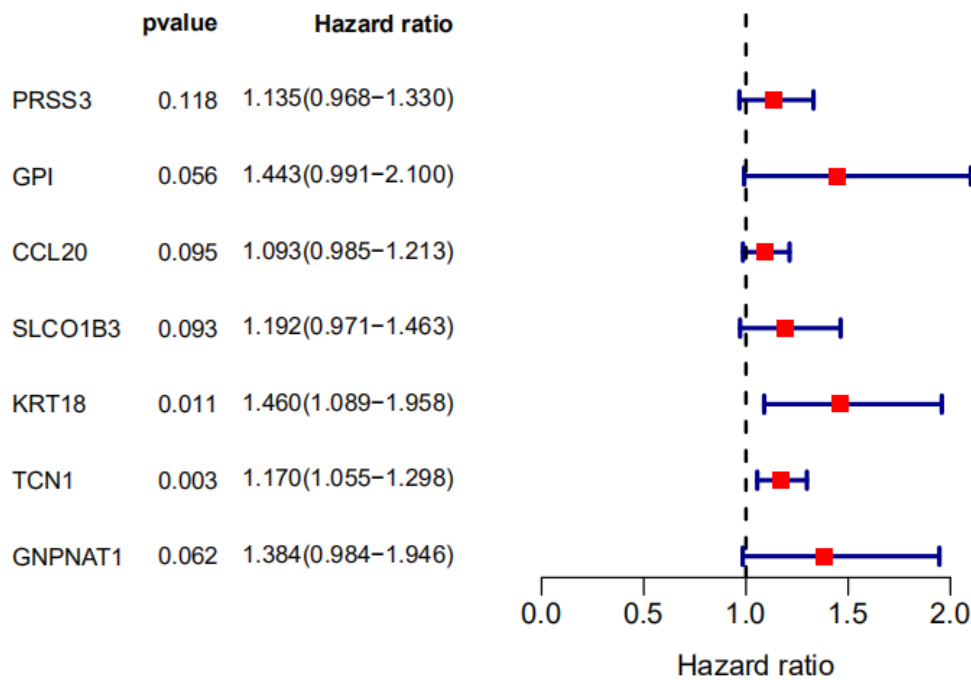


Figure 4. Construction of the 7-MAG signature using multivariate Cox regression analysis.

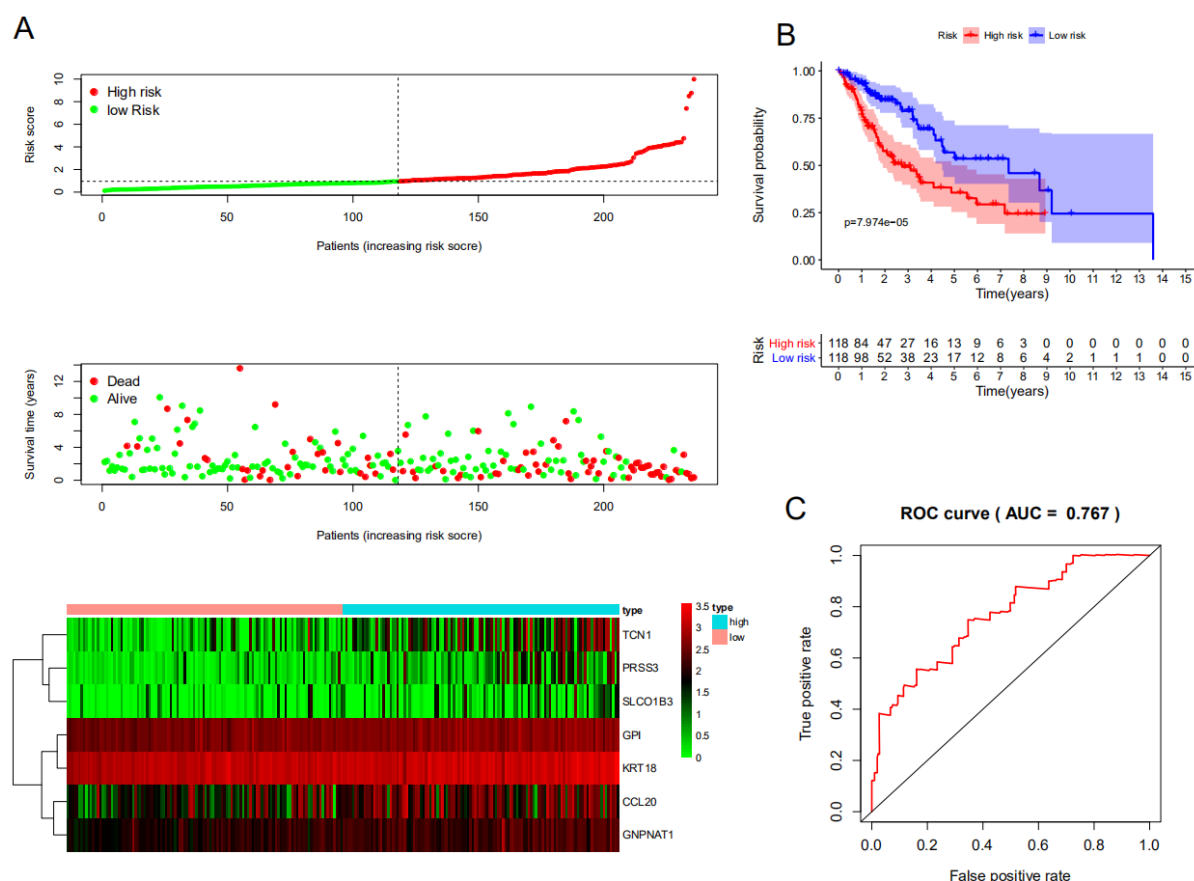


Figure 5. Identification of the 7-MAG signature in the training group ($n = 236$). (A) The risk score, survival distribution, and heat maps of LUAD patients stratified according to the 7-MAG signature. (B) Kaplan-Meier curves. The overall survival (OS) of LUAD patients was analyzed between high- and low-risk scores. (C) The receive operator characteristic (ROC) curve. The prognostic value of the 7-MAG signature was evaluated using the ROC curve.

3.4. Validation of the MAG signature in prediction of LUAD prognosis in the testing group and the entire TCGA-LUAD cohort

The 7-MAG signature to predict OS of LUAD patients was validated in the testing group ($n = 232$) and the entire TCGA-LUAD cohort ($n = 468$). The 7-MAG signature was able to effectively distinguish LUAD patients into two groups with better or worse OS in the testing group ($P = 9.465e-03$, $AUC = 0.682$; Figure 6A–C). Similarly, the entire TCGA-LUAD cohort also further confirmed the results, and LUAD patients with high risk scores had a shorter OS than those with low risk scores ($P = 1.47e-05$, $AUC = 0.727$; Figure 6D–F). Univariate- and multivariate Cox analysis of the 7-MAG signature in the entire TCGA-LUAD cohort confirmed it to be an independent prognostic predictor for LUAD patients (Figure 7).

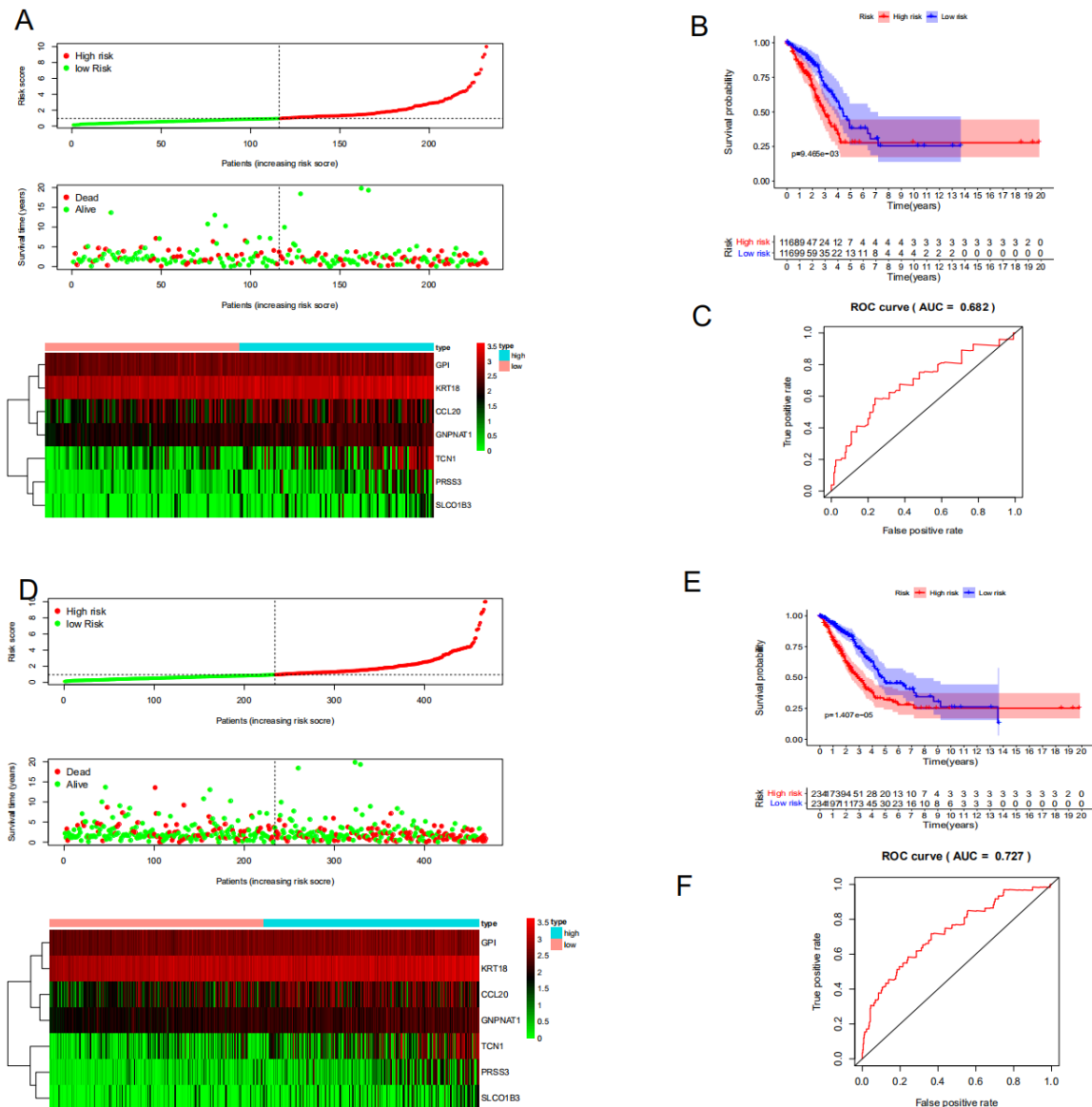


Figure 6. The performance of the 7-MAG signature in the testing group ($n = 232$) and the entire TCGA-LUAD cohort ($n=468$). (A, D) The risk score, survival distribution, and heat maps of LUAD patients stratified according to the 7-MAG signature. (B, E) Kaplan-Meier curves. Kaplan-Meier curves were used to analyze the survival significance stratified by the high- and low-risk scores. (C, F) The receive operator characteristic (ROC) curves. The prognostic value of the 7-MAG signature was evaluated using the ROC curves.

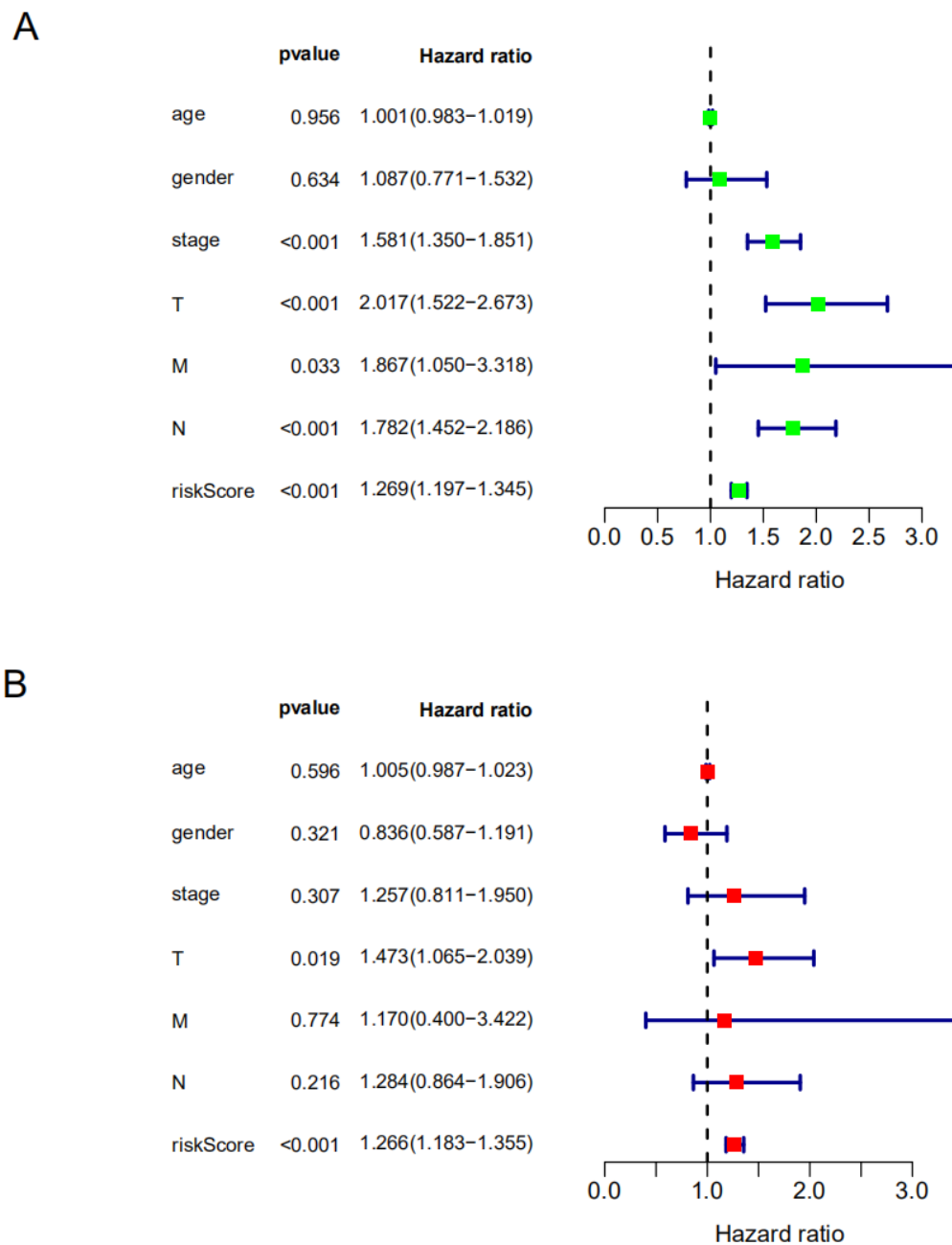


Figure 7. The univariate and multivariate Cox regression analyses of the 7-MAG signature against the clinicopathological data. (A) The univariate Cox regression analysis. (B) The multivariate Cox regression analysis.

3.5. Construction of a MAG nomogram model in accordance with the 7-MAG signature

A quantitative model of the seven DEMAGs to predict OS of LUAD patients was developed by integrating the seven DEMAGs into a nomogram (Figure 8). In the nomogram model, we first assigned points to each variable using a point scale based on the multivariate Cox analysis. Next, a horizontal

line was drawn to determine the point of each variable, and the total points of each LUAD patient, which were distributed between 0 and 100, was calculated by summing the points of all variables. Finally, a vertical line between the total points axis and each prognostic axis was utilized to evaluate the 1-, 2-, and 3-year OS of LUAD patients.

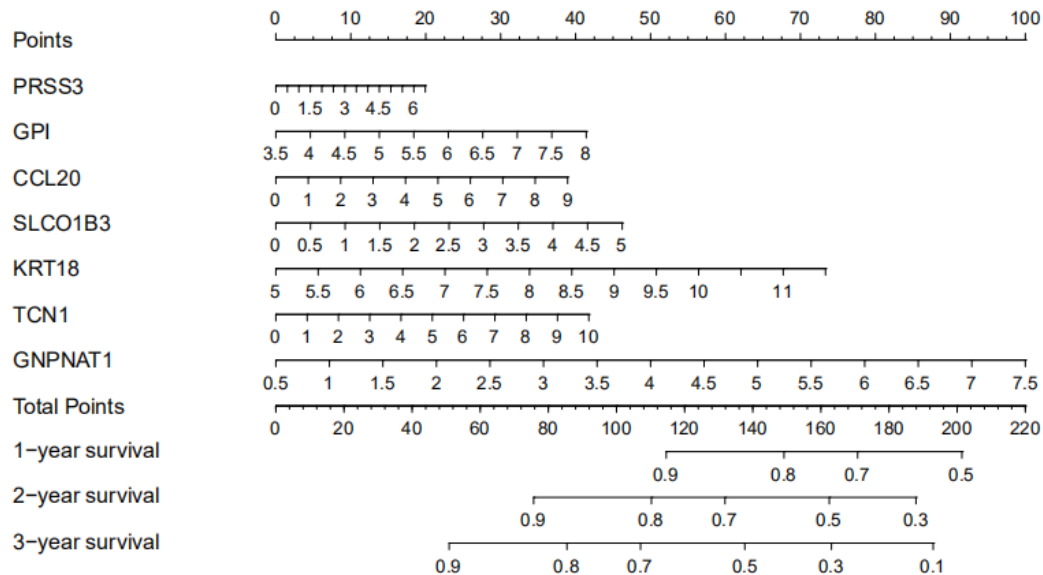


Figure 8. The nomogram model. The nomogram model was constructed to predict the 1-, 2-, and 3-year survival of LUAD patients.

3.6. Bioinformatical enrichment of the 7-MAG signature in the entire TCGA-LUAD cohort

To gain specific biological insights of the 7-MAG signature in LUAD, the GSEA was performed that revealed alterations of the cell cycle, DNA replication, mismatch repair, pentose phosphate pathway, proteasome, and the p53 signaling pathway were significantly enriched in LUAD samples with high risk scores, whereas the vascular smooth muscle contraction was significantly enriched in LUAD samples with low risk scores ($FDR < 0.05$; Figure 9).

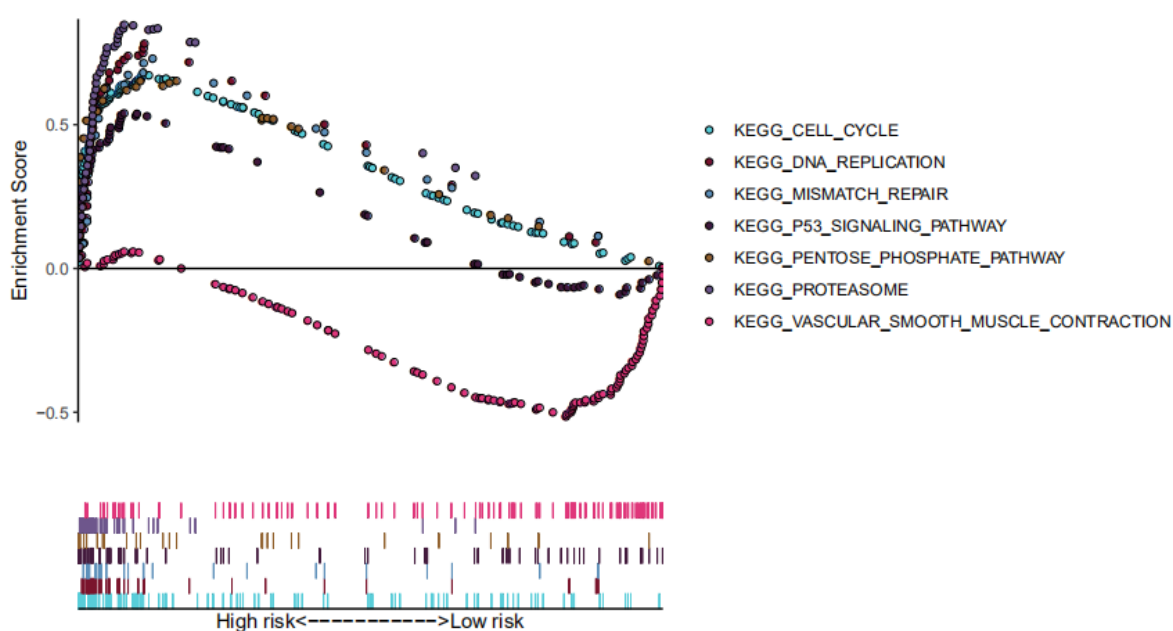


Figure 9. Bioinformatic analysis of the MAG signature in LUAD patients (n = 468). The gene set enrichment analysis (GSEA) was performed to significantly enrich the biological processes of the MAG signature with the high- and low-risk scores.

4. Discussion

LUAD is a heterogeneously malignant disease and often shows early-stage tumor metastasis leading to a poor prognosis. scRNA-seq technology could help to effectively dissect tumor cell heterogeneity and identify potential prognosis biomarkers. Therefore, we first screened LUAD scRNA-seq data to identify 414 MAGs from the GSE69405 and found 22 prognostic DEMAGs from the entire TCGA-LUAD cohort. After that, we successfully developed a 7-MAG signature, which could serve as novel indicator for prognosis of LUAD patients and provide potential novel therapeutic targets and molecular mechanism for metastatic LUAD.

To attain potential MAGs, we performed PCA and tNSE analysis on scRNA-seq data of primary LUAD and metastatic LUAD. At the same time, we characterized the transcriptome and heterogeneity between primary and metastatic LUADs and identified a total of 414 MAGs from the GSE69405 and combined with LUAD Bulk RNA-data for subsequent analysis. In the entire TCGA-LUAD cohort, 114 DEMAGs were identified from 414 MAGs using the Limma package in accordance to $\log_{2}FC > 1.0$ and $FDR < 0.01$, and 52 out of 414 MAGs were identified to be associated with OS of LUAD patients via univariate Cox analysis, among which 22 genes were considered as prognostic DEMAGs. To ensure the construction of a MAG signature that was more reasonable and reliable, LASSO regression analysis was executed to narrow down these prognostic DEMAGs. Finally, a 7-MAG signature, including PRSS3, GPI, CCL20, KRT18, TCN1, SLCO1B3, and GNPAT1, was established through multivariate Cox proportional hazards regression analysis in the training group. To elucidate the prognostic value of the 7-MAG signature, the Kaplan-Meier with log-rank analysis and ROC curve analysis confirmed that the 7-MAG signature showed a good predictive ability for OS of LUAD

patients in the training ($n = 236$, $P = 7.947 \times 10^{-5}$, $AUC = 0.767$; Figure 5A–C) and the testing ($n = 232$, $P = 9.465 \times 10^{-3}$, $AUC = 0.682$; Figure 6A–C) groups, as well as the entire TCGA-LUAD cohort ($n = 468$, $P = 1.47 \times 10^{-5}$, $AUC = 0.727$; Figure 6D–F). Additionally, univariate- and multivariate Cox analysis also demonstrated that the 7-MAG signature was an independent factor in the prediction of LUAD prognosis. Importantly, the nomogram based on the 7-MAG signature was established to more intuitively help the prediction of OS in LUAD patients at one, two, and three years. Finally, to show validity of the MAG signature, GSEA illustrated that the cell cycle, DNA replication, mismatch repair, pentose phosphate pathway, proteasome, and the p53 signaling pathway were involved in progression of LUAD.

Importantly, components of the 7-MAG signature possess different biological functions and various patterns of expression in different human cancers. For example, PRSS3 was reported to be involved in tumor metastasis in prostate, pancreatic, and gastric cancers [23–25], while Ma et al. revealed that PRSS3 could predict prognosis of LUAD patients and promote growth and invasion in LUAD cells [26]. Ma *et al.* showed that knockdown of GPI inhibited cancer cell proliferation, invasion, and migration in gastric cancer [27], whereas the high GPI expression was associated with NSCLC metastatic potential [28]. CCL20 was shown to be expressed in various human tissues, including the lung, liver, and lymph nodes [29–31]. Moreover, CCL20 could mediate the migration of epithelial cells and likely participates in cancer cell migration and metastasis in a variety of human cancers, such as breast, colorectal, prostate, and pancreatic cancers [32–35]. Similarly, Wang et al. indicated that CCL20 was significantly overexpressed in NSCLC tissues, and it contributed to cancer cell proliferation and migration through the PI3K pathway in lung cancer [36]. KRT18, also known as cytokeratin 18 (CK18), was shown to be abnormally expressed in various human cancers and has been associated with poor disease progression and prognosis [37,38]. For instance, Ma et al. reported that KRT18 was highly expressed and correlated with poor prognosis in NSCLC, while KRT18 knockdown was prone to inhibit NSCLC cell migration [39]. TCN1 is a type of vitamin B12-binding protein that transports vitamin B12 from the stomach to the intestine. It was reported that TCN1 overexpression was correlated with poor biological behavior and tumorigenesis of various tumor tissues [40]. Liu et al. found that TCN1 was overexpressed in colon cancer, and TCN1 overexpression was a poor prognostic biomarker and predicted neoadjuvant chemosensitivity in colon cancer [41]. Moreover, SLCO1B3, also known as the organic anion transporting polypeptide (AOTP), localizes at chromosome 12p12-31.7 to 12p12-37.2 and plays important roles in transporting various components to cells. Recent studies have shown that a truncated form of SLCO1B3 occurred in human cancer tissues and cells lines [42–44]. Hase et al. demonstrated that SLCO1B3 promoted NSCLC progression via mediating epithelial-mesenchymal transition [45]. Finally, GNPAT1 is a member of the GCN5-related N-acetyltransferase superfamily. Kaushik et al. reported that genetic loss-of-function of GNPAT1 in castration-resistant prostate cancer (CRPC)-like cells contributed to proliferation and increased tumor cell aggressiveness through the PI3K/AKT signaling pathway [46]. Taken together, these seven genes have been shown to be involved in the occurrence and progression of various human cancers.

In the current study, it was confirmed that the 7-MAG signature possessed the ability to predict LUAD prognosis. However, our current study did have some limitations. For example, we only assessed the available online data but did not use data from our own patients; thus, it is a retrospective study. Future prospective research is warranted to verify our results.

5. Conclusions

In the current study, we characterized the transcriptome and heterogeneity between primary LUAD and metastatic LUAD based on scRNA-seq, and 414 MAGs were identified from LUAD scRNA-seq data and a 7-MAG signature was established in the training group. The 7-MAG signature was able to predict OS of LUAD patients in the training group, the testing group, the entire TCGA-LUAD cohort and the nomogram model. Furthermore, the GSEA results revealed that LUAD progression was due to alterations in the cell cycle, DNA replication, mismatch repair, pentose phosphate pathway, the proteasome, and the p53 signaling pathway. Future studies will need to verify the usefulness of the 7-MAG signature for the prediction of LUAD prognosis and for the potential to target these biomarkers as novel strategies to control LUAD with metastasis.

Data availability

All data in this study are already available from the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) and The Cancer Genome Atlas (<https://portal.gdc.cancer.gov/>) databases.

Conflicts of interest

The authors declare that there is no conflict of interest in this work.

References

1. R. J. Scheff, B. J. Schneider, Non-small-cell lung cancer: treatment of late stage disease: chemotherapeutics and new frontiers, in *Seminars in interventional radiology*, **30** (2013), 191–198.
2. L. L. Humphrey, M. Deffebach, M. Pappas, C. Baumann, K. Artis, J. P. Mitchell, et al., Screening for lung cancer with low-dose computed tomography: a systematic review to update the US Preventive services task force recommendation, *Ann. Intern. Med.*, **159** (2013), 411–420.
3. C. A. Ridge, A. M. McErlean, M.S. Ginsberg, Epidemiology of lung cancer, in *Seminars in interventional radiology*, **30** (2013), 93–98.
4. H. Satoh, K. Kurishima, R. Nakamura, H. Ishikawa, K. Kagohashi, G. Ohara, et al., Lung cancer in patients aged 80 years and over, *Lung Cancer*, **65** (2009), 112–118.
5. N. L. Kibrinsky, M. G. Klug, P. J. Hokanson, D. E. Sjolander, L. Burd, Impact of smoking on cancer stage at diagnosis, *J. Clin. Oncol.*, **21** (2003), 907–913.
6. J. Olak, Surgical strategies for metastatic lung cancer, *Surg. Oncol. Clin.*, **8** (1999), 245–257.
7. J. Pfannschmidt, H. Dienemann, Surgical treatment of oligometastatic non-small cell lung cancer, *Lung Cancer*, **69** (2010), 251–258.
8. H. Ishikawa, H. Satoh, K. Kurishima, Y. T. Yamashita, M. Ohtsuka, K. Sekizawa, Lung cancer with synchronous brain and bone metastasis, *Clin. Oncol.*, **12** (2000), 136–137.
9. A. Oikawa, H. Takahashi, H. Ishikawa, K. Kurishima, K. Kagohashi, H. Satoh, Application of conditional probability analysis to distant metastases from lung cancer, *Oncol. Lett.*, **3** (2012), 629–634.

10. T. Tamura, K. Kurishima, H. Watanabe, T. Shiozawa, K. Nakazawa, H. Ishikawa, et al., Characteristics of clinical N0 metastatic non-small cell lung cancer, *Lung Cancer*, **89** (2015), 71–75.
11. S. L. Wood, M. Pernemalm, P. A. Crosbie, A. D. Whetton, The role of the tumor-microenvironment in lung cancer-metastasis and its relationship to potential therapeutic targets, *Cancer Treat. Rev.*, **40** (2014), 558–566.
12. N. E. Navin, The first five years of single-cell cancer genomics and beyond, *Genome Res.*, **25** (2015), 1499–1507.
13. A. A. Powell, A. H. Talasz, H. Zhang, M. A. Coram, A. Reddy, G. Deng, et al., Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines, *PLoS One*, **7** (2012), e33788.
14. H. Gong, Y. Li, Y. Yuan, W. Li, H. Zhang, Z. Zhang, et al., EZH2 inhibitors reverse resistance to gefitinib in primary EGFR wild-type lung cancer cells, *BMC Cancer*, **20** (2020), 1189.
15. Y. Liu, G. Ye, L. Huang, C. Zhang, Y. Sheng, B. Wu, et al., Single-cell transcriptome analysis demonstrates inter-patient and intra-tumor heterogeneity in primary and metastatic lung adenocarcinoma, *Aging*, **12** (2020), 21559–21581.
16. D. He, D. Wang, P. Lu, N. Yang, Z. Xue, X. Zhu, Single-cell RNA sequencing reveals heterogeneous tumor and immune cell populations in early-stage lung adenocarcinomas harboring EGFR mutations, *Oncogene*, **40** (2021), 355–368.
17. Z. Chen, M. Zhao, M. Li, Q. Sui, Y. Bian, J. Liang, et al., Identification of differentially expressed genes in lung adenocarcinoma cells using single-cell RNA sequencing not detected using traditional RNA sequencing and microarray, *Lab. Invest.*, **100** (2020), 1318–1329.
18. A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, et al., Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, *Science*, **344** (2014), 1396–1401.
19. I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, et al., Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq, *Science*, **352** (2016), 189–196.
20. K. T. Kim, H. W. Lee, H. O. Lee, S. C. Kim, Y. J. Seo, W. Chung, et al., Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells, *Genome Biol.*, **16** (2015), 127.
21. A. Iasonos, D. Schrag, G. V. Raj, K. S. Panageas, How to build and interpret a nomogram for cancer prognosis, *J. Clin. Oncol.*, **26** (2008), 1364–1370.
22. V. P. Balachandran, M. Gonen, J. J. Smith, R. P. DeMatteo, Nomograms in oncology: more than meets the eye, *Lancet Oncol.*, **16** (2015), e173–180.
23. G. Jiang, F. Cao, G. Ren, D. Gao, V. Bhakta, Y. Zhang, et al., PRSS3 promotes tumour growth and metastasis of human pancreatic cancer, *Gut*, **59** (2010), 1535–1544.
24. A. Hockla, E. Miller, M. A. Salameh, J. A. Copland, D. C. Radisky, E. S. Radisky, PRSS3/mesotrypsin is a therapeutic target for metastatic prostate cancer, *Mol. Cancer Res.*, **10** (2012), 1555–1566.
25. F. Wang, Y. L. Hu, Y. Feng, Y. B. Guo, Y. F. Liu, Q. S. Mao, et al., High-level expression of PRSS3 correlates with metastasis and poor prognosis in patients with gastric cancer, *J. Surg. Oncol.*, **119** (2019), 1108–1121.

26. C. H. Hsu, C.W. Hsu, C. Hsueh, C. L. Wang, Y. C. Wu, C. C. Wu, et al., Identification and Characterization of Potential Biomarkers by Quantitative Tissue Proteomics of Primary Lung Adenocarcinoma, *Mol. Cell. Proteomics*, **15** (2016), 2396–2410.
27. Y. T. Ma, X. F. Xing, B. Dong, X. J. Cheng, T. Guo, H. Du, et al., Higher autocrine motility factor/glucose-6-phosphate isomerase expression is associated with tumorigenesis and poorer prognosis in gastric cancer, *Cancer Manag. Res.*, **10** (2018), 4969–4980.
28. Y. Dobashi, H. Watanabe, Y. Sato, S. Hirashima, T. Yanagawa, H. Matsubara, et al., Differential expression and pathological significance of autocrine motility factor/glucose-6-phosphate isomerase expression in human lung carcinomas, *J. Pathol.*, **210** (2006), 431–440.
29. W. Xiao, Z. Jia, Q. Zhang, C. Wei, H. Wang, Y. Wu, Inflammation and oxidative stress, rather than hypoxia, are predominant factors promoting angiogenesis in the initial phases of atherosclerosis, *Mol. Med. Rep.*, **12** (2015), 3315–3322.
30. K. Hieshima, T. Imai, G. Opdenakker, J. Van Damme, J. Kusuda, H. Tei, et al., Molecular cloning of a novel human CC chemokine liver and activation-regulated chemokine (LARC) expressed in liver. Chemotactic activity for lymphocytes and gene localization on chromosome 2, *J. Biol. Chem.*, **272** (1997), 5846–5853.
31. C. A. Power, D. J. Church, A. Meyer, S. Alouani, A. E. Proudfoot, I. Clark-Lewis, et al., Cloning and characterization of a specific receptor for the novel CC chemokine MIP-3alpha from lung dendritic cells, *J. Exp. Med.*, **186** (1997), 825–835.
32. A. Muscella, C. Vetrugno, S. Marsigliante, CCL20 promotes migration and invasiveness of human cancerous breast epithelial cells in primary culture. *Mol. Carcinog.*, **56** (2017), 2461–2473.
33. S. Brand, T. Olszak, F. Beigel, J. Diebold, J. M. Otte, S. T. Eichhorst, et al., Cell differentiation dependent expressed CCR6 mediates ERK-1/2, SAPK/JNK, and Akt signaling resulting in proliferation and migration of colorectal cancer cells, *J. Cell Biochem.*, **97** (2006), 709–723.
34. K. Beider, M. Abraham, M. Begin, H. Wald, I.D. Weiss, O. Wald, et al., Interaction between CXCR4 and CCL20 pathways regulates tumor growth, *PLoS One*, **4** (2009), e5125.
35. G. Z. Wang, X. Cheng, X. C. Li, Y. Q. Liu, X. Q. Wang, X. Shi, et al., Tobacco smoke induces production of chemokine CCL20 to promote lung cancer, *Cancer Lett.*, **363** (2015), 60–70.
36. B. Wang, L. Shi, X. Sun, L. Wang, X. Wang, C. Chen, Production of CCL20 from lung cancer cells induces the cell migration and proliferation through PI3K pathway, *J. Cell. Mol. Med.*, **20** (2016), 920–929.
37. Y. C. Lai, C. C. Cheng, Y. S. Lai, Y. H. Liu, Cytokeratin 18-associated Histone 3 Modulation in Hepatocellular Carcinoma: A Mini Review, *Cancer Genomics Proteomics*, **14** (2017), 219–223.
38. A. M. Fortier, E. Asselin, M. Cadrin, Keratin 8 and 18 loss in epithelial cancer cells increases collective cell migration and cisplatin sensitivity through claudin1 up-regulation, *J. Biol. Chem.*, **288** (2013), 11555–11571.
39. B. Zhang, J. Wang, W. Liu, Y. Yin, D. Qian, H. Zhang, et al., Cytokeratin 18 knockdown decreases cell migration and increases chemosensitivity in non-small cell lung cancer, *J. Cancer Res. Clin. Oncol.*, **142** (2016), 2479–2487.
40. M. Martinelli, L. Scapoli, G. Mattei, G. Ugolini, I. Montroni, D. Zattoni, et al., A candidate gene study of one-carbon metabolism pathway genes and colorectal cancer risk, *Br. J. Nutr.*, **109** (2013), 984–989.

41. G. J. Liu, Y. J. Wang, M. Yue, L. M. Zhao, Y. D. Guo, Y. P. Liu, et al., High expression of TCN1 is a negative prognostic biomarker and can predict neoadjuvant chemosensitivity of colon cancer, *Sci. Rep.*, **10** (2020), 11951.
42. M. Nagai, T. Furihata, S. Matsumoto, S. Ishii, S. Motohashi, I. Yoshino, et al., Identification of a new organic anion transporting polypeptide 1B3 mRNA isoform primarily expressed in human cancerous tissues and cells, *Biochem. Biophys. Res. Commun.*, **418** (2012), 818–823.
43. N. Thakkar, K. Kim, E. R. Jang, S. Han, K. Kim, D. Kim, et al., A cancer-specific variant of the SLCO1B3 gene encodes a novel human organic anion transporting polypeptide 1B3 (OATP1B3) localized mainly in the cytoplasm of colon and pancreatic cancer cells, *Mol. Pharm.*, **10** (2013), 406–416.
44. T. Furihata, Y. Sun, K. Chiba, Cancer-type Organic Anion Transporting Polypeptide 1B3: Current Knowledge of the Gene Structure, Expression Profile, Functional Implications and Future Perspectives, *Curr. Drug Metab.*, **16** (2015), 474–485.
45. H. Hase, M. Aoki, K. Matsumoto, S. Nakai, T. Nagata, A. Takeda, et al., Cancer type-SLCO1B3 promotes epithelial-mesenchymal transition resulting in the tumour progression of non-small cell lung cancer, *Oncol. Rep.*, **45** (2021), 309–316.
46. A. K. Kaushik, A. Shojaie, K. Panzitt, R. Sonavane, H. Venghatakrishnan, M. Manikkam, et al., Inhibition of the hexosamine biosynthetic pathway promotes castration-resistant prostate cancer, *Nat. Commun.*, **7** (2016), 11612.



AIMS Press

©2021 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)