Machine learning-based screening of the diagnostic genes and their relationship with immune-cell infiltration in patients with lung adenocarcinoma

Shuying Wang[1,2#], Qiong Wang[3#], Bin Fan[2#], Jiao Gong[4], Liping Sun[2], Bo Hu[4], Deqing Wang[1,2]

[1]Department of Laboratory Medicine, Medical School of Chinese PLA, Beijing, China; [2]Department of Blood Transfusion Medicine, The First Medical Center of PLA general Hospital, Beijing, China; [3]Department of Pathology, The First Medical Center of PLA General Hospital, Beijing, China; [4]Department of Laboratory Medicine, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

Contributions: (I) Conception and design: S Wang, Q Wang, B Fan, D Wang, L Sun, B Hu; (II) Administrative support: S Wang, Q Wang, B Fan, D Wang, L Sun, B Hu; (III) Provision of study materials or patients: S Wang, Q Wang; (IV) Collection and assembly of data: B Fan, S Wang, Q Wang, J Gong; (V) Data analysis and interpretation: S Wang, Q Wang, J Gong; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Deqing Wang. Department of Blood Transfusion Medicine, The First Medical Center of PLA General Hospital, Beijing 100853, China. Email: deqingw@vip.sina.com; Bo Hu. Department of Laboratory Medicine, The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510635, China. Email: hubo@mail.sysu.edu.cn; Liping Sun. Department of Blood Transfusion Medicine, The First Medical Center of PLA General Hospital, Beijing 100853, China. Email: sunlp2012@sina.com.

Background: Lung adenocarcinoma (LUAD) is the most common type of lung cancer, and has a dismal mortality rate of 80%, mainly due to diagnosis at an advanced stage. Biomarkers with high specificity and sensitivity for the early diagnosis of LUAD are sparse. This study aimed to identify markers for the early diagnosis of LUAD.

Methods: The GSE32863 and GSE75037 data sets were standardized and merged to screen for differentially expressed genes (DEGs). Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were conducted. The intersected DEGs from the least absolute shrinkage and selection operator (LASSO) and support vector machine (SVM) regression analyses were considered the hub genes. Then the diagnostic ability and expression of hub genes was tested in GSE63459 data set, Finally, CIBERSORT was used to analyze the correlation between the immune-infiltrating cells and hub genes.

Results: The following 7 DEGs were intersected by the LASSO and SVM regression analyses: Locus 401286 (*LOC401286*), flavin-containing monooxygenase 2 (*FMO2*), *XLKD1*, Ras homolog family member J (*RHOJ*), scavenger receptor Class A member 5 (*SCARA5*), heat shock protein beta-2 (HSPB2), and serine incorporator 2 (*SERINC2*). The area under the receiver operating characteristic curve (AUC) of *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, *HSPB2*, and *SERINC2* was 0.99, 1.00, 0.99, 1.00, 0.99, 0.99, and 0.98, respectively in the training groups. The AUC of *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, *HSPB2*, and *SERINC2* was 0.97, 0.96, 0.94, 0.88, 0.85, 0.94 and 0.89, respectively in the validation group. The immune-cell infiltrations of naive B cells, memory B cells, plasma cells, naive cluster of differentiation (CD) 4 T cells, T follicular helper cells, regulatory T cells, gamma delta T cells, monocytes, M0 macrophages, M1 macrophages, resting mast cells, activated mast cells, and neutrophils were different between the normal and tumor tissues. Notably, these immune cells were correlated with the above-mentioned 7 diagnostic genes.

Conclusions: We identified 7 DEGs in LUAD tissue that can be considered diagnostic genes based on 2 machine-learning regression methods, which could be very helpful for the early diagnosis of LUAD in clinical practice.

Keywords: Lung adenocarcinoma (LUAD); immune-cell infiltration; diagnosis; bioinformatic analysis; machine learning

Submitted Jan 30, 2022. Accepted for publication Mar 16, 2022.
doi: 10.21037/jtd-22-206
View this article at: https://dx.doi.org/10.21037/jtd-22-206

Introduction

It is widely acknowledged that early detection and treatment can improve patient outcomes for any disease, and cancer is no exception. Lung adenocarcinoma (LUAD) is the most common type of lung cancer, and has a dismal mortality rate of 80% (1). Significant progress in the screening and diagnostic methods, such as computed tomography (CT) imaging, has been made in recent years (2). However, most patients still miss the optimal therapeutic window, as they are only diagnosed at an advanced stage (3). Previously reported non-invasive approaches for early diagnosis of LUAD included microRNAs (4), DNA methylation markers (5), and autoantibody combined with CT (6). Nevertheless, biomarkers with high specificity, simplicity, and convenience for test in clinical practice are sparse. Thus, novel biomarkers need to be explored and identified.

Immune cells and the immune response have been shown to play very important roles in the occurrence and development of LUAD (7,8). As reported in the literature, tumor cells and immune cells interact with the tumor microenvironment (TME) and affect tumorigenesis (8,9). Notably, different levels of immune-cell infiltration have various effects on prognosis (10). In recent years, machine learning has been used to screen diagnostic genes, which has the ability to decipher complicated connections between multiple sets of test data and diseases (11). However, in some studies, the screened genes were not associated with the absolute value and proportion of the infiltrating immune cells (12,13) and only 1 type of machine-learning method was used (14).

The aim of the present study is to screen the diagnostic genes and analyze their relationship with immune-cell infiltration based on machine-learning in patients with lung adenocarcinoma. We hypotheses that novel diagnostic genes for LUAD could be identified. We present the following article in accordance with the STARD reporting checklist (available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-206/rc).

Methods

Data download and preliminary process

The Gene Expression Omnibus (GEO) data sets GSE32863 and GSE75037 were downloaded, normalized, and merged using R packages "limma" and "Sva." The differential expression analysis was conducted on the merged data using the screening criteria |logFC| >2, and an adjusted P value <0.05. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Functional analysis

The differentially expressed genes (DEGs) were analyzed by GO, KEGG and GSEA R packages with "clusterProfiler", "org.Hs.eg.db" and "c5.go.v7.4.symbols.gmt". GO (http://geneontology.org) is a standard recognized classification system for defining the biological processes (BPs), molecular functions (MFs), and cellular components (CCs) of DEGs (15). The KEGG (https://www.kegg.jp/) is a database that provides a manual curation of the pathways associated with genes (16). The screening conditions for the GO annotation and KEGG analysis included P values <0.05 and adjusted P values <0.05. The enrichment of the upregulated or downregulated sets of genes from the REACTOME pathway database was computed by running GSEA against the fold-change ranked list of genes in the experiment (17) The filter conditions were a P value <1 and, and an adjusted P value filter <0.05.

Immune-cell infiltration analysis and the correlations between the immune cells and DEGs

The CIBERSORT deconvolution algorithm is a method used to characterize the cell composition of complex tissues from their gene expression profiles (18). The immune-cell infiltration of the GSE32863 data set was analyzed. The correlations between the immune cells and *DEGs* are displayed in a lollipop chart. The abscissa represents the

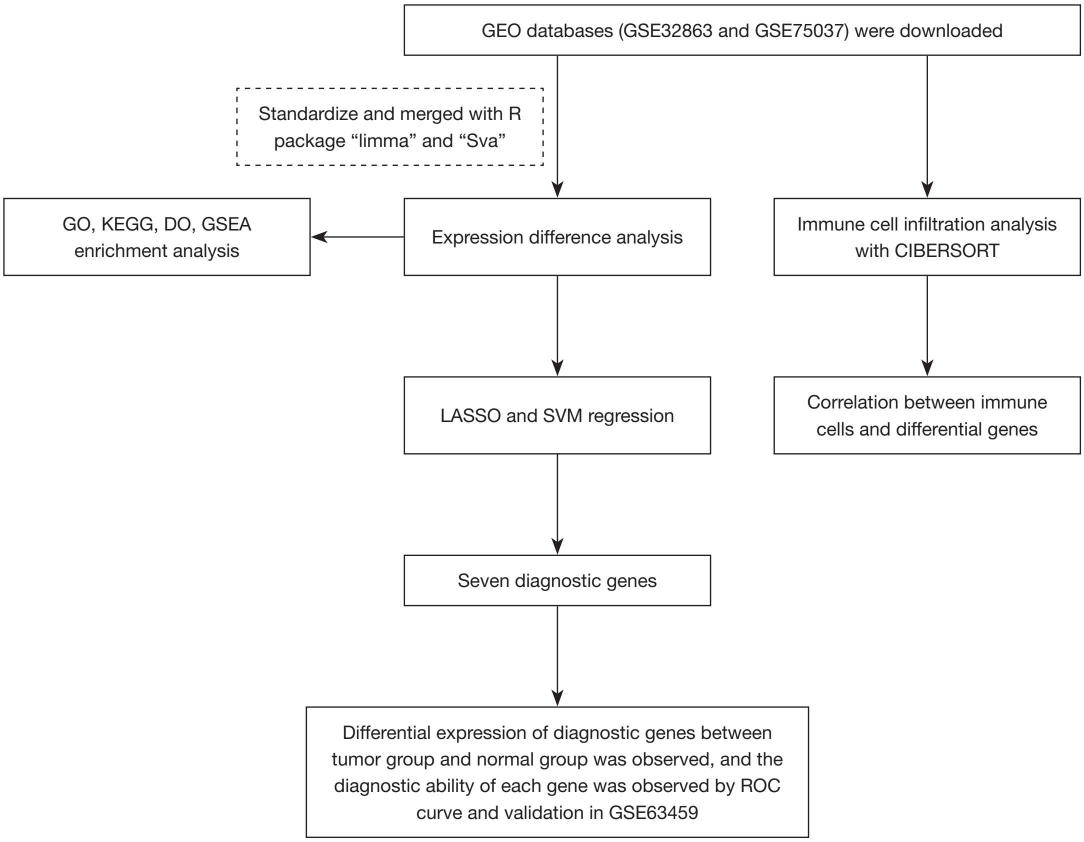
GEO databases (GSE32863 and GSE75037) were downloaded
Standardize and merged with R package “limma” and “Sva”
GO, KEGG, DO, GSEA enrichment analysis
Expression difference analysis
Immune cell infiltration analysis with CIBERSORT
LASSO and SVM regression
Correlation between immune cells and differential genes
Seven diagnostic genes
Differential expression of diagnostic genes between tumor group and normal group was observed, and the diagnostic ability of each gene was observed by ROC curve and validation in GSE63459

Figure 1 Study flowchart. GEO, Gene Expression Omnibus; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, gene set enrichment analysis; LASSO, least absolute shrinkage and selection operator; SVM, support vector machine; ROC, receiver operating characteristic.

correlation coefficient, the left ordinate represents the names of immune cells, and the right ordinate represents the P values. The size of the lollipop head indicates the correlation coefficient, and the color of the lollipop head indicates if the differences were significant (green indicates a P value <0.05, and yellow indicates a P value >0.05).

Statistical analysis

The diagnostic performance of the genes was assessed using area under the receiver operating characteristic curve (AUC). The distribution of the differentially expressed genes was shown by heatmap and volcano map. The differences of gene expression between the two groups were compared by *t*-test and expressed by boxplot. All the statistical analyses were performed by using R software (Version 4.1.1). A P value <0.05 was considered as statistical significance.

Results

Results of the DEG analysis

In total, 384 DEGs were identified from the GSE32863 data set, including 91 upregulated and 293 downregulated genes. The flowchart of the study is shown in *Figure 1*. The expression of the screened DEGs and the differences in details in each sample showed in https://cdn.amegroups.cn/static/public/jtd-22-206-1.xls, https://cdn.amegroups.cn/static/public/jtd-22-206-2.xls.

The top 50 DEGs are shown in *Figure 2A* (heat map) and *Figure 2B* (volcano map). We found that the genes recombinant glutathione peroxidase 2 (*GPX2*), Purkinje cell protein 4 (*PCP4*), locus649841 (*LOC649841*), fucosyltransferase 3 (*FUT3*), and transmembrane protein 45B (*TMEM45B*) were upregulated, while the genes intelectin 1 (*ITLN1*), metallothionein 1M (*MT1M*), nterleukin-6 (*IL-6*), surfactant protein A (*SFTPA*), and

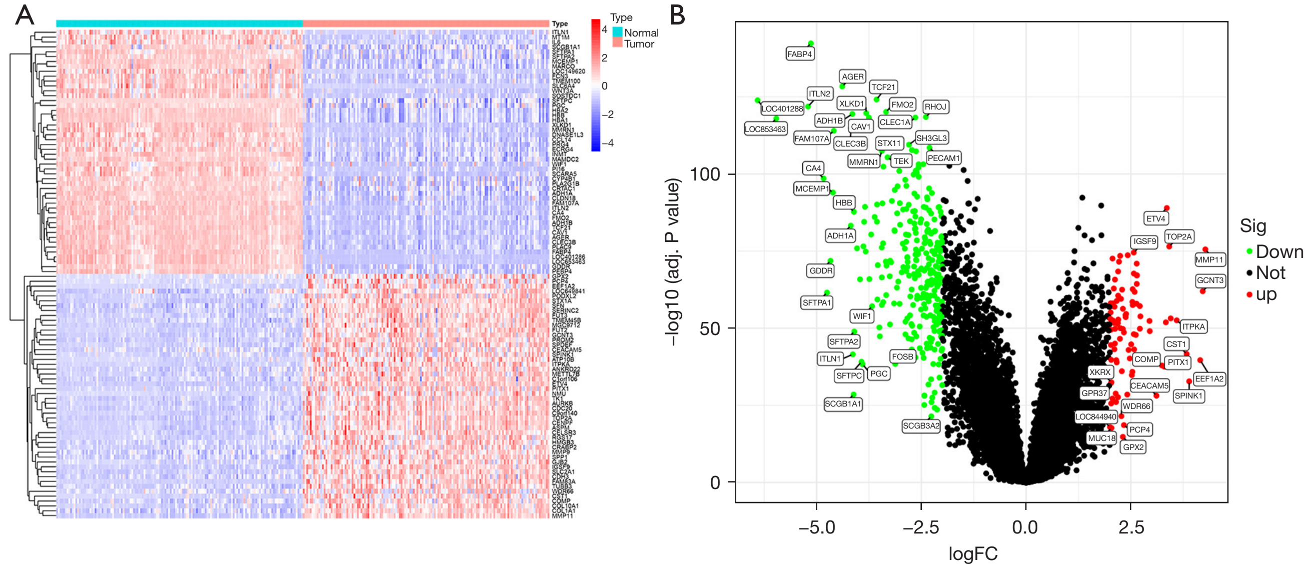

Figure 2 Heatmap (A) and volcano map (B) of the top 50 DEGs. DEGs, differentially expressed genes.

ficolin 3 (*FCN3*) were downregulated in the tumor group.

GO, KEGG and GSEA enrichment

During the GO annotation, the DEGs were found to be significantly enriched in terms of the CCs, including the endocytic vesicles, extracellular matrix, and collagen-containing extracellular matrix, BPs, including the extracellular matrix organization, humoral-immune response and neutrophil activation, and the MFs, including glycosaminoglycan binding, oxygen carrier activity, and haptoglobin binding (see *Figure 3A*). The KEGG analysis showed that these genes were significantly enriched in the signaling pathways related to malaria, complement and coagulation cascades, and leukocyte transendothelial migration (see *Figure 3B*). GO and KEGG pathway enrichment analyses were conducted by GSEA, and the GO terms "adaptive-immune-response", "humoral-immune-response", and "extracellular-signal-regulated kinases (ERK) 1 and 2 cascade" were significantly expressed in the normal group (see Figure S1A), while the GO terms "nuclear-chromosome", "DNA-conformation-change", and "chromosomal region" were significantly expressed in the tumor group (see Figure S1B). In the KEGG pathway analysis, the terms "chemokine-signaling-pathway", "cytokine-cytokine-receptor-interaction", and "graft-versus-host-disease" were significantly expressed in the normal group (see Figure S1C), while the terms "base-excision-repair", "cell-cycle", and "DNA-replication" were significantly expressed in the tumor group (see Figure S1D).

Screening of the diagnostic genes by LASSO and SVM regression analyses and the validation

The intersection results of the LASSO (see *Figure 4A*) and SVM (*Figure 4B*) regression analyses revealed 7 DEGs that were considered diagnostic genes (see *Figure 4C*); that is, *LOC401286*, flavin-containing monooxygenase 2 (*FMO2*), *XLKD1*, *RHOJ*, *SCARA5*, *HSPB2*, and serine incorporator 2 (*SERINC2*) (see *Table 1*). There were significant differences in the expression of these 7 genes between the normal and tumor groups (see *Figure 4D*). SERINC2 was significantly upregulated in the tumor group, but the remaining 6 genes were significantly downregulated.

The diagnostic performance of *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, *HSPB2*, and *SERINC2* for LUAD was assessed by a ROC curve analysis, which yielded area under the curve (AUC) values of 0.99, 1.00, 0.99, 1.00, 0.99, 0.99, and 0.98, respectively (see *Figure 5*). Similar results were obtained during the validation using the GSE63459 data set (see *Figure 6*). Consistently, SERINC2 was significantly upregulated in the tumor group, while the remaining 6 genes were significantly downregulated. The ROC curve analyses for *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, *HSPB2*, and *SERINC2* yielded AUC values of 0.97, 0.96, 0.94, 0.88, 0.85, 0.94 and 0.89, respectively.

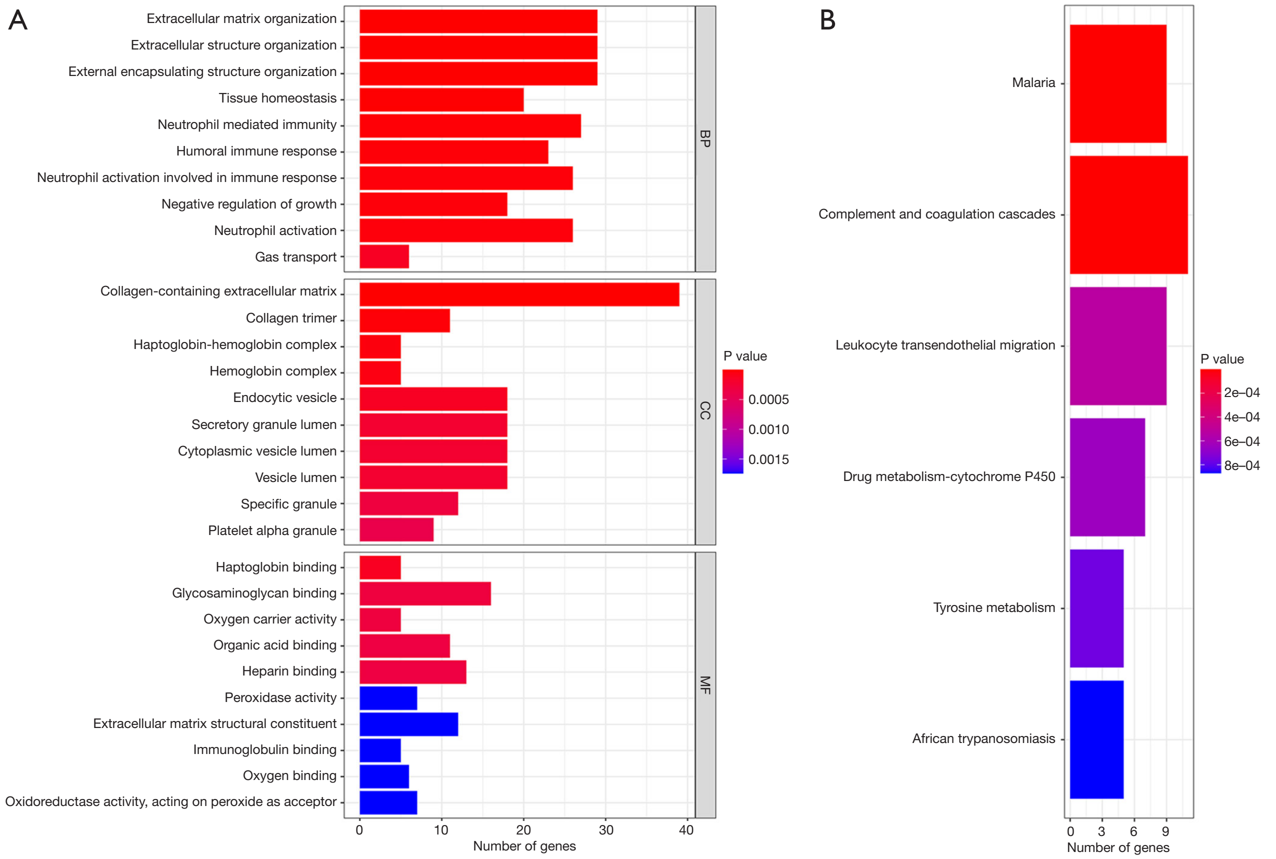

Figure 3 The GO (A) and KEGG (B) analyses of the diagnostic genes. BP, biological processes; CC, cellular component; MF, molecular function; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Immune-cell infiltration analysis and the correlations between immune cells and DEGs

The proportion of immune cells infiltrated in normal and tumor tissues (see *Figure 7A*), and the correlation between the immune cells (see *Figure 7B*) was analyzed. Significant differences in the content of the immune cells, naive B cells, memory B cells, plasma cells, naive cluster of differentiation (CD) 4 T cells, T follicular helper cells, regulatory T cells (Tregs), gamma delta T cells, monocytes, M0 macrophages, M1 macrophages, resting mast cells, activated mast cells, and neutrophils were found in normal and tumor tissues (see *Figure 7C*). The correlations between the infiltrating immune cells and the expression of the 7 diagnostic genes are shown in *Figure 8*. A negative correlation was found between M0 macrophages, and monocytes and resting mast cells (r=−0.66 and −0.73), while a positive correlation was found between monocytes and resting mast cells (r=0.64). *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, and *HSPB2* were positively correlated with monocytes and resting mast cells, and negatively correlated with Tregs and macrophages. The opposite results were found for SERINC2.

Discussion

In the present study, the functional enrichment analysis showed that the identified DEGs were enriched in the GO terms of extracellular matrix, glycosaminoglycan binding, complement and coagulation cascades, and leukocyte transendothelial migration. The intersection of the LASSO and SVM regression results identified 7 diagnostic genes (i.e., *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, *HSPB2*, and *SERINC2*), which shown significant performance for the early diagnosis of LUAD in clinical practice. We also estimated the infiltration of immune cells, and analyzed their correlations with the 7 diagnostic DEGs.

The SVM and LASSO regression methods are machine-

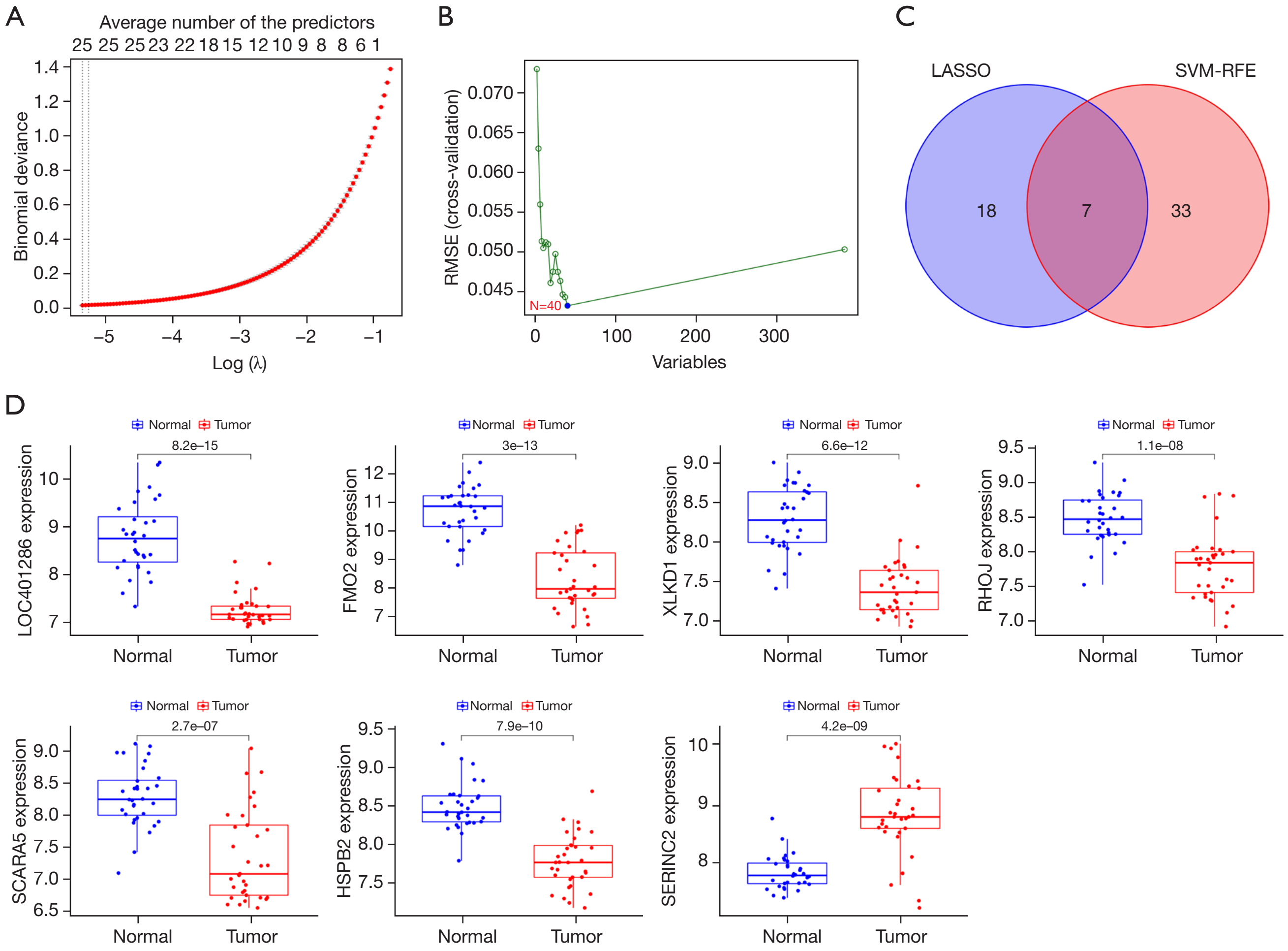

Figure 4 The intersection of genes obtained from the LASSO regression and SVM regression analyses, and the comparison of the expression of the diagnostic genes between the normal and tumor groups in the training (GSE32863) and validation data sets (GSE75037). (A) Venn plot of the intersecting genes from the LASSO regression and SVM regression; (B) comparison of the expression of the diagnostic genes between the normal and tumor groups; (C) Venn of LASSO and SVM regression; and (D) the differences of seven heat genes between normal group and tumor group. LASSO, least absolute shrinkage and selection operator; SVM, support vector machine.

learning methods that have been extensively used to screen diagnostic and prognosis-related indicators in recent years. In the present study, we integrated and intersected the regression results. Ultimately, 7 DEGs (i.e., *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, *HSPB2* and *SERINC2*) were identified. Consistent with the literature (19), only SERINC2 was upregulated in the tumor group. SERINC2 is a member of the SERINC family of transmembrane proteins that incorporates serine into membrane lipids, including phosphatidylserine, and sphingolipids, during synthesis. These membrane lipids thus act as important indicators of tumorigenesis and cancer progression (20,21). Further, SERINC2 reportedly promotes LUAD proliferation, migration, and invasion, and may involve the phosphatidylinositol 3 kinase (PI3K)/serine threonine kinase (Akt) signaling pathway (22) Additionally, SERINC2 has been correlated with alcohol dependence in Europeans (23). Interestingly, the knockdown of SERINC2 in hepatocellular carcinoma reportedly inhibits cell-cycle progression via the transcriptional activation of Kirsten rat sarcoma (k-Ras) (22). However, the role of SERINC2 in cancer has been largely understudied; thus, further studies and animal experiments

Table 1 The 7 intersection genes of the LASSO and SVM regression models

Different genes of the LASSO regression model	Different genes of the SVM regression model	Intersection genes
FABP4	*HBA2*	*LOC401286*
LOC401286	*MFAP4*	*FMO2*
FMO2	*FAM107A*	*XLKD1*
XLKD1	*AGER*	*RHOJ*
RHOJ	*XLKD1*	*SCARA5*
MMRN1	*TEK*	*HSPB2*
SCARA5	*HSPB2*	*SERINC2*
LGI3	*JAM2*	
ETV4	*RASL12*	
HSPB2	*HBB*	
C10orf67	*LDB2*	
SOX17	*TCF21*	
CCL23	*RHOJ*	
MT1M	*STX11*	
IGSF9	*SH3GL3*	
C5AR1	*STX1A*	
MGC34774	*FMO2*	
SERINC2	*LOC401286*	
PROM2	*C2orf32*	
GCNT3	*MS4A7*	
RHBDL1	*ANKRD47*	
DES	*EDG1*	
IL6	*PECAM1*	
C18orf34	*LOC653463*	
SPDEF	*LIMS2*	
	SERINC2	
	FHL1	
	SCN4B	
	NAP5	
	MMP11	
	CAV2	
	DNASE1L3	
	COX7A1	
	CFD	
	GIMAP5	
	ADH1A	
	SCARA5	
	DPEP2	
	PLAC9	
	GSTM5	

LASSO, least absolute shrinkage and selection operator; SVM, support vector machine.

need to be conducted to assess the value of SERINC2 as an early diagnostic or therapeutic marker for LUAD.

Human FMO2 is expressed in the lungs in 2 isoforms (i.e., FMO2*2A and FMO2*1) (24), and acts as a tumor suppressor in LUAD. XLKD1, RHOJ and SCARA5 are reportedly downregulated in patients with LUAD, and their high expression can inhibit the occurrence and development of cancer (25,26). Additionally, HSPB2 has been shown to be correlated with pancreatic cancer and hepatocellular carcinoma (27) via the activation of protein 53. To the best of our knowledge, no study has uncovered a relationship between HSPB2 and LUAD; thus, further investigations need to be conducted.

Over the years, due to unprecedented technological progress, the focus of research has shifted from tumor cells to the TME, and our understanding of tumorigenesis has been refined (28). It is now well established that immune-cell infiltration is an important part of the TME (29,30), and the immune system plays a dual role in tumor cells. At present, immunosuppressive therapy, such as programmed cell death protein 1 (PD-1) inhibitor therapy, has become the mainstay of LUAD treatment along with chemotherapy and surgery (29,31)

Consistent with the literature (28), our immune-cell infiltration analysis showed that naive B cells, memory B cells, plasma cells, naive CD4 T cells, T follicular helper cells, Tregs, gamma delta T cells, monocytes, M0 macrophages, M1 macrophages, resting mast cells, activated mast cells, and neutrophils were significantly different between LUAD and healthy subjects. $CD4^+$ and $CD8^+$ T cells and their secreted cytokines participate in adaptive immunity; PD-1 inhibition has been reported to result in the increased proliferation of all T cell subsets and

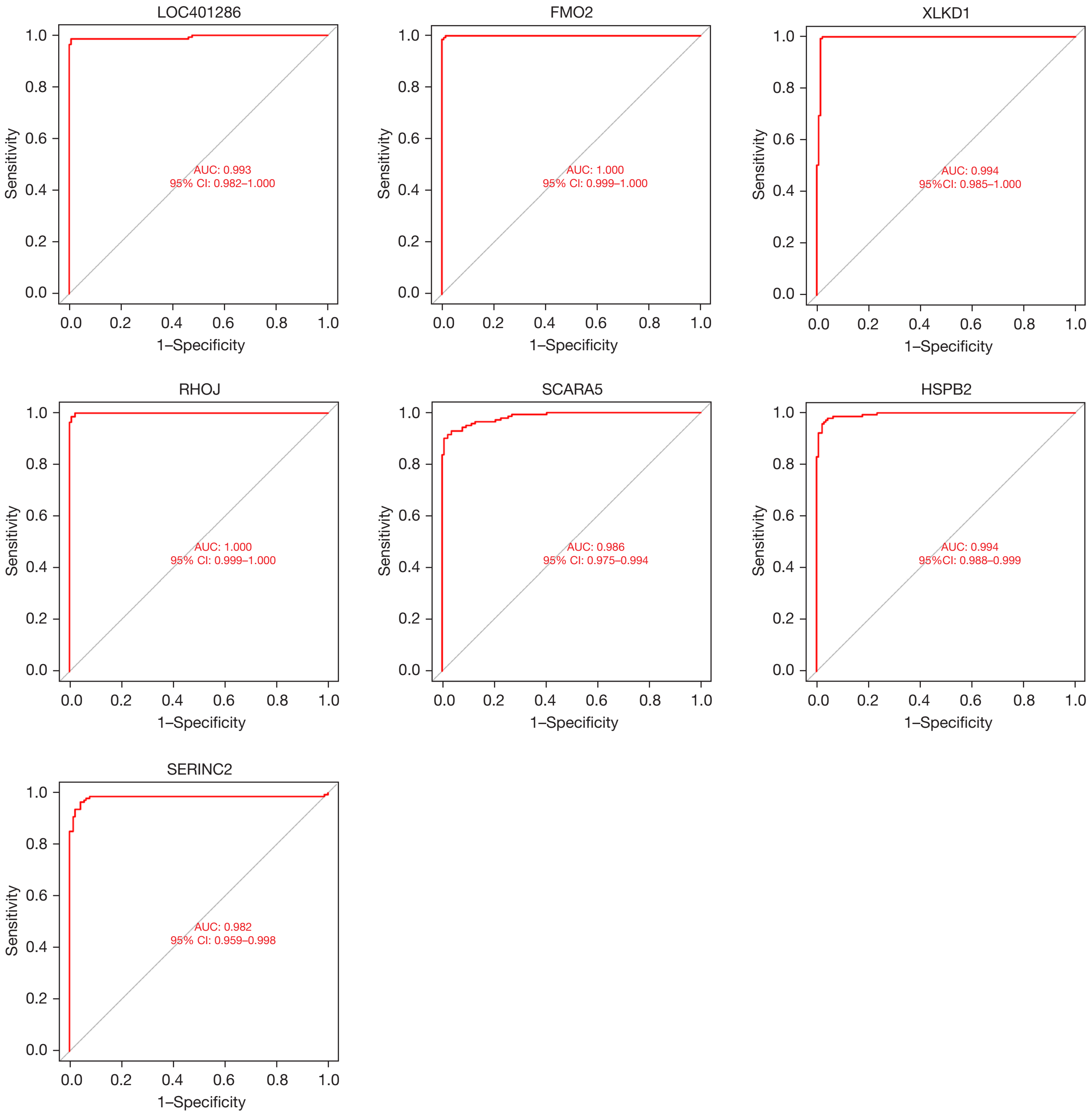

Figure 5 ROC curves of the diagnostic genes in the training data set. AUC, area under the receiver operating characteristic curve; ROC, receiver operating characteristic.

effector cytokine production by CD4+ T helper 1 cells (32). Our study found that M0 macrophages were negatively correlated with monocytes and resting mast cells (r=–0.66 and –0.73), but a positive correlation was found between monocytes and resting mast cells (r=0.64). *LOC401286*, *FMO2*, *XLKD1*, *RHOJ*, *SCARA5*, and *HSPB2* were positively correlated with monocytes and resting mast cells, and negatively correlated with Tregs and macrophages. Interestingly, the opposite results were observed for *SERINC2*. These results are consistent with other reports (33), but further experimental research at the *in-vivo* and *in-vitro* levels needs to be conducted to increase

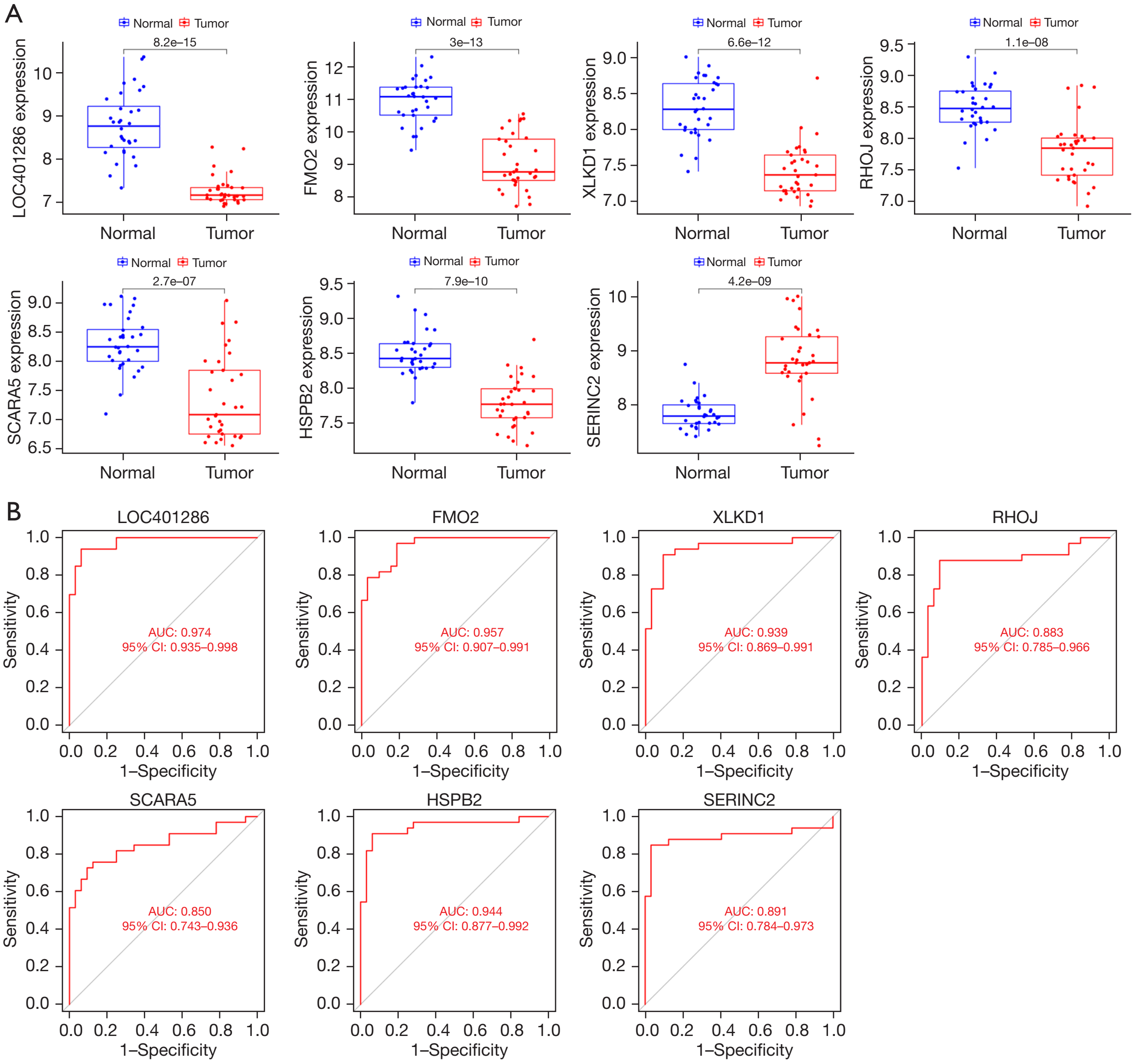

Figure 6 Comparison of the expression of the diagnostic genes between the normal and tumor groups (A) and the ROC curves of the diagnostic genes in the validation data set (GSE63459) (B). AUC, area under the receiver operating characteristic curve; ROC, receiver operating characteristic.

the robustness of our finding. The limitation of this study was that the stage of patients from validation dataset was not exactly the same as the stage in training dataset, because the paper needs to identify the diagnostic gene, which was more meaningful in the early stages, and therefore made up for this deficiency.

In conclusion, we identified 7 DEGs in LUAD tissue that can be considered diagnostic genes based on 2 machine-learning regression methods (i.e., the SVM and LASSO regression models). Our findings were successfully validated using another independent data set that contained data from patients with stage I LUAD. Importantly, the infiltrating immune cells were analyzed, and a significant correlation was found to the 7 DEGs, which suggests that these genes affect the occurrence and development of tumors via their interaction with immune cells. Accordingly, our findings could be very helpful for the early diagnosis of LUAD in clinical practice.

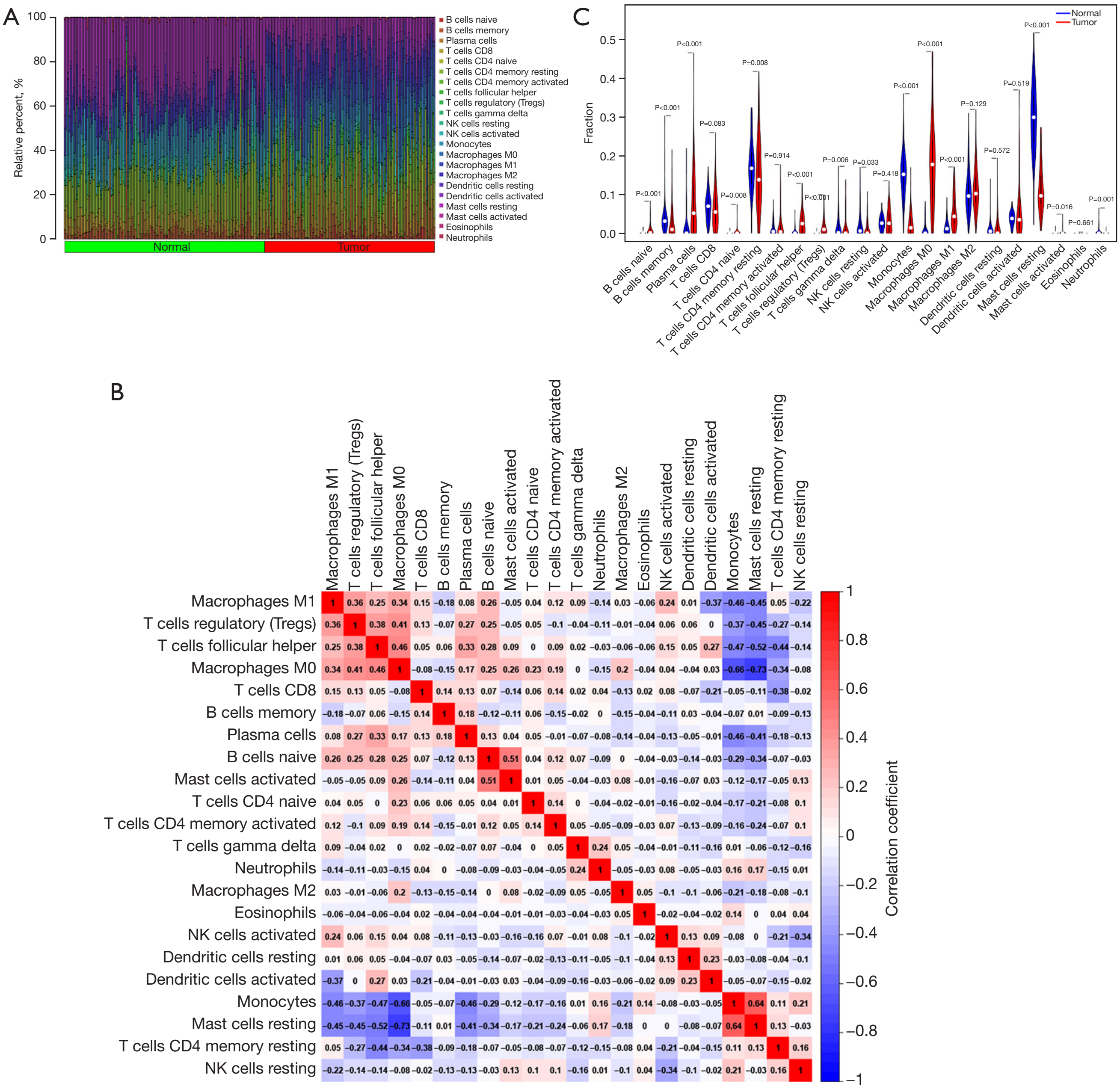

Figure 7 Immune-cell infiltration analysis of patients with LUAD. (A) Comparison of the infiltration of immune cells in normal and tumor tissues; (B) heatmap of the correlations between the immune cells in the tumor tissues; (C) comparison of the infiltrating immune cells in normal and tumor tissues. LUAD, lung adenocarcinoma.

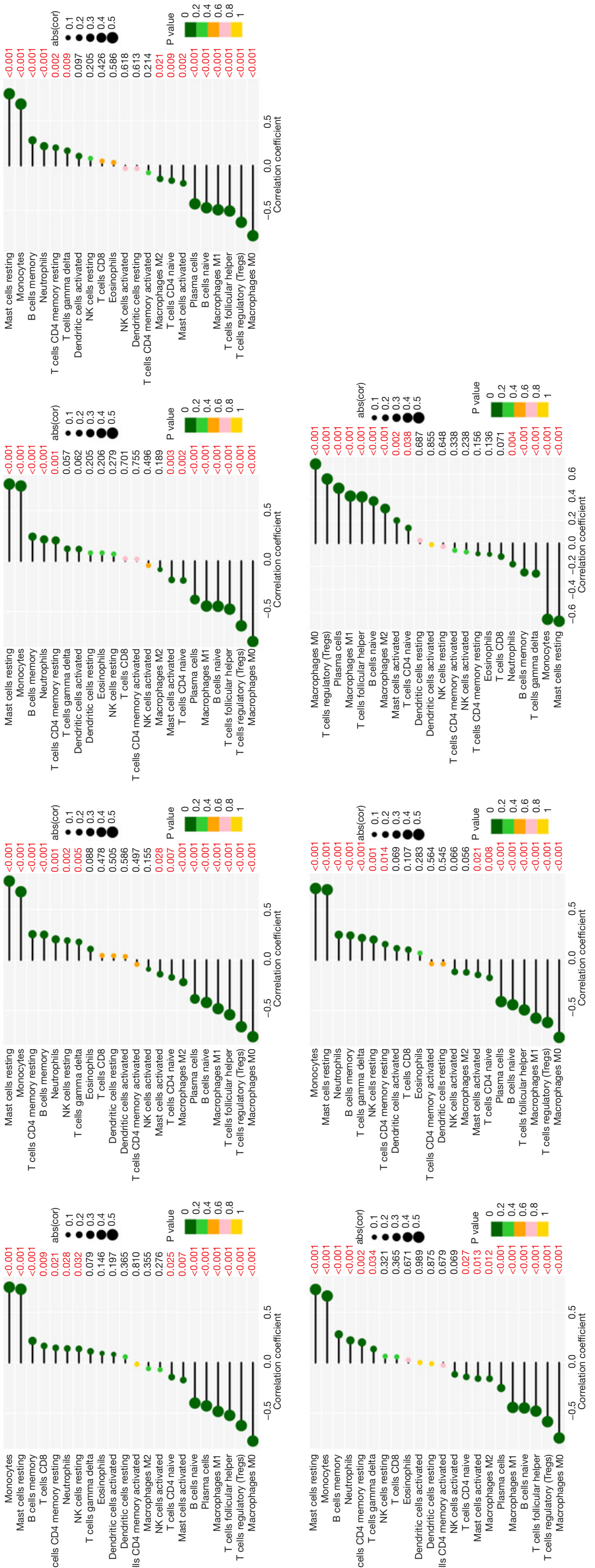

Figure 8 Correlations between the infiltration of immune cells and the expression of the 7 diagnostic genes.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (Nos. 81970167 and 81800108).

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-206/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jtd.amegroups.com/article/view/10.21037/jtd-22-206/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Reck M, Rabe KF. Precision Diagnosis and Treatment for Advanced Non-Small-Cell Lung Cancer. N Engl J Med 2017; 377:849-61.
2. Shao X, Niu R, Jiang Z, et al. Role of PET/CT in Management of Early Lung Adenocarcinoma. AJR Am J Roentgenol 2020; 214:437-45.
3. Zhang M, Yue J, Cui R, et al. Bright quantum dots emitting at ~1,600 nm in the NIR-IIb window for deep tissue fluorescence imaging. Proc Natl Acad Sci U S A 2018; 115:6590-5.
4. Wang Y, Zhao H, Gao X, et al. Identification of a three-miRNA signature as a blood-borne diagnostic marker

for early diagnosis of lung adenocarcinoma. Oncotarget 2016;7:26070-86.

5. Li M, Zhang C, Zhou L, et al. Identification and validation of novel DNA methylation markers for early diagnosis of lung adenocarcinoma. Mol Oncol 2020;14:2744-58.
6. Meng QC, Gao PR, Ren PF, et al. Early diagnosis of subtype in early clinical stage lung adenocarcinoma by using an autoantibody panel and computed tomography. Zhonghua Yi Xue Za Zhi 2019;99:204-8.
7. Zhang C, Zhang J, Xu FP, et al. Genomic Landscape and Immune Microenvironment Features of Preinvasive and Early Invasive Lung Adenocarcinoma. J Thorac Oncol 2019;14:1912-23.
8. Petitprez F, Meylan M, de Reyniès A, et al. The Tumor Microenvironment in the Response to Immune Checkpoint Blockade Therapies. Front Immunol 2020;11:784.
9. Katsuta E, Rashid OM, Takabe K. Clinical relevance of tumor microenvironment: immune cells, vessels, and mouse models. Hum Cell 2020;33:930-7.
10. Xie M, Wei J, Xu J. Inducers, Attractors and Modulators of CD4+ Treg Cells in Non-Small-Cell Lung Cancer. Front Immunol 2020;11:676.
11. Wu J, Zan X, Gao L, et al. A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study. JMIR Med Inform 2019;7:e13476.
12. Sun S, Guo W, Wang Z, et al. Development and validation of an immune-related prognostic signature in lung adenocarcinoma. Cancer Med 2020;9:5960-75.
13. Chen H, Carrot-Zhang J, Zhao Y, et al. Genomic and immune profiling of pre-invasive lung adenocarcinoma. Nat Commun 2019;10:5472.
14. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47:D330-8.
15. Kanehisa M, Furumichi M, Sato Y, et al. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res 2021;49:D545-51.
16. Canzler S, Hackermüller J. multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. BMC Bioinformatics 2020;21:561.
17. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453-7.
18. Wu K, Zhang X, Li F, et al. Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. Nat Commun 2015;6:10131.
19. Digumarti R, Bapsy PP, Suresh AV, et al. Bavituximab plus paclitaxel and carboplatin for the treatment of advanced non-small-cell lung cancer. Lung Cancer 2014;86:231-6.
20. Proia RL, Hla T. Emerging biology of sphingosine-1-phosphate: its role in pathogenesis and therapy. J Clin Invest 2015;125:1379-87.
21. Zeng Y, Xiao D, He H, et al. SERINC2-knockdown inhibits proliferation, migration and invasion in lung adenocarcinoma. Oncol Lett 2018;16:5916-22.
22. Zuo L, Wang KS, Zhang XY, et al. Rare SERINC2 variants are specific for alcohol dependence in individuals of European descent. Pharmacogenet Genomics 2013;23:395-402.
23. Cashman JR, Zhang J. Human flavin-containing monooxygenases. Annu Rev Pharmacol Toxicol 2006;46:65-100.
24. Hsu YL, Hung JY, Lee YL, et al. Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. Oncotarget 2017;8:104831-54.
25. Zeng T, Chen C, Yang P, et al. A Protective Role for RHOJ in NonSmall Cell Lung Cancer Based on Integrated Bioinformatics Analysis. J Comput Biol 2020;27:1092-103.
26. Liu J, Hu G, Chen D, et al. Suppression of SCARA5 by Snail1 is essential for EMT-associated cell migration of A549 cells. Oncogenesis 2013;2:e73.
27. Yu Z, Wang H, Fang Y, et al. Molecular chaperone HspB2 inhibited pancreatic cancer cell proliferation via activating p53 downstream gene RPRM, BAI1, and TSAP6. J Cell Biochem 2020;121:2318-29.
28. Yang Z, Zhuang L, Szatmary P, et al. Upregulation of heat shock proteins (HSPA12A, HSP90B1, HSPA4, HSPA5 and HSPA6) in tumour tissues is associated with poor outcomes from HBV-related early-stage hepatocellular carcinoma. Int J Med Sci 2015;12:256-63.
29. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell 2012;21:309-22.
30. Zhang C, Wang H, Wang X, et al. CD44, a marker of cancer stem cells, is positively correlated with PD-L1 expression and immune cells infiltration in lung adenocarcinoma. Cancer Cell Int 2020;20:583.
31. Zuo S, Wei M, Wang S, et al. Pan-Cancer Analysis of Immune Cell Infiltration Identifies a Prognostic Immune-Cell Characteristic Score (ICCS) in Lung Adenocarcinoma. Front Immunol 2020;11:1218.
32. Saab S, Zalzale H, Rahal Z, et al. Insights Into Lung Cancer Immune-Based Biology, Prevention, and

Treatment. Front Immunol 2020;11:159.

33. Fan T, Lu Z, Liu Y, et al. A Novel Immune-Related Seventeen-Gene Signature for Predicting Early Stage Lung Squamous Cell Carcinoma Prognosis. Front Immunol 2021;12:665407.

(English Language Editor: L. Huleatt)

Cite this article as: Wang S, Wang Q, Fan B, Gong J, Sun L, Hu B, Wang D. Machine learning-based screening of the diagnostic genes and their relationship with immune-cell infiltration in patients with lung adenocarcinoma. J Thorac Dis 2022;14(3):699-711. doi: 10.21037/jtd-22-206

Supplementary

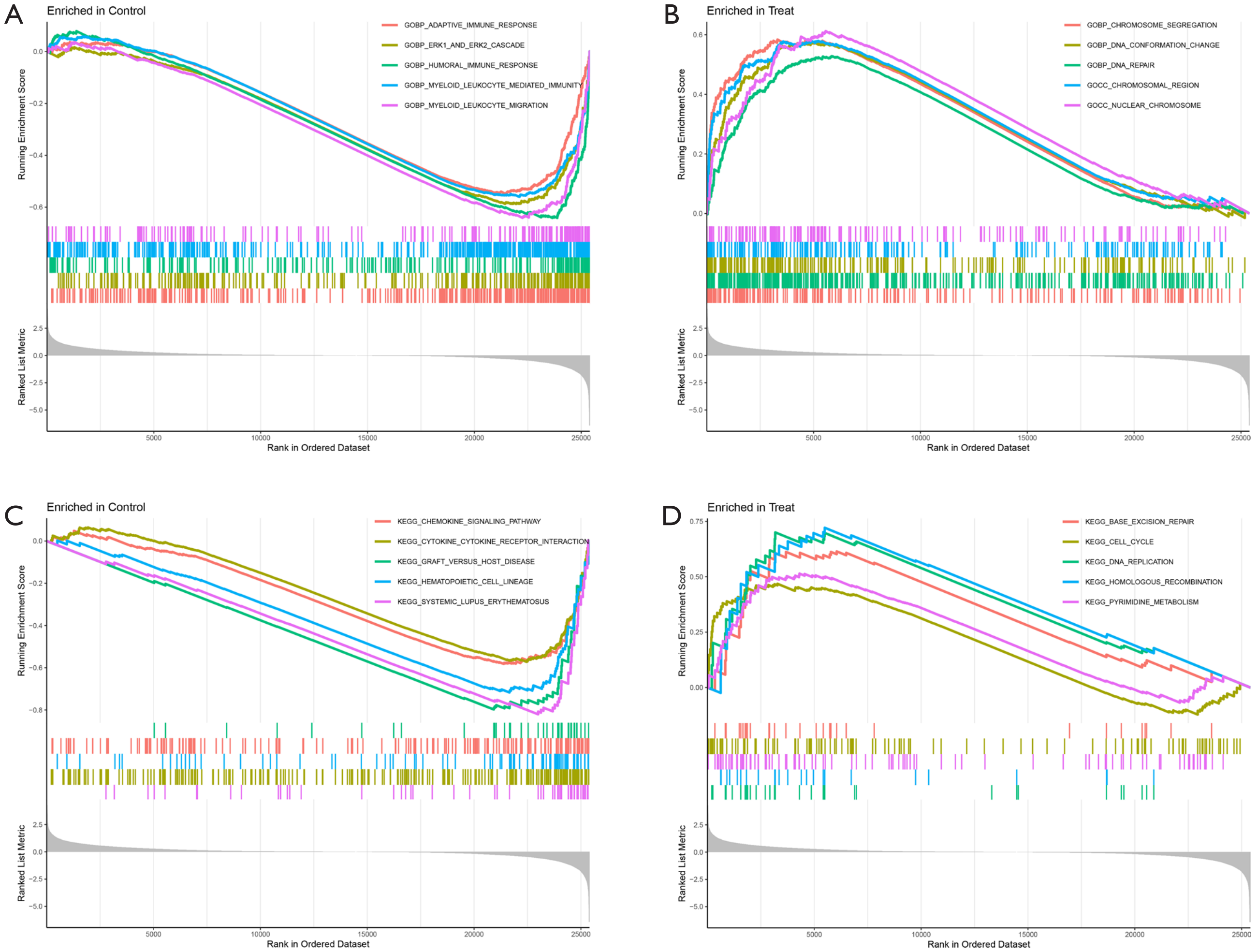

Figure S1 The GO and KEGG pathway enrichment analysis was conducted by GSEA. (A) GSEA enrichment of GO terms in the normal group; (B) GSEA enrichment of GO terms in the tumor group; (C) GSEA of the KEGG pathways in the normal group; and (D**)** GSEA of the KEGG pathways in the tumor group. GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, gene set enrichment analysis.