scientific reports

Check for updates

OPEN

Immune, metabolic landscapes of prognostic signatures for lung adenocarcinoma based on a novel deep learning framework

Shimei Qin[1,2], Shibin Sun[1,2], Yahui Wang[1,2], Chao Li[1], Lei Fu[1], Ming Wu[1], Jinxing Yan[1], Wan Li[1], Junjie Lv[1]✉ & Lina Chen[1]✉

Lung adenocarcinoma (LUAD) is a malignant tumor with high lethality, and the aim of this study was to identify promising biomarkers for LUAD. Using the TCGA-LUAD dataset as a discovery cohort, a novel joint framework VAEjMLP based on variational autoencoder (VAE) and multilayer perceptron (MLP) was proposed. And the Shapley Additive Explanations (SHAP) method was introduced to evaluate the contribution of feature genes to the classification decision, which helped us to develop a biologically meaningful biomarker potential scoring algorithm. Nineteen potential biomarkers for LUAD were identified, which were involved in the regulation of immune and metabolic functions in LUAD. A prognostic risk model for LUAD was constructed by the biomarkers HLA-DRB1, SCGB1A1, and HLA-DRB5 screened by Cox regression analysis, dividing the patients into high-risk and low-risk groups. The prognostic risk model was validated with external datasets. The low-risk group was characterized by enrichment of immune pathways and higher immune infiltration compared to the high-risk group. While, the high-risk group was accompanied by an increase in metabolic pathway activity. There were significant differences between the high- and low-risk groups in metabolic reprogramming of aerobic glycolysis, amino acids, and lipids, as well as in angiogenic activity, epithelial-mesenchymal transition, tumorigenic cytokines, and inflammatory response. Furthermore, high-risk patients were more sensitive to Afatinib, Gefitinib, and Gemcitabine as predicted by the pRRophetic algorithm. This study provides prognostic signatures capable of revealing the immune and metabolic landscapes for LUAD, and may shed light on the identification of other cancer biomarkers.

Cancer is still a global major public health problem[1]. According to the latest global cancer statistics estimates, lung cancer remains the leading cause of cancer death and the second most commonly diagnosed cancer, accounting for approximately 20% of cancer-related mortality and 10% of incidence[2]. Non-small cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer cases[3], with lung adenocarcinoma (LUAD) as its main histologic subtype and its incidence is still increasing[4]. Despite significant advances in diagnosis and treatment, the 5-year survival rate of LUAD is only 4–17%[5]. Therefore, screening and poor prognosis of lung adenocarcinoma remains an ongoing challenge. Accumulating evidence suggests that biomarker identification and application are of major importance for timely diagnosis and accurate prognosis of cancer[6,7].

Machine learning has been widely used in biomarker discovery studies for cancer[8–10]. For example, based on breast cancer gene expression datasets, Zare et al.[11] identified 59 novel inflammatory breast cancer-specific gene signatures using a random forest approach. Zhang et al.[12] used a support vector machine algorithm to excavate a m6A target miRNAs diagnostic signature for cancer detection and successfully implemented it for lung cancer. Machine learning algorithms were deployed to develop tumor-infiltrating immune cell associated RNAs to predict survival outcomes in LUAD patients[13]. Deep learning has gained increasing attention in the research of cancer signatures identification[14,15]. The denoising autoencoder, an unsupervised deep learning algorithm, has been successfully applied to LUAD molecular signature mining[16]. Divate et al.[17] presented a robust neural network model providing gene signatures for accurate classification of cancers based on gene expression data. Three feed-forward neural networks were employed to RNA-seq samples from 18 solid tumor types and recognized

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150000, China. [2]These authors contributed equally: Shimei Qin, Shibin Sun and Yahui Wang. ✉email: lvjunjie525@126.com; chenlina@ems.hrbmu.edu.cn

transcriptome signatures that were consistent across tumors[18]. Deep learning methods are capable of identifying highly complex patterns in large datasets compared to common machine learning techniques, which help to efficiently identify molecular signatures associated with cancer[19].

Variational autoencoder (VAE) is a deep generative model based on variational Bayesian inference designed to learn nonlinear latent representations of high-dimensional data[20]. VAE has shown encouraging results in capturing biologically meaningful low-dimensional representations of multi-omics data. Daniel et al.[21] utilized VAE to learn a generalized latent representation of large-scale metabolomics data, and VAE representation outperformed both linear and nonlinear principal component analysis. VAE models trained on gene expression data have good ability to identify generalizable biological representations[22]. In addition, VAE has been applied in multi-omics representation learning such as proteomics and epigenomics[23,24]. Evidently, VAE is a powerful method for dimensionality reduction. Neural networks are commonly employed in the current deep learning field, among which multilayer perceptron (MLP) is widely used in cancer diagnosis research. Lorencin et al.[25] combined MLP and Laplacian edge detector to achieve bladder cancer detection. Deep learning methods including MLP have revealed salivary glycopatterns as biomarkers for the diagnosis of papillary thyroid cancer[26]. MLP has also been applied to screening for breast, colorectal, and prostate cancers[27–29]. Shapley Additive Explanations (SHAP) is derived from the Shapley value in cooperative game theory, which aims to tackle the problem of lack of interpretability faced by machine learning[30]. In a study by Chakraborty et al.[31], SHAP facilitated the identification of prognostic factors for breast cancer by enhancing the interpretability of the extreme gradient boosting model. Multiple studies have utilized SHAP to interpret the output of machine learning models to screen out important features[32–34]. SHAP is considered a state-of-the-art machine learning interpreter[35]. The magnitude and direction of the influence of features on the output variables can be assessed by calculating SHAP values.

Here, we developed a LUAD biomarker potential scoring algorithm based on a novel deep learning framework, variational autoencoder joint multilayer perceptron (VAEjMLP), which combined the feature dimensionality reduction and classification prediction tasks and efficiently evaluated the importance of each feature gene by applying SHAP. This work successfully identified 19 LUAD biomarkers (LABs) and constructed a LUAD prognostic risk model. The functions mediated by the biomarkers in LUAD were investigated. The metabolic and immune landscapes of the prognostic risk model and its associations with angiogenesis, epithelial-mesenchymal transition (EMT), tumorigenic cytokines, and inflammation were explored. Our study provided reliable biomarkers for LUAD and may help to more accurately determine the survival of LUAD patients.

Materials and methods

LUAD data sources and preprocessing

The RNA-seq expression data and clinical information for LUAD were retrieved from The Cancer Genome Atlas (TCGA) (https://tcga-data.nci.nih.gov/tcga/) data portal. In constructing the prognostic risk model, the samples were filtered by clinical information including age, sex, survival time, overall survival (OS) status, pathologic T, N, M, stage, and history of previous cancer diagnosis, and a total of 482 patients with LUAD were included in the study. The GSE72094 derived from The Gene Expression Omnibus (GEO)[36] (http://www.ncbi.nlm.nih.gov/geo) database was used as a validation set for the prognostic risk model, in which 393 LUAD samples with complete clinical information were included in this study. Single-cell RNA sequencing (scRNA-seq) data for 11 cases of distal normal lung tissue and 15 cases of primary LUAD were acquired from the GEO database, with accession number GSE131907. The GSE131907 dataset was used to analyze the expression of prognostic risk factors at the single-cell transcriptome level.

Methods

Identifying reliable biomarkers is crucial for improving the prognosis of LUAD. Deep learning has been widely used in the exploration of biological problems. In this study, we proposed a novel deep learning framework based on VAE, MLP, and SHAP to identify biomarkers and construct a prognostic risk model for LUAD (the workflow of our investigation is shown in Fig. 1).

LUAD biomarker identification framework

In order to mine biomarkers for LUAD, it is first necessary to identify genes that are highly altered between cancer and normal states. Based on the TCGA LUAD expression data consisting of read counts, the differentially expressed genes between normal and cancer samples were identified using the negative binomial distribution model of the R package "edgeR" (version 3.40.2), with a threshold set at $|\log2(\text{fold change})| > 1$ and $p\text{-value} < 0.05$[37]. To identify more stable differentially expressed genes, random sampling was performed on normal and cancer samples in a 1:6 ratio for 1000 times, where all normal samples were taken in each sampling. Differential expression analysis was conducted on each of the 1000 sets of samples and then the frequency of differential expression of each gene was counted. All differentially expressed genes, as well as genes with a frequency not less than 950 or 750, were selected as subsequent input features, respectively. The joint framework VAEjMLP was developed by combining the VAE model with MLP to perform the feature dimensionality reduction and classification prediction tasks simultaneously, allowing efficient representation learning with the help of the classification task. VAE consists of an encoder and a decoder and is an important generative model for dealing with feature representation capabilities. In this study, the encoder of the VAE is an inference model modelled by a neural network with the variational parameter φ as a learnable parameter, defined as an approximate posterior distribution, which is further assumed to be a Gaussian distribution:

$$z = encoder(x) \sim q_{\varphi}(z|x) \tag{1}$$

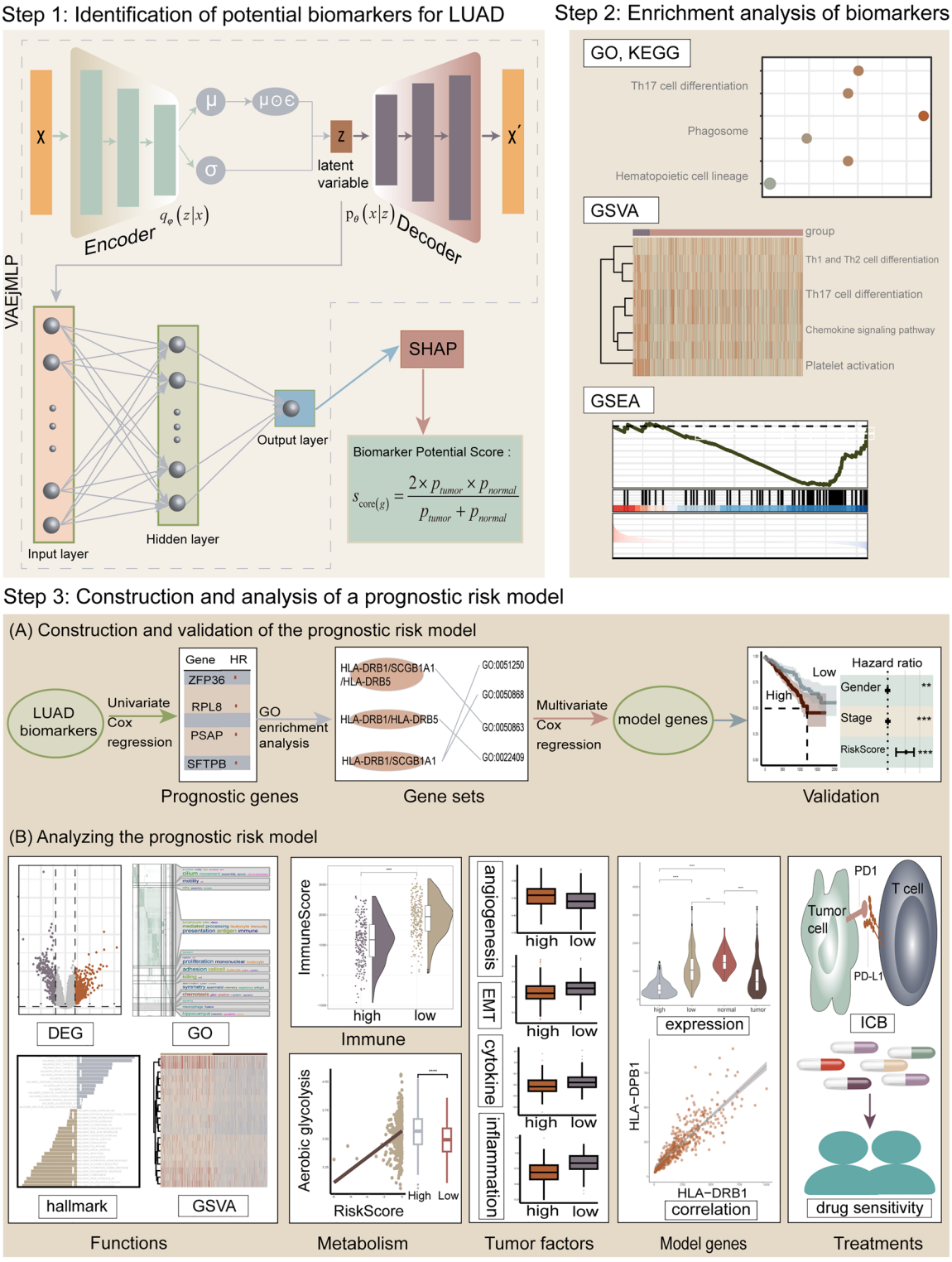

Figure 1. Flowchart of this study. Step 1, Constructing a joint framework VAEjMLP and then using the output of VAEjMLP as the input to Shapley Additive Explanations (SHAP), based on which a biomarker potential scoring algorithm was developed. Step 2, Enrichment analysis of biomarkers. Step 3, Construct a prognostic risk model for Lung adenocarcinoma (LUAD) by examining the prognostic significance and biological functions of the potential biomarkers, and analyze the model in multiple aspects. "DEG" means differentially expressed gene and "ICB" means immune checkpoint blockade.

where z is the latent variable and x is the original input. The encoder will output the mean, μ, and variance, σ, of the latent distribution, allowing for sampling from the hidden variable space. Since this operation of sampling is non-differentiable, to make the model trainable, sampling was performed by re-parameterization:

$$z = \mu + \sigma\epsilon \tag{2}$$

where $\epsilon \sim N(0, 1)$. The decoder still modelled by a neural network uses z as input to reconstruct the original data to generate x':

$$x' = decoder(x) \sim p_\theta(x|z) \tag{3}$$

where θ is the learnable parameter of the decoder and $p_\theta(x|z)$ follows a Bernoulli distribution. The objective function of the VAE is given by:

$$L_{VAE} = BCE\left(x, x'\right) + D_{KL}(N(\mu, \sigma)\|N(0, 1)) \tag{4}$$

where BCE is the binary cross entropy used to compute the distance between the original input x and the reconstructed data $x\prime$, and D_{KL} is the difference between the learned distribution and the standard Gaussian distribution. The VAE maps the original input to a low-dimensional latent space, which achieves feature dimensionality reduction. The latent space representation was then used as an input to the MLP to perform the classification task. The MLP adopted the cross entropy as a loss function, which can be expressed as:

$$L_{MLP} = CE\left(y, y'\right) \tag{5}$$

where y is the true label and y' is the predicted label. Ultimately, the loss function of the VAEjMLP can be expressed as:

$$L_{VAE-MLP} = \lambda L_{VAE} + L_{MLP} \tag{6}$$

where λ is the hyperparameter that balances the training weights of VAE and MLP and was set to 0.001. The training objective of VAEjMLP is namely to minimize the loss function $L_{VAE-MLP}$. To train the VAEjMLP, we had randomly split the sample into an 80% training set and a 20% test set. The trained VAEjMLP was then passed into SHAP to quantify the contribution of each feature to the prediction results. SHAP interprets the predicted value $y^{(i)}$ of sample $x^{(i)}$ as the sum of the attributions of each feature in the sample, where the attributions are the SHAP values, which satisfy the following formula:

$$\mathrm{y^{(i)}} = \mathrm{y^{base}} + \sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{f}\left(\mathrm{x_j^{(i)}}\right) \tag{7}$$

where y^{base} is the model's baseline, which is the mean prediction of all training samples, N represents the number of feature genes, and $f\left(x_j^{(i)}\right)$ is the SHAP value of the j-th feature of the sample $x^{(i)}$. $f\left(x_j^{(i)}\right) > 0$ indicates that the feature is a positive contributor to the prediction, and vice versa, indicating that the feature is a negative contributor to the prediction. SHAP outputs the contribution rate of each feature gene in each sample. Based on the positive or negative impact of feature genes on the predicted output, this study proposed a biomarker potential scoring algorithm to assess the potential of the feature genes as markers for LUAD. First, the distribution of the data formed by all contribution rate values for all feature genes was examined, and a threshold for significant contribution was determined based on the distribution. In this study, a positive contribution rate indicates that the feature gene tends to classify the sample as a cancer sample and, conversely, a negative contribution rate when classified as a normal sample. For each feature gene g with significant contribution, there exists the number of times C_P counted as a significant positive contribution in cancer samples and the number of times C_N counted as a significant negative contribution in normal samples. The number of cancer samples is N_{tumor} and the number of normal samples is N_{normal}, then the proportion of feature gene g that make a significant positive contribution to cancer samples is $P_{tumor} = \frac{C_P}{N_{tumor}}$, and the proportion of feature gene g that make a significant negative contribution to normal samples is $P_{normal} = \frac{C_N}{N_{normal}}$. Then, the biomarker potential score $S_{core}\left(g\right)$ for feature gene g is calculated as follows,

$$S_{core}\left(g\right) = \frac{2 \times P_{tumor} \times P_{normal}}{P_{tumor} + P_{normal}} \tag{8}$$

The biomarker potential scores of all feature genes with significant contributions were calculated, and the 25th percentile of all scores in descending order was used as the lower bound to screen for potential LABs. The framework for identifying LABs was implemented in python using the PyTorch SHAP software packages[38,39].

Enrichment analysis of LABs

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed using the R package "clusterProfiler"[40]. Biological pathways of target mitochondrial genes were annotated by MitoCarta 3.0 (http://www.broadinstitute.org/mitocarta)[41] database. Gene set variation analysis (GSVA) is an unsupervised method that calculates enrichment scores for specific gene sets in each sample[42]. We used the GSVA to assess and compare differences in enrichment of functional gene sets across samples, with reference gene sets from the KEGG database resource (http://www.genome.jp/kegg/)[43]. Gene set enrichment analysis (GSEA) is used to assess the distribution trend of genes in a predefined gene set in a gene list ranked by phenotypic relatedness to determine their contribution to the phenotype[44]. In this study, GSEA was applied to identify biological

functions associated with LUAD. | NES (normalized enrichment score) | > 1, Permutation test p-value < 0.05 and false discovery rate (FDR) < 0.25 were considered statistically significant.

Constructing a prognostic risk model for LUAD

Univariate Cox regression analysis was performed on each LABs to screen genes significantly associated with OS in LUAD. Then, the screened genes were further filtered for biological functions and the regression coefficients of the genes were calculated using multivariate Cox regression analysis. A prognostic risk score was generated for each patient using the following formula,

$$RiskScore = \sum_{i=1}^{M} Coef_{Gene(i)} \times Exp_{Gene(i)} \quad (9)$$

where M represents the number of samples, $Coef_{Gene(i)}$ represents the regression coefficient of $Gene(i)$ and $Exp_{Gene(i)}$ represents the expression level of $Gene(i)$. The samples were divided into high-risk and low-risk groups based on the median risk score.

Enrichment analysis of the high- and low-risk groups

GO enrichment analysis was performed for differentially expressed genes between the high- and low-risk groups. Functional differences between the high- and low-risk groups were further analyzed by calculating enrichment scores for the hallmark and immunologic gene sets between the two groups using the GSVA method. The hallmark gene sets and immunologic signature gene sets were obtained from "h.all.v2023.1" and "c7.immunesigdb.v2023.1" of The Molecular Signatures Database (MSigDB) (https://www.gsea-msigdb.org/gsea/msigdb)[44], respectively.

Characterization of immune, metabolic, and other tumor-related factors

The relative infiltration abundance of 22 immune cell types in each sample was estimated based on gene expression data using the CIBERSORTx[45] online tool. The R package "estimate" calculated stromal and immune scores to estimate the abundance of stromal cells and the level of immune cell infiltration in the sample's tumor microenvironment (TME) and combined the two scores to infer tumor purity[46]. Metabolic reprogramming is recognized as one of the key features of malignant tumors, with reprogramming of glycolysis, amino acid metabolism and lipid metabolism providing a tremendous energy demand for cancer cell proliferation[47]. Angiogenesis, EMT, tumorigenic cytokines and inflammation play crucial roles in tumor growth and progression[48–51]. Angiogenesis, EMT and tumorigenic cytokines-related genes were collected from the published literature of Qiu et al.[52]. Genes related to aerobic glycolysis, metabolism of glutamine, serine, glycine, arginine, methionine, tryptophan, fatty acids and sphingolipids, and inflammatory response were downloaded from MSigDB (https://www.gsea-msigdb.org/gsea/msigdb) database. Reference collections include "h.all.v2023.1", "c2.cp.wikipathways.v2023.1", "c2.cp.reactome.v2023.1", "c2.cp.kegg.v2023.1" and "c5.go.bp.v2023.1". Metabolic reprogramming, angiogenic activity, EMT, tumorigenic cytokines and inflammatory response scores were calculated for each sample by single-sample gene set enrichment analysis (ssGSEA) using the R package "GSVA"[42].

Prediction of immunotherapy response

The R package "EaSIeR" was used to predict the likelihood of patient response to immune checkpoint inhibitors (ICI). "EaSIeR" employs five types of features describing the immune TME to construct a model and uses transcriptome-based immune response scores (cytolytic activity (CYT), Roh immune score (Roh_IS), chemokines, Davoli immune signature (Davoli_IS), IFNy signature (IFNy), Expanded immune signature (Ayers_expIS), T-cell inflamed signature (Tcell_inflamed), T-cell inflamed signature (RIR), Tertiary lymphoid structures signature (TLS)) as learning targets to predict patients' ICI response[53]. Higher scores indicate a higher likelihood that a patient will respond to ICI therapy.

Analysis of prognostic risk factors

The expression of prognostic risk factors and their correlations with immune-related genes and immune checkpoints were analyzed. Immune-related genes were downloaded from the TISIDB (http://cis.Hku.hk/TISIDB/) database[54] and 79 immune checkpoints were from Hu et al.[55].

Statistical analysis

All statistical analyses were performed based on R software (4.2.1). Kolmogorov-Smirnov test was used for testing normality. Two-tailed Student's t-test was used to estimate the difference between the two groups when the data obeyed normal distribution, otherwise the Wilcoxon rank-sum test was performed. Spearman's rank correlation analysis was employed to explore the correlation between the variables. If not mentioned, $p < 0.05$ was considered statistically significant. If not specifically labelled, for the symbolic marking of statistical significance, we used the following convention, "ns": $p > 0.05$; "*": $p <= 0.05$; "** ": $p <= 0.01$; "***": $p <= 0.001$; "****": $p <= 0.0001$.

Results

Potential biomarkers for LUAD

The TCGA database contains 539 cancer samples and 59 normal samples corresponding to LUAD. To identify stable differentially expressed genes, 1000 random samplings were performed between normal and cancer samples in a ratio of 1:6 (59:354). A total of 8836 differentially expressed genes were screened in the 1000 sample sets. Among them, 80.24% of the genes were differentially expressed with a frequency of 1000, 7471 genes had a

frequency not less than 950, and 7723 genes had a frequency not less than 750 (Fig. 2A). For each of the above 1000 sets of samples, three groups of differentially expressed genes screened according to frequency thresholds were respectively used as feature inputs for VAEjMLP. For the VAEjMLP model, it used a VAE model consisting of an encoder and decoder modeled by a neural network, respectively, to achieve feature dimensionality reduction, and input the reduced features into an artificial neural network MLP for classification prediction. In VAEjMLP, the encoder part of VAE consisted of four hidden layers, with 1024 neurons in hidden layer 1 and hidden layer 2, 512 in hidden layer 3, and 128 in hidden layer 4, and the decoder had a mirrored structure with the encoder (Fig. 2B). The latent vector of dimension 128 obtained from VAE training was passed into a three-layer MLP with 128 neurons in the first layer, 64 in the second layer and the third layer as the output layer (Fig. 2B). Both VAE and MLP used the Leaky ReLU activation function. The number of epochs was set to 250, with the first 50 epochs for unsupervised training of VAE, the 51st–100th epochs for classification training of MLP, and the 101st–250th epochs for joint training of VAEjMLP. The area under the curve (AUC) of the trained VAEjMLP model reached over 0.999, and the accuracy, precision, and recall all reached over 0.99, 0.955, and 0.994, respectively. For each set of input samples, three groups of significant contributing genes were screened using ±5%, ±1%, and ±0.5% of the distribution formed by all the contribution values of the SHAP output as thresholds, respectively. Using the biomarker potential scoring algorithm, each significant contributing gene received its scores in each set of samples. The quartiles of the descending biomarker potential scores of the three groups of significantly contributing genes were used as lower bounds to screen the genes, respectively, and the intersection was considered as candidates. When all differentially expressed genes, genes with a frequency not less than 750, and genes with a frequency not less than 950 were used as features, respectively, the number of candidate genes identified in 1000 sample sets was 127–183, 109–137, and 92–132, respectively (Fig. 2C). Eventually, the 19 shared genes of all candidates were identified as potential LABs (Fig. 2D).

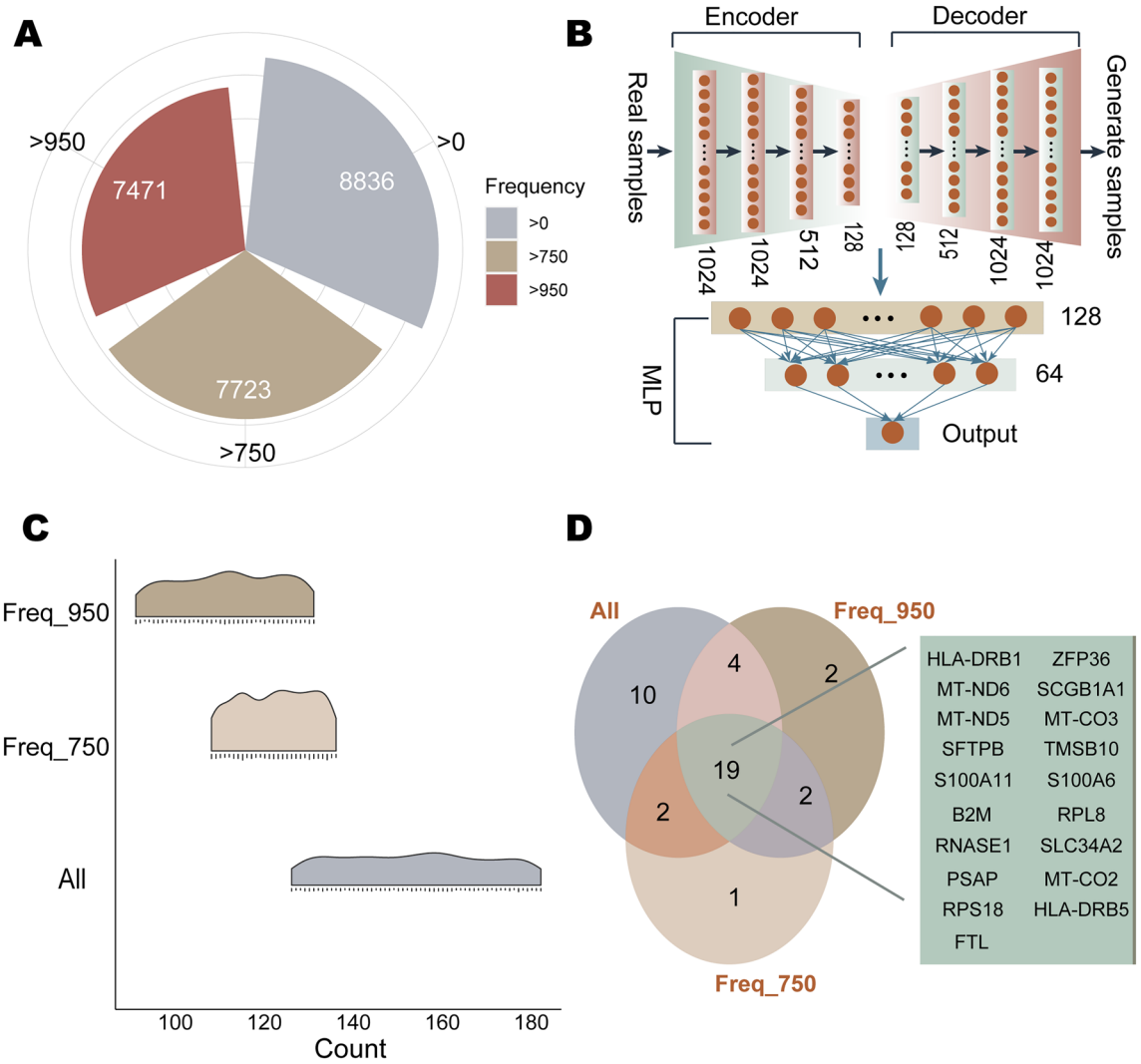

Figure 2. Potential biomarkers for LUAD. (**A**) Number of input features for the 3 sets. (**B**) Number of neurons in each hidden layer of VAEjMLP. (**C**) Number of candidate genes identified in each of the 1000 sample sets corresponding to the three sets of input features. Where "All", "Freq_750", "Freq_950" represent three sets of input features respectively. (**D**) The identified biomarkers for LUAD.

Association analysis of LABs with LUAD

LABs consist of 15 nuclear and 4 mitochondrial genes. The biological functions of nuclear genes were explored using GO and KEGG enrichment analysis. The results demonstrated that nuclear genes were mainly enriched in immune pathways, such as MHC protein complex assembly, Th1 and Th2 cell differentiation, antigen processing and presentation, regulation of T cell activation, etc. (Fig. 3A,B). In addition, four mitochondrial genes (*MT-ND6*, *MT-ND5*, *MT-CO3*, *MT-CO2*) were all involved in driving mitochondrial oxidative phosphorylation (OXPHOS) (Fig. 3C). The OXPHOS system serves as the center of cellular metabolism and is critical for energy production in eukaryotic cells[56]. Besides, the GSVA results showed that immune pathways, such as platelet activation, leukocyte transendothelial migration, natural killer cell mediated cytotoxicity, and Th1 and Th2 cell differentiation, were mostly enriched in the normal group as compared to the LUAD group (Fig. 3D). Simultaneously, the normal and LUAD groups presented significant differences in the enrichment of metabolic pathways (e.g., fructose and mannose metabolism, arginine and proline metabolism, and cholesterol metabolism) (Fig. 3E). The GSEA was also employed to evaluate the signaling pathways involved in nuclear genes. The results indicated that nuclear genes of LABs were negatively linked to immune pathways (Th17 cell differentiation, Th1 and Th2 cell differentiation, Phagosome, intestinal immune network for IgA production, and hematopoietic cell lineage) in LUAD (Fig. 3F–J). These results illustrated that LABs were implicated in the regulation of immune and metabolic functions in LUAD.

Construction and validation of a prognostic risk model

The univariate Cox regression analysis showed that *HLA-DRB1*, *SCGB1A1*, *SFTPB*, *RNASE1*, *SLC34A2*, and *HLA-DRB5* among the LABs were significantly associated with LUAD survival (Fig. 4A). Further analysis was conducted using survival-related biomarker sets that were enriched in the same GO entries. There were four sets of genes enriched in the same GO entries (Fig. 4B): *HLA-DRB1*/*HLA-DRB5*, *HLA-DRB1*/*SCGB1A1*/*HLA-DRB5*, *HLA-DRB1*/*SCGB1A1*, and *HLA-DRB1*/*SFTPB*/*HLA-DRB5*. The Kaplan-Meier survival curve analysis for all four gene sets showed significant associations with LUAD survival. More specifically, the significant *p*-values for the *HLA-DRB1*/*HLA-DRB5* (Fig. 4C), *HLA-DRB1*/*SCGB1A1* (Fig. 4D), and *HLA-DRB1*/*SFTPB*/*HLA-DRB5* (Fig. 4E) sets were 0.01, 0.0024, and 0.0014, respectively. The *HLA-DRB1*/*SCGB1A1*/*HLA-DRB5* set presented the most significant difference in survival rates (*p*-value=0.00055) (Fig. 4F). Ultimately, the set *HLA-DRB1*/*SCGB1A1*/*HLA-DRB5*, which was most correlated with LUAD survival, was selected to construct the LUAD prognostic risk model and risk scores for patients were calculated. The regression coefficients for *HLA-DRB1*, *SCGB1A1*, and *HLA-DRB5* in the LUAD prognostic risk model were −7.36e−05, −1.51e−03, and −7.43e−04, respectively. It can be observed that the patients with higher risk scores had poorer OS (Fig. 4F). The univariate Cox regression analysis suggested that pathological T, N, M and stage staging were significantly associated with survival in LUAD patients (Fig. 4G). A multivariate Cox regression analysis was performed to explore the predictive independence of the prognostic risk model by combining these survival-related clinicopathological factors with the risk score. The results demonstrated that the risk score (HR = 2.33, 95% CI = 1.38–3.9, *P* = 1e−3) served as an independent risk factor for OS in patients with LUAD (Fig. 4H). The GSE72094 dataset derived from the GEO database validated the accuracy of the prognostic risk model (Fig. 4I,J). The association of the risk score with survival status and prognostic risk factors was further explored (Fig. 4K). Overall, a higher risk score was associated with a higher number of deaths, indicating a poorer prognosis and higher risk of death in the high-risk group. Compared to the low-risk group, the prognostic risk factors *SCGB1A1*, *HLA-DRB1*, *HLA-DRB5* were all down-regulated in the high-risk group.

Enrichment analysis of the prognostic risk model

Genes differentially expressed in the high- and low-risk groups were identified, and 459 up-regulated genes and 352 down-regulated genes were recognized in the high-risk group (Fig. 5A). The up-regulated genes were mainly enriched in various metabolic pathways, secretion and transportation of substances, and pathways related to development (Fig. 5B). The down-regulated genes were predominantly enriched in various immune-related pathways (Fig. 5C). To further explore the differences in biological functions of different risk groups, we conducted GSVA analysis. Among the 50 cancer hallmark gene sets, significant enrichment differences were observed in 37 gene sets between the high and low-risk groups (Fig. 5D). Compared with the low-risk group, the high-risk group was mainly accompanied by the enrichment of cell proliferation pathways (such as E2F targets, G2M checkpoint, MYC targets, etc.) and metabolic pathways (such as MTORC1 signaling, glycolysis, oxidative phosphorylation, etc.) (Fig. 5D). Also of concern was the significant down-regulation of the activity of immune-related pathways in the high-risk group compared to the low-risk group, such as IL2/STAT5 signaling, IL6/JAK/STAT3 signaling, inflammatory response, and other pathways (Fig. 5D). Both GO and GSVA analyses illustrated that metabolism-related pathways were primarily enriched in the high-risk group and immune-related pathways were mostly enriched in the low-risk group. The characteristics of unlimited tumor proliferation often require alterations in energy metabolism[57], and studies have indicated that increased expression of glycolytic enzymes correlates with poor prognosis in lung cancer patients[58,59]. The immune system plays a dual role of both promotion and inhibition of cancer development[60], and the signaling pathways IL2/STAT5 and IL6/JAK/STAT3 were closely associated with the prognosis of pancreatic ductal adenocarcinoma and olfactory neuroblastoma[61,62]. The immunologic signature gene sets derived from the MSigDB database represent immune states and perturbations signatures. Therefore, we used GSVA analysis to compare the differences in the enrichment of immunologic signature gene sets between the high- and low-risk groups. Among the 4,872 immunologic signature gene sets, there were differences in enrichment in 3,574 gene sets between the two risk groups. Specifically, the high-risk group exhibited significantly lower enrichment scores in 2,585 immunologic signature gene sets compared to the low-risk group.

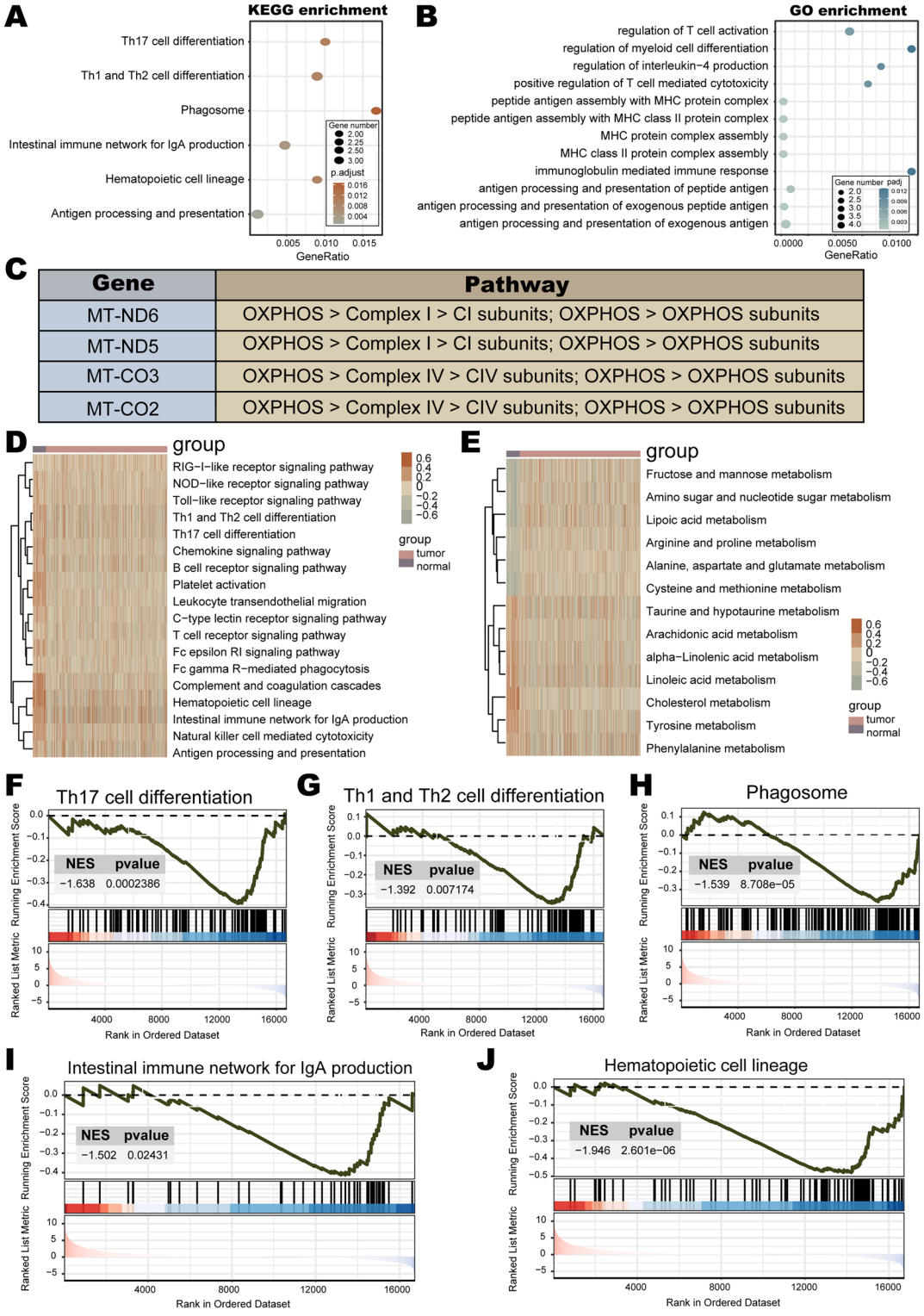

Figure 3. Association analysis of LUAD biomarkers (LABs) with LUAD. (**A**) KEGG, and (**B**) GO enrichment analysis of LABs. (**C**) Pathway annotation of mitochondrial genes. (**D**–**E**) GSVA analyses in TCGA LUAD. (**F**–**J**) GSEA analyses based on TCGA LUAD.

Not only that, the top 100 gene sets with the most significant differences were all showed down-regulation of activity in the high-risk group (Fig. 5E presents the top 20 gene sets with the most significant differences).

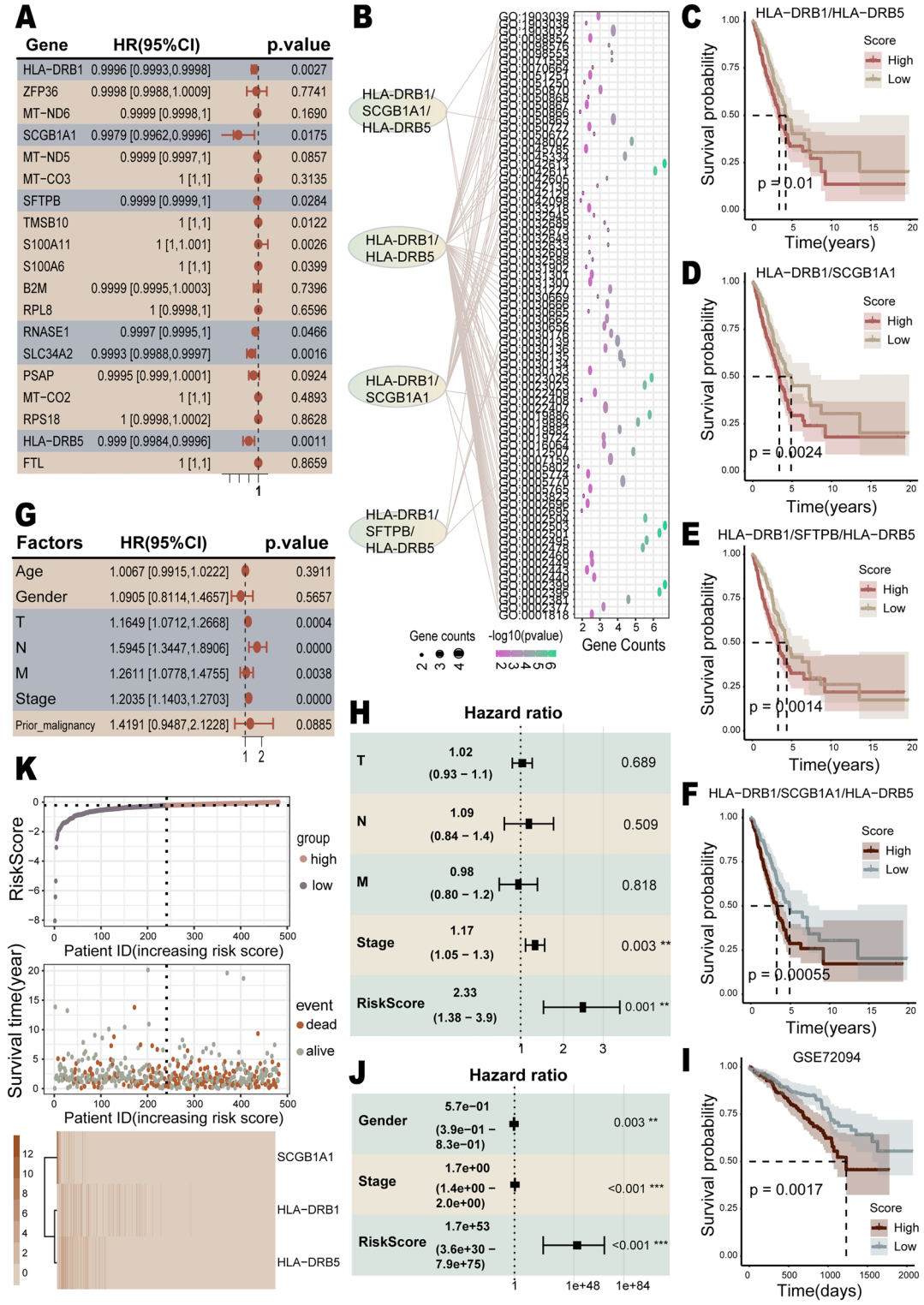

Figure 4. The construction of the prognostic risk model. (**A**) Forest plot of univariate Cox regression analysis for the LABs. Genes with *p*-value < 0.05 are highlighted with a grayish blue background. (**B**) Enriched for genes with the same GO entry. (**C**–**F**) Kaplan–Meier survival curve analysis for the gene sets *HLA-DRB1/HLA-DRB5*, *HLA-DRB1/SCGB1A1*, *HLA-DRB1/SFTPB/HLA-DRB5*, and *HLA-DRB1/SCGB1A1/HLA-DRB5*. (**G**) Forest plot of univariate Cox regression analysis for clinical environmental factors. Factors with *p*-value < 0.05 are highlighted with a grayish blue background. (**H**) Forest plot of multivariate Cox regression analysis of the risk score combined with survival-related clinical factors. (**I**) Kaplan–Meier survival curve analysis of the validation set. (**J**) Multivariate Cox regression analysis of the validation set. (**K**) Upper panel: risk scores of patients in ascending order. The horizontal dashed line represents the median risk score. Middle panel: survival time of patients in ascending order of risk score. Bottom panel: heatmap of gene expression levels in patients sorted by ascending risk score.

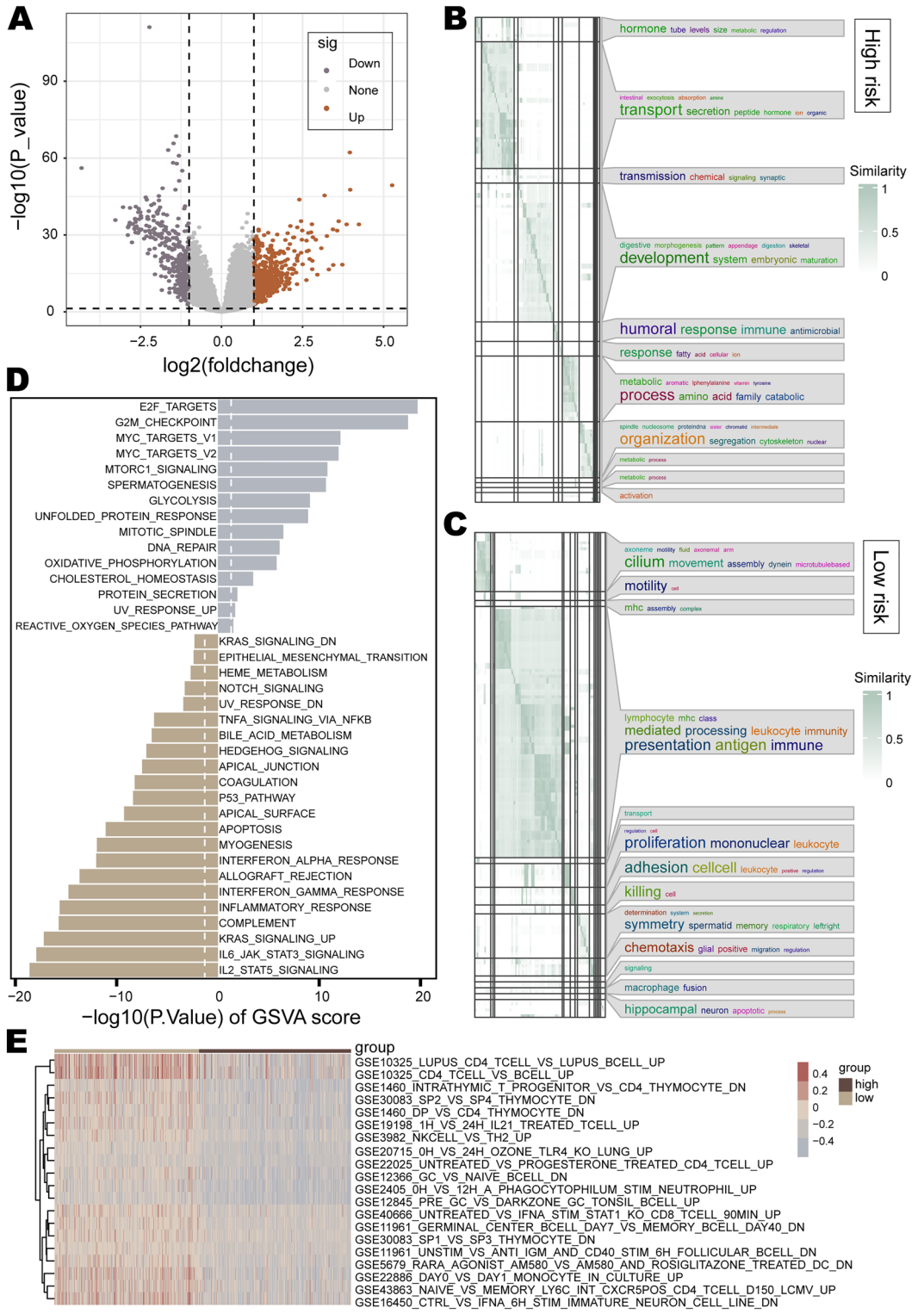

Figure 5. Enrichment analysis of the prognostic risk model. (**A**) Differentially expressed genes between the high- and low-risk groups. (**B**, **C**) Clustering maps of GO entries enriched by up-regulated genes in the high-risk and low-risk groups. Word cloud annotations are appended on the right, and font sizes reflect word frequencies. (**D**) Enrichment analysis of cancer hallmark gene sets with differential enrichment between high-risk and low-risk groups. Vertical dashed lines represent significance thresholds. (**E**) Heatmap of enrichment scores for immunologic signature gene sets differentially enriched in the high-risk and low-risk groups.

Immune landscapes for the prognostic risk model

The high-risk group presented significantly down-regulation of immune pathway activity compared to the low-risk group. Therefore, the CIBERSORT algorithm was employed to calculate the immune infiltration of the high-risk and low-risk groups, while the ESTIMATE algorithm was utilized to assess the immune TME in both groups. The level of infiltration of 22 immune cell types in LUAD patients was estimated using the CIBERSORT algorithm. The proportions of 13 immune cell types infiltrated showed significant differences between the two risk groups. The infiltration proportions of B cells naive, eosinophils, macrophages M0, mast cells activated, plasma cells, and T cells CD4 memory activated were higher in the high-risk group than in the low-risk group (Fig. 6A). The correlation between the risk score and the level of infiltration of these six immune cells was significant and positive (Fig. 6B). B cells memory, dendritic cells resting, macrophages M2, mast cells resting, monocytes, T cells CD4 memory resting, and T cells regulatory (Tregs) infiltration levels were higher in the low-risk group (Fig. 6A). The infiltration proportions of these 7 immune cell types showed a significant and negative correlation with the risk score (Fig. 6B). Several LUAD survival studies have demonstrated that macrophages M0, mast cells activated and T cells CD4 memory activated were accompanied by higher levels of infiltration in high-risk groups, and infiltration levels of B cells memory, dendritic cells resting, mast cells resting, monocytes, T cells CD4 memory resting and T cells regulatory (Tregs) were higher in low-risk groups[63–65]. This aligns with the findings of our study. Stromal scores for stromal cell abundance, immune scores for the extent of immune cell infiltration and tumor purity in the TME were estimated using ESTIMATE for different risk groups. The results indicated that patients in the high-risk group had higher tumor purity than those in the low-risk group ($p < 2.2e{-}16$) (Fig. 6C). The risk score was positively correlated with tumor purity in LUAD patients (R = 0.55, $p < 2.2e{-}16$) (Fig. 6D). Compared to the low-risk group, patients in the high-risk group had significantly lower stromal scores ($p = 3.2e{-}16$), immune scores ($p < 2.2e{-}16$), and ESTIMATE scores ($p < 2.2e{-}16$) (Fig. 6E,G,I). The risk score showed negative correlations with the stromal score (R = − 0.43, $p < 2.2e{-}16$), the immune score (R = − 0.59, $p < 2.2e{-}16$), and the ESTIMATE score (R = − 0.55, $p < 2.2e{-}16$) (Fig. 6F,H,J). Taken together, we thought that the immune response effect in the high-risk group may be inferior to that of the low-risk group. Multiple studies have separately constructed survival models for LUAD, and their findings consistently demonstrated that LUAD patients with high-risk scores exhibited lower immune scores, stromal scores, ESTIMATE scores, and higher tumor purity[64,66–68]. This provides valid support for our findings.

Metabolism landscapes and tumor-related factors analysis for the prognostic risk model

Cancer cells reprogram their metabolic pathways to meet the needs of tumor initiation and progression[69,70]. GO and GSVA analyses revealed significant metabolic differences between the high- and low-risk groups. The metabolic reprogramming differences between the high- and low-risk groups were investigated by calculating the reprogramming scores of glycolysis, amino acid metabolism, and lipid metabolism in LUAD patients. Aerobic glycolysis is a key metabolic hallmark of cancer, promoting the survival and proliferation of cancer cells. The preference of cancer cells for glycolysis is often associated with poorer clinical outcomes[71]. The risk score was positively correlated with the aerobic glycolysis reprogramming score (R = 0.29, $p = 1.1e{-}10$) (Fig. 7A). The patients in the high-risk group had significantly higher aerobic glycolysis reprogramming scores ($p = 2.9e{-}06$) than those in the low-risk group (Fig. 7A). Amino acid metabolism is one of the pivotal nutrients supporting cancer cell growth, and the metabolism of glutamine, methionine, serine, glycine, arginine, and tryptophan is deregulated in many cancers[72–76]. The risk score was positively correlated with the metabolic reprogramming scores of glutamine (R = 0.58, $p < 2.2e{-}16$) and methionine (R = 0.15, $p = 1.2e{-}3$) (Fig. 7B,C). The metabolic reprogramming scores of glutamine ($p < 2.2e{-}16$) and methionine ($p = 2.4e{-}2$) were both higher in the high-risk group than in the low-risk group (Fig. 7B,C). Studies have shown that the metabolism of glutamine and methionine not only drives tumor cell growth, but also creates a TME that benefits from tumor immune escape[77–79]. Meanwhile, significant positive correlations were observed between the risk score with the metabolic reprogramming scores of serine (R = 0.37, $p < 2.2e{-}16$) and glycine (R = 0.39, $p < 2.2e{-}16$) (Fig. 7D,E). The high-risk group presented higher metabolic reprogramming scores of serine ($p = 3e{-}09$) and glycine ($p = 3.6e{-}13$) (Fig. 7D,E). The upregulation of serine\glycine metabolism is associated with poor clinical outcomes in several cancers[80]. While the patients' metabolic reprogramming scores for arginine (R = − 0.32, $p = 8e{-}13$) and tryptophan (R = − 0.24, $p = 5.8e{-}08$) were negatively correlated with the risk scores (Fig. 7F,G). The metabolic reprogramming scores of arginine ($p = 3.2e{-}08$) and tryptophan ($p = 1.3e{-}4$) were lower in the high-risk group (Fig. 7F,G). Lipids are acting as central players in cancer biology[81], and metabolic imbalances in fatty acids and sphingolipids assume a crucial role in cancer development[82,83]. There were significant negative correlations between the risk scores with the metabolic reprogramming scores of fatty acids (R = − 0.15, $p = 8.2e{-}4$) and sphingolipids (R = − 0.28, $p = 2.8e{-}10$) (Figure 7H,I). The patients with higher metabolic reprogramming scores for fatty acids ($p = 5e{-}2$) and sphingolipids ($p = 8.9e{-}07$) tended to be accompanied by lower risk scores (Fig. 7H,I).

In addition, angiogenesis, EMT, tumorigenic cytokines and inflammation play important and complex roles in the progression of malignant tumors. Thereby, angiogenic activity score, EMT score, tumorigenic cytokine score and inflammatory response score were evaluated for the prognostic risk model, respectively. The angiogenesis activity score in the high-risk group was significantly higher than that in the low-risk group ($p = 2.9e{-}08$) (Fig. 7J). There was a significant trend of positive correlation between the risk score and the angiogenesis activity score (R = 0.32, $p = 9.5e{-}13$) (Fig. 7J). Angiogenesis is associated with poor prognosis in several cancers, including NSCLC, and anti-angiogenic drugs are being developed[84,85]. Patients in the high-risk group had significantly lower EMT scores ($p = 2e{-}05$), tumorigenic cytokine scores ($p = 9.4e{-}08$) and inflammatory response scores ($p < 2.2e{-}16$) than those in the low-risk group (Fig. 7K–M). The risk score was negatively correlated with the EMT score (R = − 0.2, $p = 6.5e{-}06$), tumorigenic cytokine score (R = − 0.3, $p = 3.9e{-}11$), and inflammatory response score (R = − 0.51, $p < 2.2e{-}16$) (Fig. 7K–M).

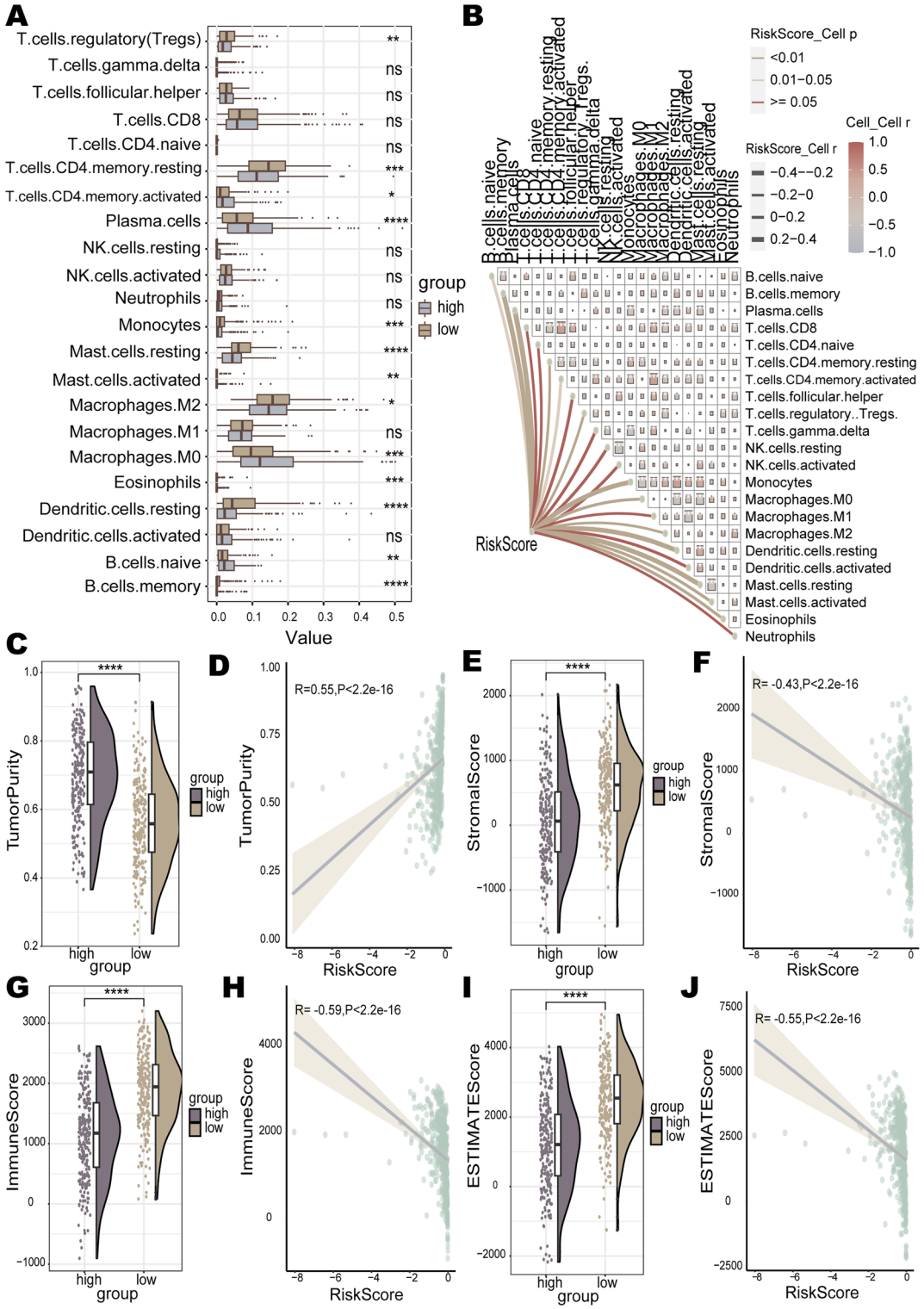

Figure 6. Immune landscape of the prognostic risk model. (**A**) Infiltration levels of 22 immune cells in high- and low-risk groups. (**B**) The correlation between risk scores and immune cell infiltration levels in patients. (**C**) Tumor purity, (**E**) stromal score, (**G**) immune score, and (**I**) ESTIMATE score in the high- and low- risk groups. Correlation of risk scores with (**D**) tumor purity, (**F**) stromal score, (**H**) immune score, and (**J**) ESTIMATE score.

In conclusion, patients with poor prognosis experienced significant changes in metabolic reprogramming, angiogenesis, EMT, tumorigenic cytokines, and inflammation.

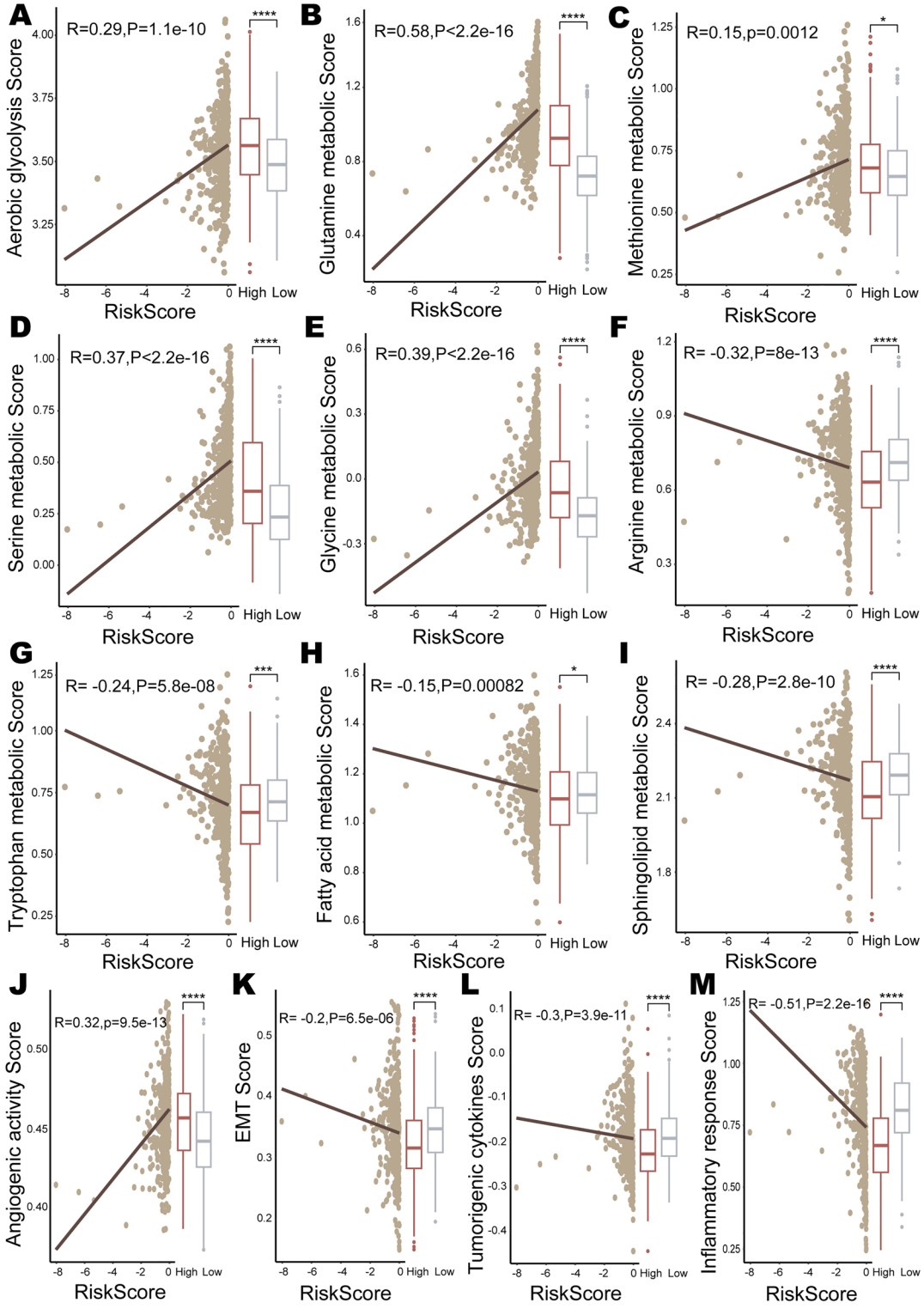

Figure 7. Metabolism landscapes and tumor-related factors analysis for the prognostic risk model. Correlations between risk scores and metabolic reprogramming scores for (**A**) aerobic glycolysis, (**B**) glutamine, (**C**) methionine, (**D**) serine, (**E**) glycine, (**F**) arginine, (**G**) tryptophan, (**H**) fatty acids, and (**I**) sphingolipids. The correlations between risk scores and scores for (**J**) angiogenesis activity, (**K**) EMT, (**L**) tumorigenic cytokines, and (**M**) inflammatory response. "High" and "Low" represent the high-risk and low-risk groups, respectively.

Characterization of prognostic risk factors

We further analyzed the prognostic risk factors *HLA-DRB1*, *SCGB1A1* and *HLA-DRB5*. They were all significantly

lower expressed in the high-risk group than in the low-risk group (Fig. 8A–C). *HLA-DRB1*, *SCGB1A1* and *HLA-DRB5* were accompanied by lower expression in LUAD tissues compared to their expression in normal lung tissues (Fig. 8A–C). The scRNA-seq dataset GSE131907 was used to explore the expression of prognostic risk factors in normal lung tissue and primary LUAD. With data preprocessing and quality control, a total of 92,951 single cells were captured and clustered into nine cell types including endothelial cells, epithelial cells, monocytes, etc. Similarly, the expression of prognostic risk factors was observed at the single-cell transcriptome level were all significantly lower in the disease state than in the normal state (Fig. 8D–F). Differential expression of prognostic risk factors was also compared in different cell types between normal and LUAD states. *HLA-DRB1* exhibited differential expression between normal and disease states in all cell types except MAST cells (Fig. 8G). In the LUAD state, *HLA-DRB1* expression was significantly higher in all cell types except B Cells and Monocyte than in the normal state (Fig. 8G). *SCGB1A1* expression was significantly higher in all cell types in the normal state

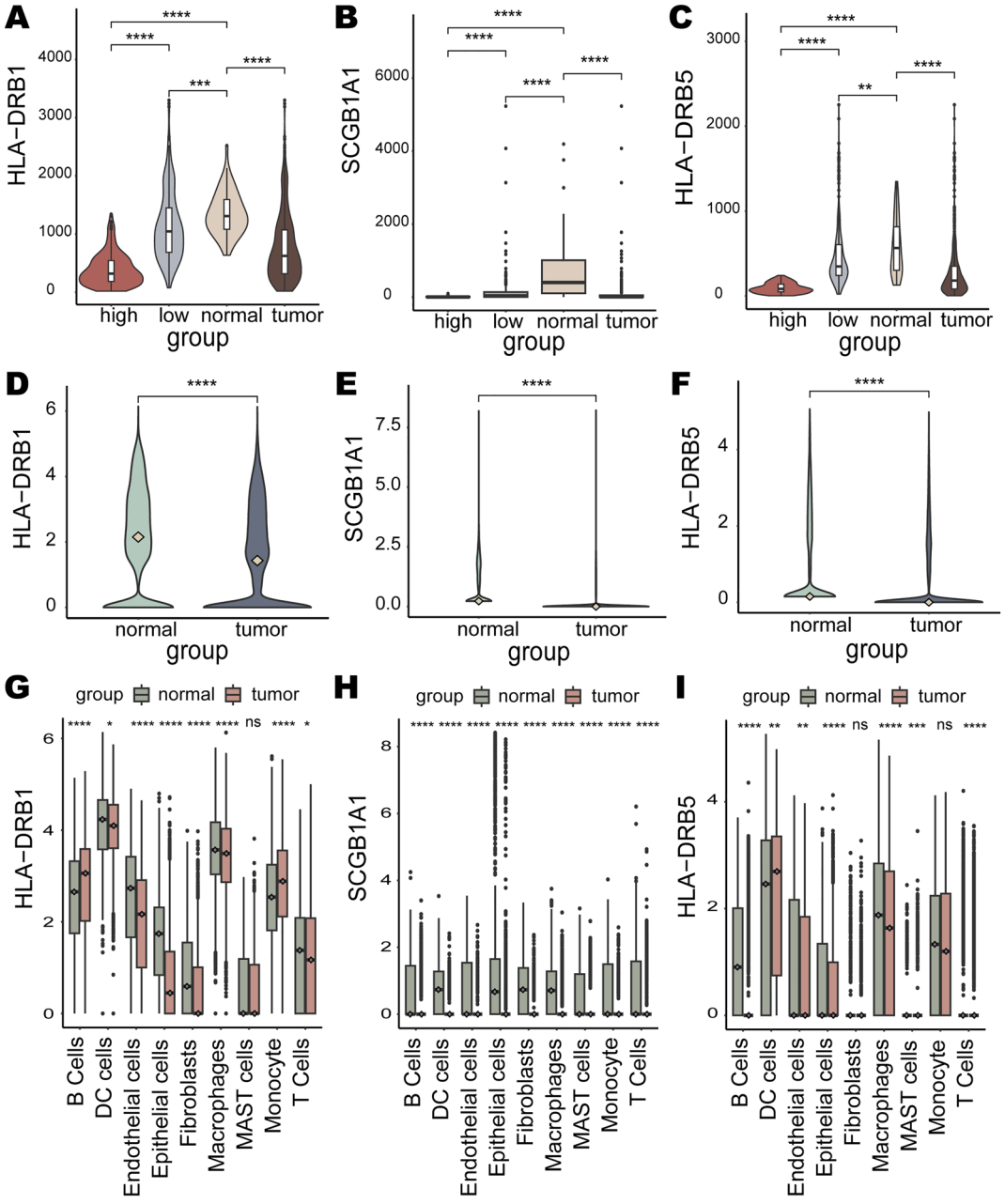

Figure 8. Characterization of the expression of prognostic risk factors. Expression of (**A**) *HLA-DRB1*, (**B**) *SCGB1A1*, and (**C**) *HLA-DRB5* in normal tissues, cancerous tissues and high- and low-risk groups. Overall expression of (**D**) *HLA-DRB1*, (**E**) *SCGB1A1*, and (**F**) *HLA-DRB5* at the single-cell level in normal and cancer states. Expression of (**G**) *HLA-DRB1*, (**H**) *SCGB1A1*, and (**I**) *HLA-DRB5* in various cell types in normal and cancer states.

than in the LUAD state (Fig. 8H). *HLA-DRB5* showed significant expression differences between normal and disease states in B cells, DC cells, endothelial cells, epithelial cells, macrophages, MAST cells, and T cells (Fig. 8I).

Recently, targeting TME has emerged as a promising method for cancer treatment[86]. Immune-related genes, as important components of TME, play a crucial role in tumorigenesis and TME homeostasis[87]. In parallel, immune checkpoints can modulate the degree of immune activation in TME[88]. Co-expression analysis was performed to explore the correlation of the prognostic risk factors *HLA-DRB1*, *SCGB1A1* and *HLA-DRB5* with immune-related genes (chemokines, chemokine receptors, MHC, immunostimulatory genes, immunosuppressive genes) and immune checkpoints. The prognostic risk factors *HLA-DRB1* and *HLA-DRB5* are immune checkpoints, of which *HLA-DRB1* is also an MHC gene. The prognostic risk factors *HLA-DRB1* (Fig. 9), *SCGB1A1* (Supplementary Fig. S1), and *HLA-DRB5* (Supplementary Fig. S2) were positively correlated with the majority of immune-related genes and immune checkpoints. PD-1 (PDCD1), PD-L1 (CD274), CTLA4 and LAG3 are common immune checkpoints focused on by researchers. *HLA-DRB1* was positively correlated with those four immune checkpoints (*PD-1*: R = 0.36, p = 2.22e−16; *PD-L1*: R = 0.43, p = 0; *CTLA4*: R = 0.42, p = 0; *LAG3*: R = 0.27, p = 1.13e−09), as was *HLA-DRB5* (*PD-1*: R = 0.30, p = 8.75e−12; *PD-L1*: R = 0.38, p = 0; *CTLA4*: R = 0.36, p = 8.88e−16; *LAG3*: R = 0.20, p = 1.44e−05), whereas *SCGB1A1* showed a positive correlation with *CTLA4* (R = 0.13, p = 4.6e−3). The findings of Aram et al.[89] showed that HLA-DRB1 was associated with anti-PD-1/PD-L1 therapy response in NSCLC. A recent study demonstrated the ability of fucosylation of HLA-DRB1 to enhance the efficacy of immune checkpoint blockade anti-PD1 drugs in melanoma[90].

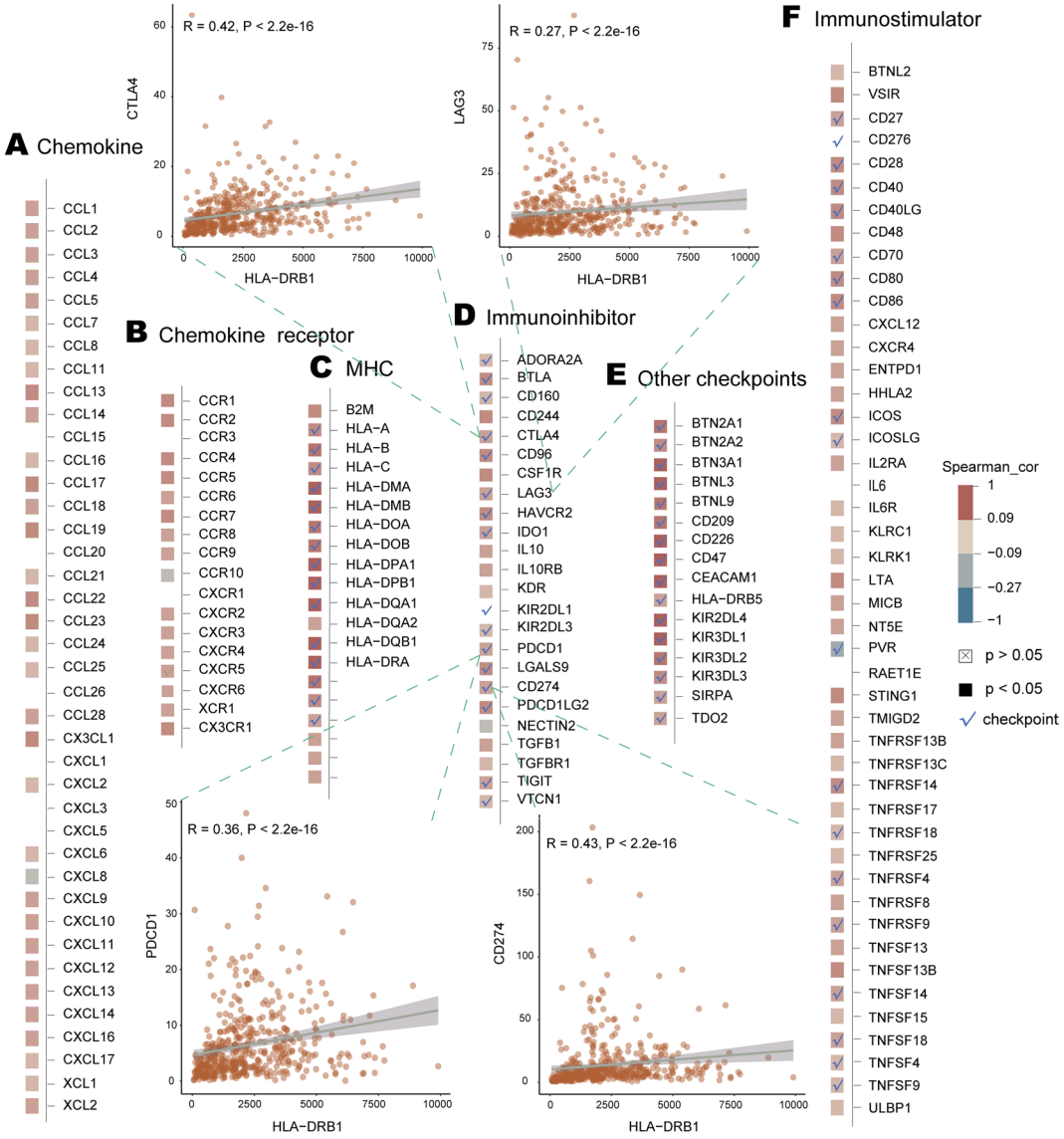

Figure 9. The correlation of *HLA-DRB1* with immune-related genes and immune checkpoints: (**A**) chemokine genes, (**B**) chemokine receptor genes, (**C**) MHC, (**D**) immunosuppressive genes, (**E**) other immune checkpoints, and (**F**) immunostimulatory genes. "other immune checkpoints" means immune checkpoints other than immune-related genes.

Prediction of immunotherapy response and analysis of therapeutic benefit in different risk groups

The immune checkpoint blockade therapy has made revolutionary progress in the treatment of human cancers, including LUAD[91]. Hence, we employed the "EaSIeR" to predict the response scores to ICI therapy in the high-risk and low-risk groups. It was shown that the ICI treatment response score of the low-risk group was significantly higher than that of the high-risk group ($p < 2e{-}16$) (Fig. 10A). And the score was significantly negatively correlated with the risk score ($R = -0.52$, $p < 2.2e{-}16$) (Fig. 10B). Additionally, "EaSIeR" further provided nine indicators (CYT, Roh_IS, chemokines, Davoli_IS, IFNy, Ayers_expIS, Tcell_inflamed, TLS, and RIR) for the evaluation of the immune response. Eight of these scores, except the RIR score, were higher in the low-risk group compared to the high-risk group, and these eight scores were negatively correlated with the risk score (Fig. 10C–K). This indicated a stronger immune response in the low-risk group. For the common immune checkpoints PD-1, PD-L1, CTLA4 and LAG3, inhibitors targeting them have already received regulatory approval or are undergoing clinical trials[92]. The expression levels of *PD-1* ($p = 3.1e{-}08$), *PD-L1* ($p = 9.7e{-}09$), *CTLA4* ($p = 1.4e{-}08$), and *LAG3* ($p = 8e{-}05$) were higher in the low-risk group (Fig. 10L–O). These portend that patients in the low-risk group may benefit more from immune checkpoint blockade.

The pRRophetic algorithm was applied to evaluate sensitivity differences between the high- and low-risk groups to drugs in the Cancer Genome Project database, in which eight NSCLC therapeutic drugs were included[93]. The results showed that the high-risk group had significantly lower IC50 values than the low-risk group for Afatinib ($p = 1.9e{-}2$), Gefitinib ($p = 7.5e{-}06$), and Gemcitabine ($p = 5e{-}2$), suggesting that high-risk patients may benefit more from these drugs (Fig. 11A–C). Crizotinib ($p = 2.8e{-}2$) and Erlotinib ($p = 2e{-}04$) had lower IC50 values in the low-risk group, indicating that these drugs may be more suitable for low-risk patients (Fig. 11D–E). However, there was no significant difference in sensitivity to Docetaxel, Paclitaxel, and Vinorelbine between the two groups (Fig. 11F–H).

Discussion

LUAD is a cancer with a poor prognosis[94], and the search for available LUAD biomarkers is important to achieve effective medical management of patients and improve their prognosis. In this study, we proposed an interpretable deep learning framework based on VAE, MLP, and SHAP that has mapped a large number of genes to a low-dimensional latent representation, enabling efficient feature selection. Considering the SHAP value reflected the influence of the feature genes as well as the positivity and negativity of the influence, a potential biomarker scoring algorithm was developed and 19 LUAD biomarkers were successfully identified. These biomarkers were enriched in the metabolic and immune pathways, revealing that LUAD patients present significant immune alterations and metabolic disturbances.

Given the importance of prognostic prediction in guiding precise therapeutic decisions, a 3-gene prognostic risk model for LUAD was developed through functional profiling and survival significance testing of identified biomarkers. The high- and low-risk groups divided by the risk score showed differences in the immune and metabolic landscapes. The low-risk group featured higher immune pathway activity, immune infiltration, and ICI treatment response scores. This indicates that high-risk patients were under an immunosuppressive state and that low-risk patients may be more suitable for ICI therapy. Interestingly, the three prognostic risk factors were positively related to the majority of immune-related genes. They were differentially expressed genes between the high-risk and low-risk groups, enriched in immune-related pathways such as regulation of T cell activation and leukocyte cell-cell adhesion, especially among them HLA-DRB1 and HLA-DRB5, which are immune checkpoints. Anti- PD-1/PD-L1 has yielded significant benefits for patients with NSCLC by inhibiting immune checkpoint activity, and the study demonstrated that HLA-DRB1 enhances the efficacy of anti-PD-1 therapy[90]. These suggest that the screened prognostic risk factors may be promising targets for immunotherapy of LUAD. Whereas the high-risk group presented significant enrichment of metabolic pathways. The high-risk patients had a higher propensity for metabolic reprogramming of aerobic glycolysis, glutamine, methionine, serine, and glycine, while having a lower propensity for metabolic reprogramming of arginine, tryptophan, fatty acids, and sphingolipids. Aerobic glycolysis and the metabolism of serine and glycine were related to poorer prognosis in cancers[71,80]. The close relationship between the prognostic risk model and metabolic reprogramming implies that targeting metabolism may contribute to improving the clinical prognosis of LUAD patients. Notably, increasing evidence suggests that altered metabolic patterns not only provide a survival advantage for tumor cells, but are also critical in modulating anti-tumor immune responses[95,96] Methionine depletion and metabolic byproducts of aerobic glycolysis led to antitumor immune impairment[79,97]. Blocking glutamine metabolism has been shown to enhance anticancer immunity[98]. The study by Omkar et al.[99] revealed that modification of fatty acid metabolism in CAR-T cells can improve their anti-tumor immune capacity. These findings revealed that the immune system and metabolism influence each other in cancer progression.

Angiogenesis, EMT, tumorigenic cytokines, and inflammation are correlates of tumor development. For the risk groups divided by risk score, the high-risk group presented higher angiogenic activity and lower EMT, tumorigenic cytokines, and inflammatory response. Studies have shown that angiogenesis is a poor prognostic indicator for survival in NSCLC[100,101]. The significant variations of these factors between the high- and low-risk groups may provide ideas for the treatment of LUAD. Of concern, EMT can stimulate cancer cells to produce pro-inflammatory factors, while inflammation can in turn induce EMT[102]. Not only that, EMT is also closely related to anti-tumor immune response[103]. In addition, researchers have found that the tumor vessel and the immune system affect each other's function, and that the combination of anti-angiogenesis therapies and immunotherapies may produce better therapeutic outcomes[104]. Aerobic glycolysis is associated with angiogenesis, a phenomenon closely associated with tumors[105]. Researchers have found that production of the tumorigenic cytokine IL-5 is associated with an enhanced response to immune checkpoint blockade in breast cancer patients[106]. Processes such

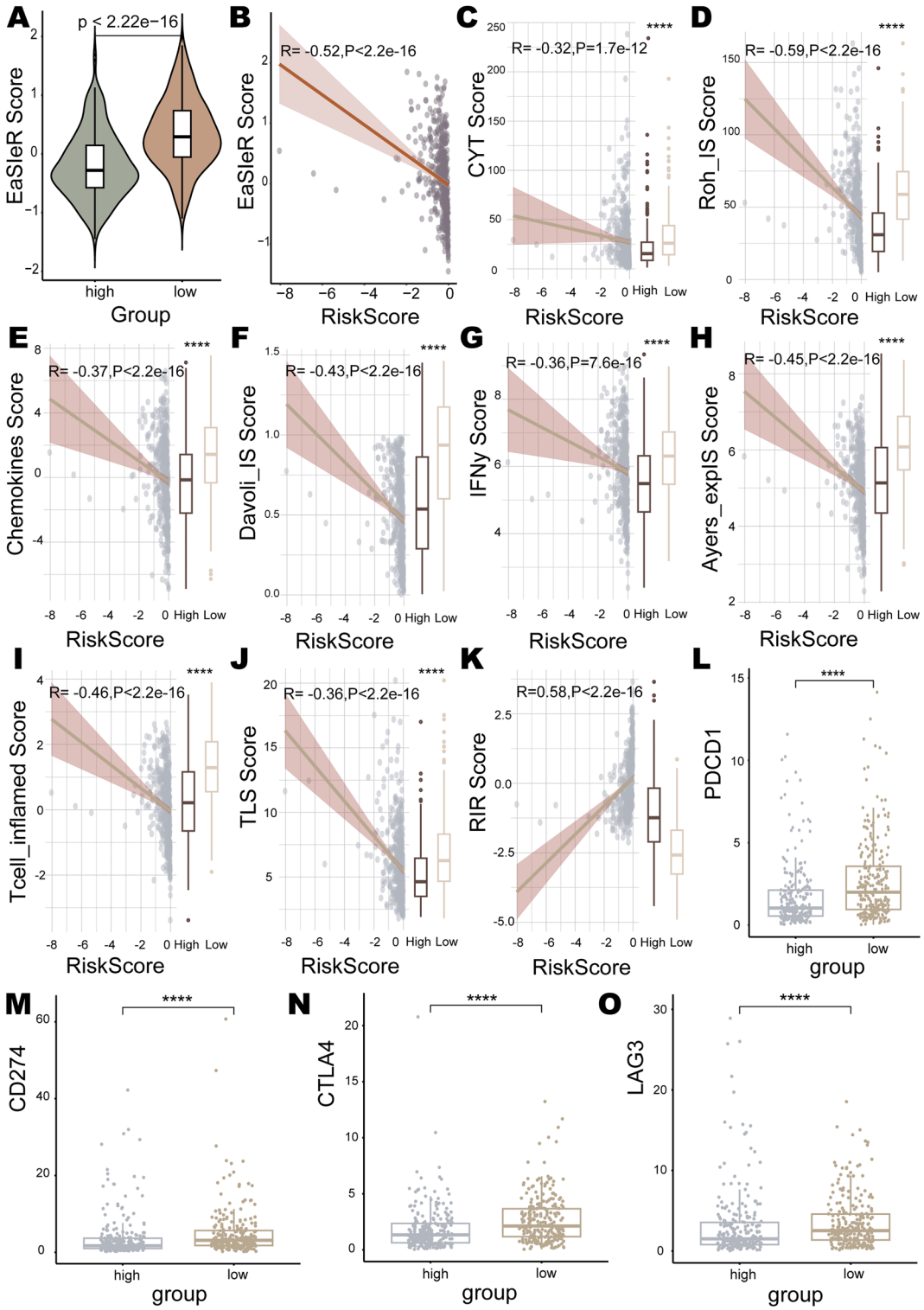

Figure 10. Immunotherapy response prediction for different risk groups. (**A**) Immune checkpoint inhibitors (ICI) treatment response scores for high and low risk groups. (**B**) Correlation of the patients' risk score with the ICI treatment response score. Correlation of the risk score with the scores of (**C**) CYT, (**D**) Roh_IS, (**E**) chemokines, (**F**) Davoli_IS, (**G**) IFNy, (**H**) Ayers_expIS, (**I**) Tcell_inflamed, (**J**) TL, and (**K**) RIR. Expression levels of immune checkpoints (**L**) PD-1, (**M**) PD-L1, (**N**) CTLA4, and (**O**) LAG3 in high-risk and low-risk groups.

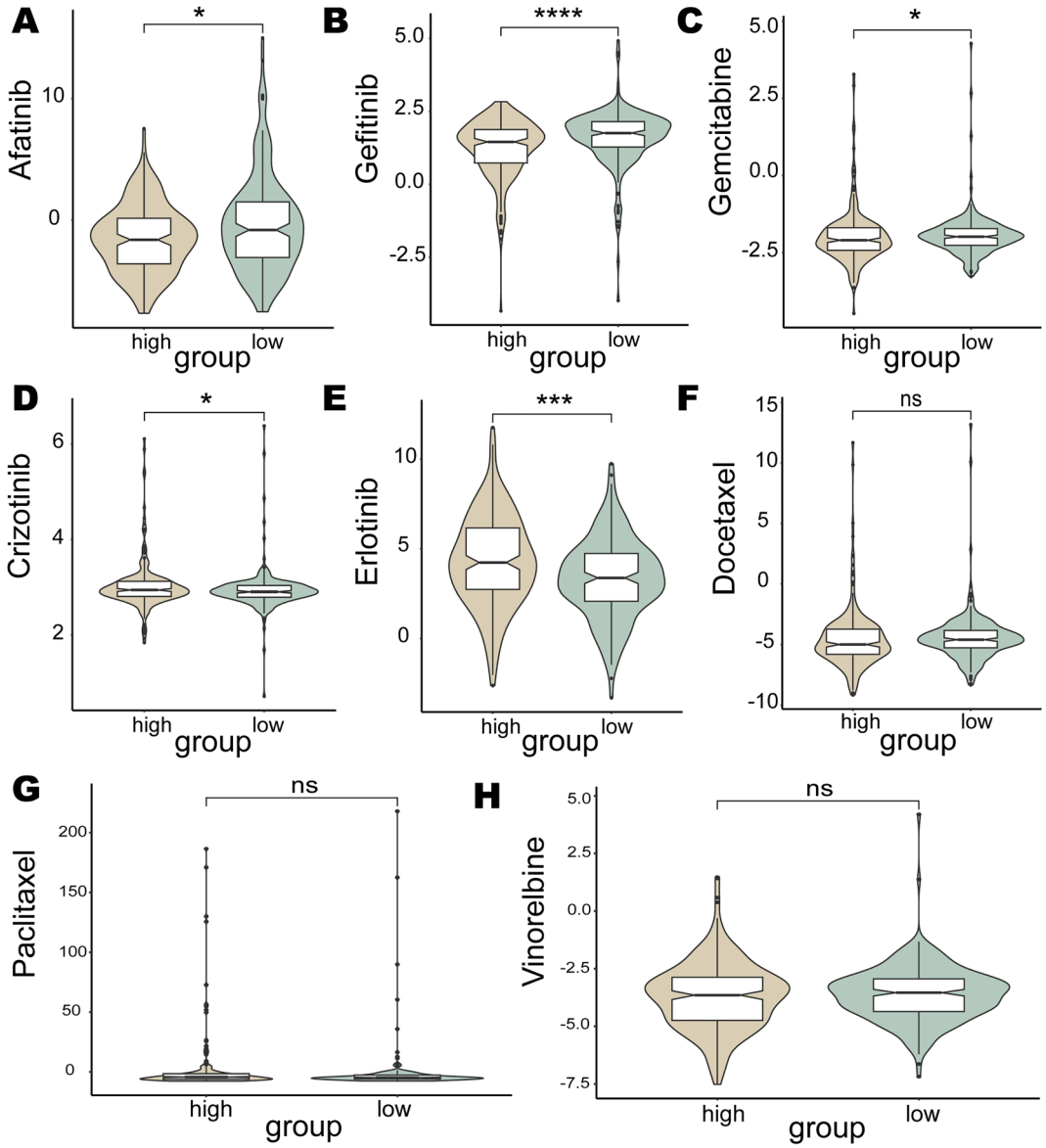

Figure 11. Benefit analysis of known therapeutic drugs in different risk groups. Sensitivity of high-risk and low-risk groups to known therapeutic drugs: (**A**) Afatinib, (**B**) Gefitinib, (**C**) Gemcitabine, (**D**) Crizotinib, (**E**) Erlotinib, (**F**) Docetaxel, (**G**) Paclitaxel, and (**H**) Vinorelbine.

as immune, metabolism, EMT, inflammation, and angiogenesis may crosstalk with each other, and combinatory interventions on them may present promising strategies for potential synergy.

To further explore the diagnostic potential of identified biomarkers, five common classification models were used, including a modified version of the integrated model AdaBoost.M1[39], the generalized linear model GLM[107], the L2 regularized linear support vector machine with class weights SVMLinearWeights2[108], the penalized logistic regression model PLR[109] and the extreme gradient boosting model XGBDART[110]. The five classification models were trained and evaluated on the TCGA dataset and another independent dataset GSE81089 from GEO (involving 108 LUAD samples and 19 normal samples) in a leave-one-out method. The models were trained using the R package "caret" and the AUC was used as the evaluation metric for the models. For the TCGA LUAD dataset, the AUC values of the AdaBoost.M1, GLM, SVMLinearWeights2, PLR, and XGBDART models were 0.9012, 0.875, 0.8929, 0.891, and 0.8918, respectively (Supplementary Fig. S3A). For the GSE81089 dataset, the AUC values for the five classification models reached 0.8855, 0.9769, 0.9242, 0.9454, and 0.8684, respectively (Supplementary Fig. S3B). To better understand the performance of the models, performance metrics including accuracy, precision, recall, and F1-score were also calculated. The five classification models achieved accuracy, precision, recall, and F1-score of over 94% on both the TCGA dataset and the GSE81089 dataset (Supplementary Table S1). The excellent performance of multiple classification models on different datasets reflected the diagnostic superiority of the biomarkers for LUAD.

Conclusion

Collectively, this study proposed a novel deep learning framework and successfully identified 19 biomarkers and 3 prognostic signatures of LUAD. The immune and metabolic landscapes as well as alterations in angiogenesis, EMT, tumorigenic cytokines, and inflammation based on prognostic signatures were investigated. In the meantime, the prognostic signatures may be indicators for predicting response to immunotherapy, but further research is needed to clarify this point. Our study may provide new insights for screening genomic markers for LUAD, provide support for clinical judgment of LUAD prognosis and have potential to be applied to other cancers.

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

Received: 17 October 2023; Accepted: 30 December 2023
Published online: 04 January 2024

References

1. Mullard, A. Addressing cancer's grand challenges. *Nat. Rev. Drug Discov.* **19**, 825–826. https://doi.org/10.1038/d41573-020-00202-0 (2020).
2. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249. https://doi.org/10.3322/caac.21660 (2021).
3. Zhang, N. *et al.* Circular RNA circSATB2 promotes progression of non-small cell lung cancer cells. *Mol. Cancer* **19**, 101. https://doi.org/10.1186/s12943-020-01221-6 (2020).
4. Shukla, S. *et al.* Development of a RNA-seq based prognostic signature in lung adenocarcinoma. *J. Natl. Cancer Inst.* https://doi.org/10.1093/jnci/djw200 (2017).
5. Hirsch, F. R. *et al.* Lung cancer: Current therapies and new targeted treatments. *Lancet* **389**, 299–311. https://doi.org/10.1016/S0140-6736(16)30958-8 (2017).
6. Sharma, A. K. Emerging trends in biomarker discovery: Ease of prognosis and prediction in cancer. *Semin. Cancer Biol.* **52**, iii–iv. https://doi.org/10.1016/j.semcancer.2018.05.008 (2018).
7. Wolrab, D. *et al.* Lipidomic profiling of human serum enables detection of pancreatic cancer. *Nat. Commun.* **13**, 124. https://doi.org/10.1038/s41467-021-27765-9 (2022).
8. Specogna, A. V. & Sinicrope, F. A. Defining colon cancer biomarkers by using deep learning. *Lancet* **395**, 314–316. https://doi.org/10.1016/S0140-6736(20)30034-9 (2020).
9. Huang, L. *et al.* Machine learning of serum metabolic patterns encodes early-stage lung adenocarcinoma. *Nat. Commun.* **11**, 3556. https://doi.org/10.1038/s41467-020-17347-6 (2020).
10. Shi, R. *et al.* Identification and validation of hypoxia-derived gene signatures to predict clinical outcomes and therapeutic responses in stage I lung adenocarcinoma patients. *Theranostics* **11**, 5061–5076. https://doi.org/10.7150/thno.56202 (2021).
11. Zare, A., Postovit, L. M. & Githaka, J. M. Robust inflammatory breast cancer gene signature using nonparametric random forest analysis. *Breast Cancer Res.* **23**, 92. https://doi.org/10.1186/s13058-021-01467-y (2021).
12. Zhang, B. *et al.* m(6)A target microRNAs in serum for cancer detection. *Mol. Cancer* **20**, 170. https://doi.org/10.1186/s12943-021-01477-6 (2021).
13. Zhang, N. *et al.* An artificial intelligence network-guided signature for predicting outcome and immunotherapy response in lung adenocarcinoma patients based on 26 machine learning algorithms. *Cell Prolif.* **56**, e13409. https://doi.org/10.1111/cpr.13409 (2023).
14. Zhang, Z. *et al.* Deep learning identifies a T-cell exhaustion-dependent transcriptional signature for predicting clinical outcomes and response to immune checkpoint blockade. *Oncogenesis* **12**, 37. https://doi.org/10.1038/s41389-023-00482-2 (2023).
15. Pei, G., Hu, R., Dai, Y., Zhao, Z. & Jia, P. Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network. *Oncogene* **39**, 5031–5041. https://doi.org/10.1038/s41388-020-1343-z (2020).
16. Wang, J. *et al.* Denoising autoencoder, a deep learning algorithm, aids the identification of a novel molecular signature of lung adenocarcinoma. *Genom. Proteom. Bioinform.* **18**, 468–480. https://doi.org/10.1016/j.gpb.2019.02.003 (2020).
17. Divate, M. *et al.* Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. *Cancers* https://doi.org/10.3390/cancers14051185 (2022).
18. Jha, A. *et al.* Identifying common transcriptome signatures of cancer by interpreting deep learning models. *Genome Biol.* **23**, 117. https://doi.org/10.1186/s13059-022-02681-3 (2022).
19. Zou, J. *et al.* A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18. https://doi.org/10.1038/s41588-018-0295-5 (2019).
20. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. arXiv:1312.6114 (2013). <https://ui.adsabs.harvard.edu/abs/2013arXiv1312.6114K>.
21. Gomari, D. P. *et al.* Variational autoencoders learn transferrable representations of metabolomics data. *Commun. Biol.* **5**, 645. https://doi.org/10.1038/s42003-022-03579-3 (2022).
22. Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S. & Greene, C. S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biol.* **21**, 109. https://doi.org/10.1186/s13059-020-02021-3 (2020).
23. Ding, X., Zou, Z. & Brooks Iii, C. L. Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **10**, 5644. https://doi.org/10.1038/s41467-019-13633-0 (2019).
24. Arslan, E., Schulz, J. & Rai, K. Machine learning in epigenomics: Insights into cancer biology and medicine. *Biochim. Biophys. Acta Rev. Cancer* **1876**, 188588. https://doi.org/10.1016/j.bbcan.2021.188588 (2021).
25. Lorencin, I., Andelic, N., Spanjol, J. & Car, Z. Using multi-layer perceptron with Laplacian edge detector for bladder cancer diagnosis. *Artif. Intell. Med.* **102**, 101746. https://doi.org/10.1016/j.artmed.2019.101746 (2020).
26. Ren, X. *et al.* Machine learning reveals salivary glycopatterns as potential biomarkers for the diagnosis and prognosis of papillary thyroid cancer. *Int. J. Biol. Macromol.* **215**, 280–289. https://doi.org/10.1016/j.ijbiomac.2022.05.194 (2022).
27. Ellis, M. *et al.* Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. *Clin. Cancer Res.* **8**, 1155–1166 (2002).
28. Battista, A., Battista, R. A., Battista, F., Iovane, G. & Landi, R. E. BH-index: A predictive system based on serum biomarkers and ensemble learning for early colorectal cancer diagnosis in mass screening. *Comput. Methods Programs Biomed.* **212**, 106494. https://doi.org/10.1016/j.cmpb.2021.106494 (2021).
29. Hu, X. *et al.* Artificial neural networks and prostate cancer–tools for diagnosis and management. *Nat. Rev. Urol.* **10**, 174–182. https://doi.org/10.1038/nrurol.2013.9 (2013).

30. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. arXiv:1705.07874 (2017). <https://ui.adsabs.harvard.edu/abs/2017arXiv170507874L>.
31. Chakraborty, D. *et al.* Explainable artificial intelligence reveals novel insight into tumor microenvironment conditions linked with better prognosis in patients with breast cancer. *Cancers* https://doi.org/10.3390/cancers13143450 (2021).
32. Rynazal, R. *et al.* Leveraging explainable AI for gut microbiome-based colorectal cancer classification. *Genome Biol.* **24**, 21. https://doi.org/10.1186/s13059-023-02858-4 (2023).
33. Cheng, J. *et al.* Multimodal disentangled variational autoencoder with game theoretic interpretability for glioma grading. *IEEE J. Biomed. Health Inform.* **26**, 673–684. https://doi.org/10.1109/JBHI.2021.3095476 (2022).
34. Ma, M. *et al.* Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms. *Eur. Radiol.* **32**, 1652–1662. https://doi.org/10.1007/s00330-021-08271-4 (2022).
35. Xu, G. *et al.* Decoding river pollution trends and their landscape determinants in an ecologically fragile karst basin using a machine learning model. *Environ. Res.* **214**, 113843. https://doi.org/10.1016/j.envres.2022.113843 (2022).
36. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Res.* **41**, D991-995. https://doi.org/10.1093/nar/gks1193 (2013).
37. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140. https://doi.org/10.1093/bioinformatics/btp616 (2010).
38. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).
39. Freund, Y. & Schapire, R. E. in *icml*. 148–156 (Citeseer).
40. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141. https://doi.org/10.1016/j.xinn.2021.100141 (2021).
41. Rath, S. *et al.* MitoCarta3.0: An updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.* **49**, D1541–D1547. https://doi.org/10.1093/nar/gkaa1011 (2021).
42. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7. https://doi.org/10.1186/1471-2105-14-7 (2013).
43. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592. https://doi.org/10.1093/nar/gkac963 (2023).
44. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550. https://doi.org/10.1073/pnas.0506580102 (2005).
45. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782. https://doi.org/10.1038/s41587-019-0114-2 (2019).
46. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612. https://doi.org/10.1038/ncomms3612 (2013).
47. Li, Z. & Zhang, H. Reprogramming of glucose, fatty acid and amino acid metabolism for cancer progression. *Cell Mol. Life Sci.* **73**, 377–392. https://doi.org/10.1007/s00018-015-2070-4 (2016).
48. Liu, Z. L., Chen, H. H., Zheng, L. L., Sun, L. P. & Shi, L. Angiogenic signaling pathways and anti-angiogenic therapy for cancer. *Signal Transduct. Target. Ther.* **8**, 198. https://doi.org/10.1038/s41392-023-01460-1 (2023).
49. Huang, Y., Hong, W. & Wei, X. The molecular mechanisms and therapeutic strategies of EMT in tumor progression and metastasis. *J. Hematol. Oncol.* **15**, 129. https://doi.org/10.1186/s13045-022-01347-8 (2022).
50. Lytle, N. K., Barber, A. G. & Reya, T. Stem cell fate in cancer growth, progression and therapy resistance. *Nat. Rev. Cancer* **18**, 669–680. https://doi.org/10.1038/s41568-018-0056-x (2018).
51. Qian, B. Z. Inflammation fires up cancer metastasis. *Semin. Cancer Biol.* **47**, 170–176. https://doi.org/10.1016/j.semcancer.2017.08.006 (2017).
52. Qiu, C. *et al.* Identification of molecular subtypes and a prognostic signature based on inflammation-related genes in colon adenocarcinoma. *Front. Immunol.* **12**, 769685. https://doi.org/10.3389/fimmu.2021.769685 (2021).
53. Lapuente-Santana, O., van Genderen, M., Hilbers, P. A. J., Finotello, F. & Eduati, F. Interpretable systems biomarkers predict response to immune-checkpoint inhibitors. *Patterns* **2**, 100293. https://doi.org/10.1016/j.patter.2021.100293 (2021).
54. Ru, B. *et al.* TISIDB: An integrated repository portal for tumor-immune system interactions. *Bioinformatics* **35**, 4200–4202. https://doi.org/10.1093/bioinformatics/btz210 (2019).
55. Hu, F. F., Liu, C. J., Liu, L. L., Zhang, Q. & Guo, A. Y. Expression profile of immune checkpoint genes and their roles in predicting immunotherapy response. *Brief Bioinform.* https://doi.org/10.1093/bib/bbaa176 (2021).
56. Vercellino, I. & Sazanov, L. A. The assembly, regulation and function of the mitochondrial respiratory chain. *Nat. Rev. Mol. Cell Biol.* **23**, 141–161. https://doi.org/10.1038/s41580-021-00415-0 (2022).
57. Yu, T. *et al.* CircRNAs in cancer metabolism: A review. *J. Hematol. Oncol.* **12**, 90. https://doi.org/10.1186/s13045-019-0776-8 (2019).
58. Cruz-Bermudez, A. *et al.* PGC-1alpha levels correlate with survival in patients with stage III NSCLC and may define a new biomarker to metabolism-targeted therapy. *Sci. Rep.* **7**, 16661. https://doi.org/10.1038/s41598-017-17009-6 (2017).
59. Majem, B., Nadal, E. & Munoz-Pinedo, C. Exploiting metabolic vulnerabilities of non small cell lung carcinoma. *Semin. Cell Dev. Biol.* **98**, 54–62. https://doi.org/10.1016/j.semcdb.2019.06.004 (2020).
60. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: Integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570. https://doi.org/10.1126/science.1203486 (2011).
61. Katsuta, E., Rashid, O. M. & Takabe, K. Fibroblasts as a biological marker for curative resection in pancreatic ductal adenocarcinoma. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms21113890 (2020).
62. Romani, C. *et al.* Gene expression profiling of olfactory neuroblastoma helps identify prognostic pathways and define potentially therapeutic targets. *Cancers* https://doi.org/10.3390/cancers13112527 (2021).
63. Zhang, C. *et al.* Comprehensive molecular analyses of a TNF family-based signature with regard to prognosis, immune features, and biomarkers for immunotherapy in lung adenocarcinoma. *EBioMedicine* **59**, 102959. https://doi.org/10.1016/j.ebiom.2020.102959 (2020).
64. Zhang, Z. *et al.* A novel basement membrane-related gene signature for prognosis of lung adenocarcinomas. *Comput. Biol. Med.* **154**, 106597. https://doi.org/10.1016/j.compbiomed.2023.106597 (2023).
65. Zhang, A., Yang, J., Ma, C., Li, F. & Luo, H. Development and validation of a robust ferroptosis-related prognostic signature in lung adenocarcinoma. *Front. Cell Dev. Biol.* **9**, 616271. https://doi.org/10.3389/fcell.2021.616271 (2021).
66. Li, Q. *et al.* Combining autophagy and immune characterizations to predict prognosis and therapeutic response in lung adenocarcinoma. *Front. Immunol.* **13**, 944378. https://doi.org/10.3389/fimmu.2022.944378 (2022).
67. Ouyang, W. *et al.* A prognostic risk score based on hypoxia-, immunity-, and epithelialto-mesenchymal transition-related genes for the prognosis and immunotherapy response of lung adenocarcinoma. *Front. Cell Dev. Biol.* **9**, 758777. https://doi.org/10.3389/fcell.2021.758777 (2021).
68. Qi, C., Ma, J., Sun, J., Wu, X. & Ding, J. The role of molecular subtypes and immune infiltration characteristics based on disulfidptosis-associated genes in lung adenocarcinoma. *Aging (Albany NY)* **15**, 5075–5095. https://doi.org/10.18632/aging.204782 (2023).
69. Martinez-Reyes, I. & Chandel, N. S. Cancer metabolism: Looking forward. *Nat. Rev. Cancer* **21**, 669–680. https://doi.org/10.1038/s41568-021-00378-6 (2021).

70. Counihan, J. L., Grossman, E. A. & Nomura, D. K. Cancer metabolism: Current understanding and therapies. *Chem. Rev.* **118**, 6893–6923. https://doi.org/10.1021/acs.chemrev.7b00775 (2018).
71. Chen, X. *et al.* Monomethyltransferase SET8 facilitates hepatocellular carcinoma growth by enhancing aerobic glycolysis. *Cell Death Dis.* **10**, 312. https://doi.org/10.1038/s41419-019-1541-1 (2019).
72. Altman, B. J., Stine, Z. E. & Dang, C. V. From Krebs to clinic: Glutamine metabolism to cancer therapy. *Nat. Rev. Cancer* **16**, 619–634. https://doi.org/10.1038/nrc.2016.71 (2016).
73. Wang, Z. *et al.* Methionine is a metabolic dependency of tumor-initiating cells. *Nat. Med.* **25**, 825–837. https://doi.org/10.1038/s41591-019-0423-5 (2019).
74. Locasale, J. W. Serine, glycine and one-carbon units: Cancer metabolism in full circle. *Nat. Rev. Cancer* **13**, 572–583. https://doi.org/10.1038/nrc3557 (2013).
75. Morris, S. M. Jr. Recent advances in arginine metabolism: Roles and regulation of the arginases. *Br. J. Pharmacol.* **157**, 922–930. https://doi.org/10.1111/j.1476-5381.2009.00278.x (2009).
76. Xue, C. *et al.* Tryptophan metabolism in health and disease. *Cell Metab.* **35**, 1304–1326. https://doi.org/10.1016/j.cmet.2023.06.004 (2023).
77. Oh, M. H. *et al.* Targeting glutamine metabolism enhances tumor-specific immunity by modulating suppressive myeloid cells. *J. Clin. Invest.* **130**, 3865–3884. https://doi.org/10.1172/JCI131859 (2020).
78. Sanderson, S. M., Gao, X., Dai, Z. & Locasale, J. W. Methionine metabolism in health and cancer: A nexus of diet and precision medicine. *Nat. Rev. Cancer* **19**, 625–637. https://doi.org/10.1038/s41568-019-0187-8 (2019).
79. Bian, Y. *et al.* Cancer SLC43A2 alters T cell methionine metabolism and histone methylation. *Nature* **585**, 277–282. https://doi.org/10.1038/s41586-020-2682-1 (2020).
80. Amelio, I., Cutruzzola, F., Antonov, A., Agostini, M. & Melino, G. Serine and glycine metabolism in cancer. *Trends Biochem. Sci.* **39**, 191–198. https://doi.org/10.1016/j.tibs.2014.02.004 (2014).
81. Butler, L. M. *et al.* Lipids and cancer: Emerging roles in pathogenesis, diagnosis and therapeutic intervention. *Adv. Drug Deliv. Rev.* **159**, 245–293. https://doi.org/10.1016/j.addr.2020.07.013 (2020).
82. Carracedo, A., Cantley, L. C. & Pandolfi, P. P. Cancer metabolism: Fatty acid oxidation in the limelight. *Nat. Rev. Cancer* **13**, 227–232. https://doi.org/10.1038/nrc3483 (2013).
83. Ogretmen, B. Sphingolipid metabolism in cancer signalling and therapy. *Nat. Rev. Cancer* **18**, 33–50. https://doi.org/10.1038/nrc.2017.96 (2018).
84. Cox, G., Jones, J. L., Walker, R. A., Steward, W. P. & O'Byrne, K. J. Angiogenesis and non-small cell lung cancer. *Lung Cancer* **27**, 81–100. https://doi.org/10.1016/s0169-5002(99)00096-3 (2000).
85. Herbst, R. S., Onn, A. & Sandler, A. Angiogenesis and lung cancer: Prognostic and therapeutic implications. *J. Clin. Oncol.* **23**, 3243–3256. https://doi.org/10.1200/JCO.2005.18.853 (2005).
86. Xiao, Y. & Yu, D. Tumor microenvironment as a therapeutic target in cancer. *Pharmacol. Ther.* **221**, 107753. https://doi.org/10.1016/j.pharmthera.2020.107753 (2021).
87. Hu, C. *et al.* Comprehensive profiling of immune-related genes in soft tissue sarcoma patients. *J. Transl. Med.* **18**, 337. https://doi.org/10.1186/s12967-020-02512-8 (2020).
88. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264. https://doi.org/10.1038/nrc3239 (2012).
89. Musaelyan, A. A. *et al.* Inflammatory and autoimmune predictive markers of response to anti-PD-1/PD-L1 therapy in NSCLC and melanoma. *Exp. Ther. Med.* **24**, 557. https://doi.org/10.3892/etm.2022.11495 (2022).
90. Lester, D. K. *et al.* Fucosylation of HLA-DRB1 regulates CD4(+) T cell-mediated anti-melanoma immunity and enhances immunotherapy efficacy. *Nat. Cancer* **4**, 222–239. https://doi.org/10.1038/s43018-022-00506-7 (2023).
91. Morad, G., Helmink, B. A., Sharma, P. & Wargo, J. A. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell* **184**, 5309–5337. https://doi.org/10.1016/j.cell.2021.09.020 (2021).
92. Zhou, F., Qiao, M. & Zhou, C. The cutting-edge progress of immune-checkpoint blockade in lung cancer. *Cell Mol. Immunol.* **18**, 279–293. https://doi.org/10.1038/s41423-020-00577-5 (2021).
93. Geeleher, P., Cox, N. & Huang, R. S. pRRophetic: An R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One* **9**, e107468. https://doi.org/10.1371/journal.pone.0107468 (2014).
94. Ma, L. *et al.* The essential roles of m(6)A RNA modification to stimulate ENO1-dependent glycolysis and tumorigenesis in lung adenocarcinoma. *J. Exp. Clin. Cancer Res.* **41**, 36. https://doi.org/10.1186/s13046-021-02200-5 (2022).
95. Xia, L. *et al.* The cancer metabolic reprogramming and immune response. *Mol. Cancer* **20**, 28. https://doi.org/10.1186/s12943-021-01316-8 (2021).
96. Wang, W. & Zou, W. Amino acids and their transporters in T cell immunity and cancer therapy. *Mol. Cell* **80**, 384–395. https://doi.org/10.1016/j.molcel.2020.09.006 (2020).
97. Reinfeld, B. I., Rathmell, W. K., Kim, T. K. & Rathmell, J. C. The therapeutic implications of immunosuppressive tumor aerobic glycolysis. *Cell Mol. Immunol.* **19**, 46–58. https://doi.org/10.1038/s41423-021-00727-3 (2022).
98. Huang, M. *et al.* Targeting glutamine metabolism to enhance immunoprevention of EGFR-driven lung cancer. *Adv. Sci.* **9**, e2105885. https://doi.org/10.1002/advs.202105885 (2022).
99. Kawalekar, O. U. *et al.* Distinct signaling of coreceptors regulates specific metabolism pathways and impacts memory development in CAR T cells. *Immunity* **44**, 380–390. https://doi.org/10.1016/j.immuni.2016.01.021 (2016).
100. Giatromanolaki, A. *et al.* Prognostic value of angiogenesis in operable non-small cell lung cancer. *J. Pathol.* **179**, 80–88. https://doi.org/10.1002/(SICI)1096-9896(199605)179:1%3c80::AID-PATH547%3e3.0.CO;2-X (1996).
101. Cai, S. *et al.* Integrative analysis and experiments to explore angiogenesis regulators correlated with poor prognosis, immune infiltration and cancer progression in lung adenocarcinoma. *J. Transl. Med.* **19**, 361. https://doi.org/10.1186/s12967-021-03031-w (2021).
102. Suarez-Carmona, M., Lesage, J., Cataldo, D. & Gilles, C. EMT and inflammation: Inseparable actors of cancer progression. *Mol. Oncol.* **11**, 805–823. https://doi.org/10.1002/1878-0261.12095 (2017).
103. Xiao, Z., Cai, Z., Deng, D., Tong, S. & Zu, X. An EMT-based risk score thoroughly predicts the clinical prognosis, tumor immune microenvironment and molecular subtypes of bladder cancer. *Front. Immunol.* **13**, 1000321. https://doi.org/10.3389/fimmu.2022.1000321 (2022).
104. Tian, L. *et al.* Mutual regulation of tumour vessel normalization and immunostimulatory reprogramming. *Nature* **544**, 250–254. https://doi.org/10.1038/nature21724 (2017).
105. Zhang, D. *et al.* Metabolic regulation of gene expression by histone lactylation. *Nature* **574**, 575–580. https://doi.org/10.1038/s41586-019-1678-1 (2019).
106. Blomberg, O. S. *et al.* IL-5-producing CD4(+) T cells and eosinophils cooperate to enhance response to immune checkpoint blockade in breast cancer. *Cancer Cell* **41**, 106–123. https://doi.org/10.1016/j.ccell.2022.11.014 (2023).
107. Nelder, J. A. & Wedderburn, R. W. Generalized linear models. *J. R. Stat. Soc. Ser. A Stat. Soc.* **135**, 370–384 (1972).
108. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
109. Park, M. Y. & Hastie, T. Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50. https://doi.org/10.1093/biostatistics/kxm010 (2008).

110. Vinayak, R. K. & Gilad-Bachrach, R. in *Artificial Intelligence and Statistics*. 489–497 (PMLR).

Author contributions

L.C. and J.L. contributed conceptualization, project administration, writing—review & editing. S.Q. contributed data curation, formal analysis, investigation, methodology, validation, writing—original draft, writing—review & editing. S.S. helped in data curation, formal analysis, investigation, methodology. Y.W. helped in data curation, investigation, formal analysis. C.L., L.F., M.W., J.Y. and W.L. carried out data curation, validation. All authors reviewed the manuscript.

Funding

This study was funded by the National Natural Science Foundation of China (61702141; 81627901), the Natural Science Foundation of Heilongjiang Province (LH2021F043) and the Heilongjiang Postdoctoral Funds for Scientific Research Initiation (LBH-Q17132).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-51108-x.

Correspondence and requests for materials should be addressed to J.L. or L.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.