Zhang *et al. Biology Direct* (2025) 20:58
https://doi.org/10.1186/s13062-025-00649-4

Biology Direct

RESEARCH Open Access

Identification of M2 macrophage-related genes associated with diffuse large B-cell lymphoma via bioinformatics and machine learning approaches

Jiayi Zhang[1], Zhixiang Jia[1], Jiahui Zhang[2], Xiaohui Mu[1] and Limei Ai[1*]

Abstract

M2 macrophages play a crucial role in the initiation and progression of various tumors, including diffuse large B-cell lymphoma (DLBCL). However, the characterization of M2 macrophage-related genes in DLBCL remains incomplete. In this study, we downloaded DLBCL-related datasets from the Gene Expression Omnibus (GEO) database and identified 77 differentially expressed genes (DEGs) between the control group and the treat group. We assessed the immune cell infiltration using CIBERSORT analysis and identified modules associated with M2 macrophages through weighted gene co-expression network analysis (WGCNA). Using the Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine Recursive Feature Elimination (SVM-RFE), and Random Forest (RF) algorithms, we screened for seven potential diagnostic biomarkers with strong diagnostic capabilities: SMAD3, IL7R, IL18, FAS, CD5, CCR7, and CSF1R. Subsequently, the constructed logistic regression model and nomogram demonstrated robust predictive performance. We further investigated the expression levels, prognostic values, and biological functions of these biomarkers. The results showed that SMAD3, IL7R, IL18, FAS and CD5 were associated with the survival of DLBCL patients and could be used as markers to predict the prognosis of DLBCL. Our study introduces a novel diagnostic strategy and provides new insights into the potential mechanisms underlying DLBCL. However, further validation of the practical value of these genes in DLBCL diagnosis is warranted before clinical application.

Keywords Bioinformatics, Diffuse large B-cell lymphoma, M2 macrophages, Immune infiltration

*Correspondence:
Limei Ai
alm121001@163.com
[1]Department of Hematology, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, China
[2]Medical College, Sanmenxia Vocational and Technical College, Sanmenxia, China

Introduction

Diffuse Large B-cell Lymphoma (DLBCL) is the most common lymphoma in adults, accounting for 30–40% of non-Hodgkin lymphomas(NHL) [1, 2]. Approximately 50–70% of patients achieve long-term remission following combined immunochemotherapy with rituximab, cyclophosphamide, doxorubicin, and prednisone (R-CHOP) [3–5]. However, a subset of patients relapse or fail to respond to the R-CHOP protocol [6, 7]. It is known that lymphoid malignancies possess clinically exploitable immune sensitivity, and their intrinsic tumor microenvironment (TME) renders them natural targets for immunotherapy [8]. In addition to novel antibodies targeting surface antigens and small molecule inhibitors aimed at oncogenic signaling pathways and tumor suppressor factors, immune checkpoint inhibitors and chimeric antigen receptor T (CAR-T) cells have also rapidly emerged as strategies targeting the tumor microenvironment [9].

The TME is a key regulatory factor in tumor growth, progression, and metastasis. Among the innate immune cells recruited to the tumor site, macrophages are the most abundant cell type and are present at various stages of tumor progression [10]. Tumor-associated macrophages (TAMs) are an important component of the TME [11] and play a pivotal role in establishing an immunosuppressive environment that promotes tumor growth and metastasis [12]. TAM can mediate immune suppression, shape and remodel the tumor immune microenvironment (TIME), and contribute to tumor immune evasion [13–15]. TAM consists of heterogeneous subpopulations, including M1 anti-tumor phenotypes and M2 pro-tumor phenotypes [16]. However, the role of M2 macrophages in the development of DLBCL remains unknown. Therefore, investigating M2 macrophage-related genes closely associated with DLBCL is of great significance for understanding its pathogenesis, enhancing diagnosis and treatment, and conducting prognostic assessments.

This study employs bioinformatics and machine learning approaches to identify M2 macrophage-related genes associated with DLBCL. These findings will aid in further exploring the role of the tumor microenvironment in DLBCL and the significance of M2 macrophages in the onset, progression, and prognosis of the disease. This work provides critical insights into the molecular mechanisms, diagnosis, treatment, and prognostic evaluation of DLBCL.

Materials and methods

Data sources

In the GEO database, we employed " DLBCL" as a keyword to filtrate gene expression profile data associated with DLBCL. In this study, the experimental group consisted of pathological tissues from patients diagnosed with DLBCL, while the control group comprised hyperplastic lymph node tissues. The datasets downloaded included GSE9327, GSE23647, GSE32018, and GSE83632. Additionally, we retrieved comprehensive clinical information for the patients from the GEO dataset GSE181063, which included details such as age, gender, disease stage, subtype, ECOG performance status, IPI score, B symptoms, LDH levels, number of extranodal sites, and survival data.

Data preprocessing and differential gene expression analysis

The platform annotation file was utilized to convert the probe expression matrix into a gene expression matrix. The "sva" packag (V.3.54.0) was employed to eliminate heterogeneity in the training dataset caused by variations in experimental platforms and batches. We assessed the effectiveness of the correction among samples through two-dimensional principal component analysis (PCA) clustering. Ultimately, we obtained a merged normalized gene expression matrix and conducted further analyses using the "limma" package (V.3.62.2). The thresholds for identifying differentially expressed genes (DEGs) were set at adj.p.Val < 0.05 and |log FC| > 0.585. Heatmaps and volcano plots for DEGs were generated using the "pheatmap"(V.1.0.12) and "ggplot2"(V.3.5.1) packages, respectively, to illustrate the patterns of differential expression.

Immune infiltration analysis

CIBERSORT is a method used to characterize the cellular composition of complex tissues based on gene expression profiles [17]. In this study, we utilized CIBERSORT software to predict the proportions of 22 infiltrating immune cell types in each tissue from the merged dataset (see Table S1). For each sample, the sum of scores for all evaluated immune cell types equaled 1 [18]. Subsequently, the results from CIBERSORT were visualized using the R packages "reshape2(V.1.4.4)," "ggpubr(V.0.6.0)," "ggplot2(V.3.5.1)," and "dplyr(V.1.1.4)."

Weighted gene co-expression network analysis (WGCNA)

WGCNA uses the correlation coefficients of normalized expression levels for each gene to assess the co-expression relationships among genes, defining genes with co-expression relationships as a module. Genes within the same module exhibit similar expression levels, while genes in different modules show significant differences in expression levels. This approach allows for the simplification of complex high-throughput data into manageable modules for dimensionality reduction analysis. Ultimately, we can explore the relationships between these gene co-expression modules and immune cells, revealing the biological significance of these modules.

We conducted the analysis using the WGCNA package (V.1.73) [19], setting the minimum module size to 50, the soft threshold to an optimal value of 10, the module merging cutoff height to 0.2, and the minimum distance to 0.2. This method was employed to derive co-expression modules containing DEGs associated with M2 macrophages in the DLBCL group. Functional enrichment analysis was performed using the clusterProfiler package (V.4.14.4) in R, with a filtering criterion of $p < 0.05$ for Gene Ontology (GO)and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of M2 macrophage-related DEGs (see Table S2). Finally, the results were visualized using the "ggplot2" package (V.3.5.1) in R.

Construction of protein-protein interaction network analysis

To analyze the interactions between M2 macrophages genes and associated proteins, we constructed a protein-protein interaction (PPI) network by utilizing the STRING online database to intersect the M2 macrophages module with DEG. To further investigate the interactions between M2 macrophages and DEGs, we employed the cytoHubba plugin in Cytoscape software to identify closely connected gene hubs [20]. For in-depth analysis, we selected the top 10 node genes based on their scoring rankings.

Identification of M2 macrophages biomarkers

We employed the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm in conjunction with the glmnet package (V.4.1.8) to perform dimensionality reduction [21], preserving the M2 macrophage-related DEGs that differentiate DLBCL patients from normal samples for feature selection. Subsequently, we established a Support Vector Machine Recursive Feature Elimination (SVM-RFE) model using the e1071 package (V.4.3.1) and compared the model's performance through the average misclassification rate obtained via 10-fold cross-validation [22]. Random Forest (RF), a recursive partitioning algorithm based on binary trees, was utilized for the selection of hub genes using the "randomForest" package (V.4.7.1.2), which ranked the DEGs and selected genes with importance scores greater than 10 for downstream analyses. Furthermore, overlapping biomarkers identified by the three algorithms were used to determine the optimal gene biomarkers for DLBCL.

Subsequently, we annotated the positions of the identified biomarker genes on human chromosomes using the "circlize" package (V.0.4.16). To assess the diagnostic ability of the optimal gene biomarkers, we calculated the Receiver Operating Characteristic (ROC) curve and measured the area under the curve (AUC), as well as accuracy, sensitivity, and specificity. Building on this, we constructed a logistic regression model based on the seven biomarker genes using the glm package in R, employing the prediction function to forecast sample types within the merged dataset, while also utilizing the ROC curve to evaluate the diagnostic performance of the logistic regression model. Additionally, we established a nomogram using the "rms" R package to predict the risk of DLBCL based on the feature genes and estimated the predictive efficacy of the nomogram through calibration curves.

Survival analysis and independent prognostic analysis

We utilized the GSE181063 dataset to analyze the prognostic value of the identified biomarker genes in DLBCL. First, we determined the threshold for each biomarker gene based on the median expression values. Subsequently, survival analysis was performed using the "survival"(V.3.8.3) and "survminer" (V.0.5.0) packages. Following this, we conducted Cox regression analysis to evaluate the potential of these biomarker genes as independent prognostic factors for DLBCL patients.

Gene set enrichment analysis (GSEA)

This analysis was conducted using the GSEA (V.4.4.1) package in R. To further explore the differentially associated pathways between the high and low infiltration groups of M2 macrophages, we calculated the correlation of M2 macrophages with all other genes in the integrated dataset. Subsequently, we ranked all genes from high to low based on their correlation and treated these ranked genes as the gene set for testing. Meanwhile, we utilized the KEGG pathway collection as a predefined set to assess the degree of enrichment within the gene set. The specific enrichment results are summarized in Table S3.

Gene set variation analysis (GSVA)

This analysis was conducted using the GSVA (V.4.4.1) package in R. GSVA is a method for analyzing gene set variation [23]. In this study, we used the KEGG pathway collection as the background gene set for GSVA analysis. Additionally, we employed the limma package to analyze the differences in GSVA scores between samples with high and low infiltration of M2 macrophages. The filtering criteria were set as follows: $t > 2$ and $p < 0.05$. If $t > 0$, we considered that the pathway was activated in the high infiltration group; conversely, if $t < 0$, we considered that the pathway was activated in the low infiltration group.

Results

Identification of DEGs in DLBCL

A total of 278 samples (158 control samples and 120 experimental samples) were used to identify DEGs in DLBCL. The batch effects of four GEO datasets (GSE9327, GSE23647, GSE32018, and GSE83632) were

addressed using the "sva" package, and normalization was performed using the "limma" package. Principal component analysis (PCA) scatter plots demonstrated the changes in normalized gene expression data before and after batch effect removal (Fig. 1A and B). Furthermore, the results of the differential expression analysis revealed a total of 77 DEGs identified in the combined dataset, including 42 upregulated genes and 35 downregulated genes (Fig. 1C and D).

Immune infiltration analysis

A substantial body of evidence indicates an inseparable link between the immune microenvironment and DLBCL [24–26]. Therefore, we applied the CIBERSORT algorithm to explore the differences in immune microenvironment between the control and experimental groups. Based on the results of the immune infiltration analysis, a bar chart was generated (Fig. 2A) that displays the infiltration status of 22 immune cell types in each cancer patient sample. Additionally, we analyzed the differential expression of immune cells between the two groups, as shown in Fig. 2B. In the experimental group, the proportions of memory B lymphocytes, naïve CD4 + T lymphocytes, follicular helper T lymphocytes, γδ T lymphocytes, activated natural killer cells, M1 macrophages, M2 macrophages, and neutrophils were higher than those in the control group. Conversely, the proportions of

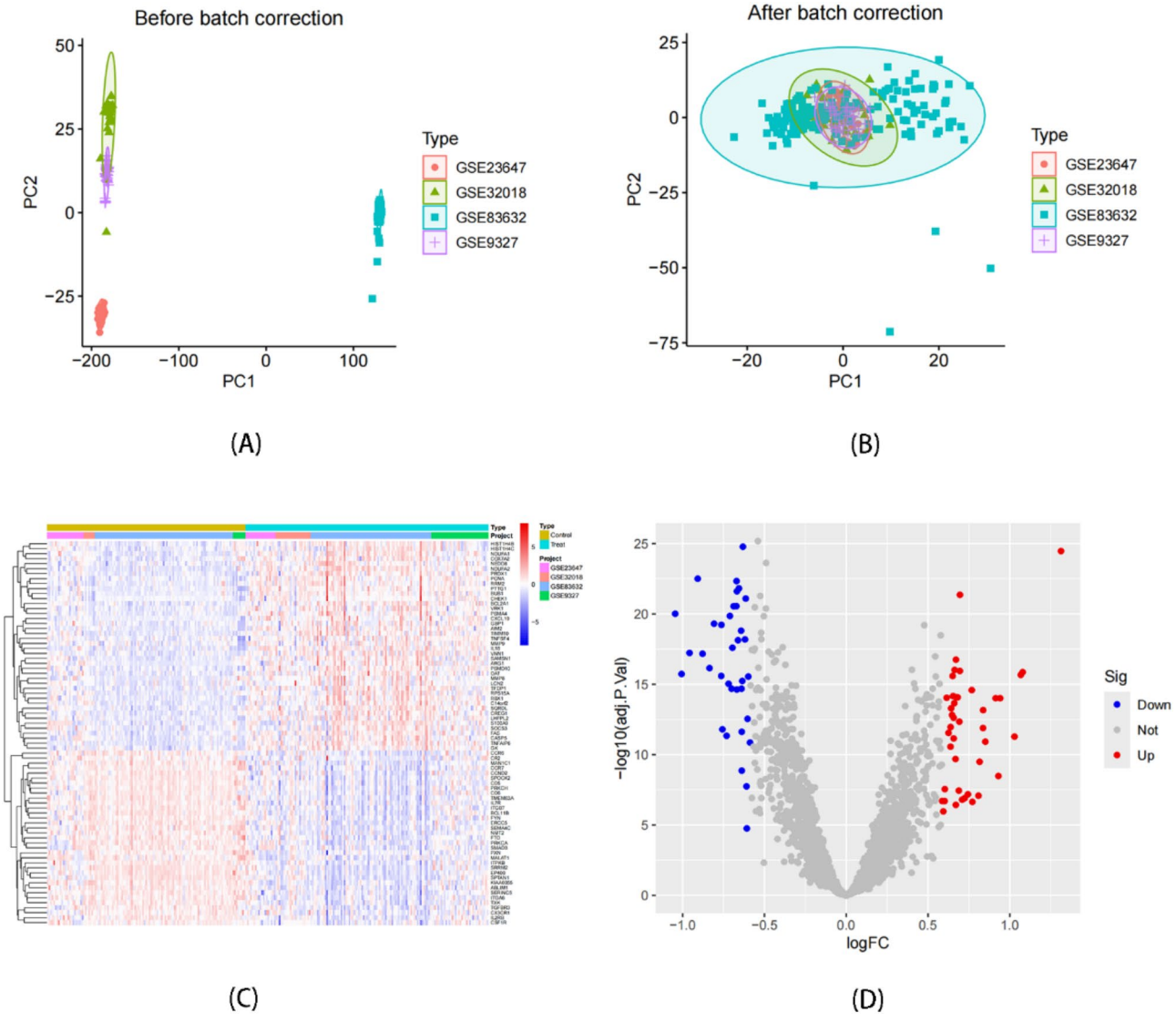

Fig. 1 Principal component analysis (PCA) showing patterns of gene expression across datasets and differential gene expression analysis. (**A**) The distribution of the four datasets before batch effect was removed. (**B**) Removed all confounding factors from the corrected samples. (**C**) A heatmap illustrating the expression patterns of DEGs across the samples. (**D**) Volcanic plots for differentially expressed genes. Red and blue dots denote significantly upregulated and downregulated genes, respectively, while black dots indicate non-significant genes

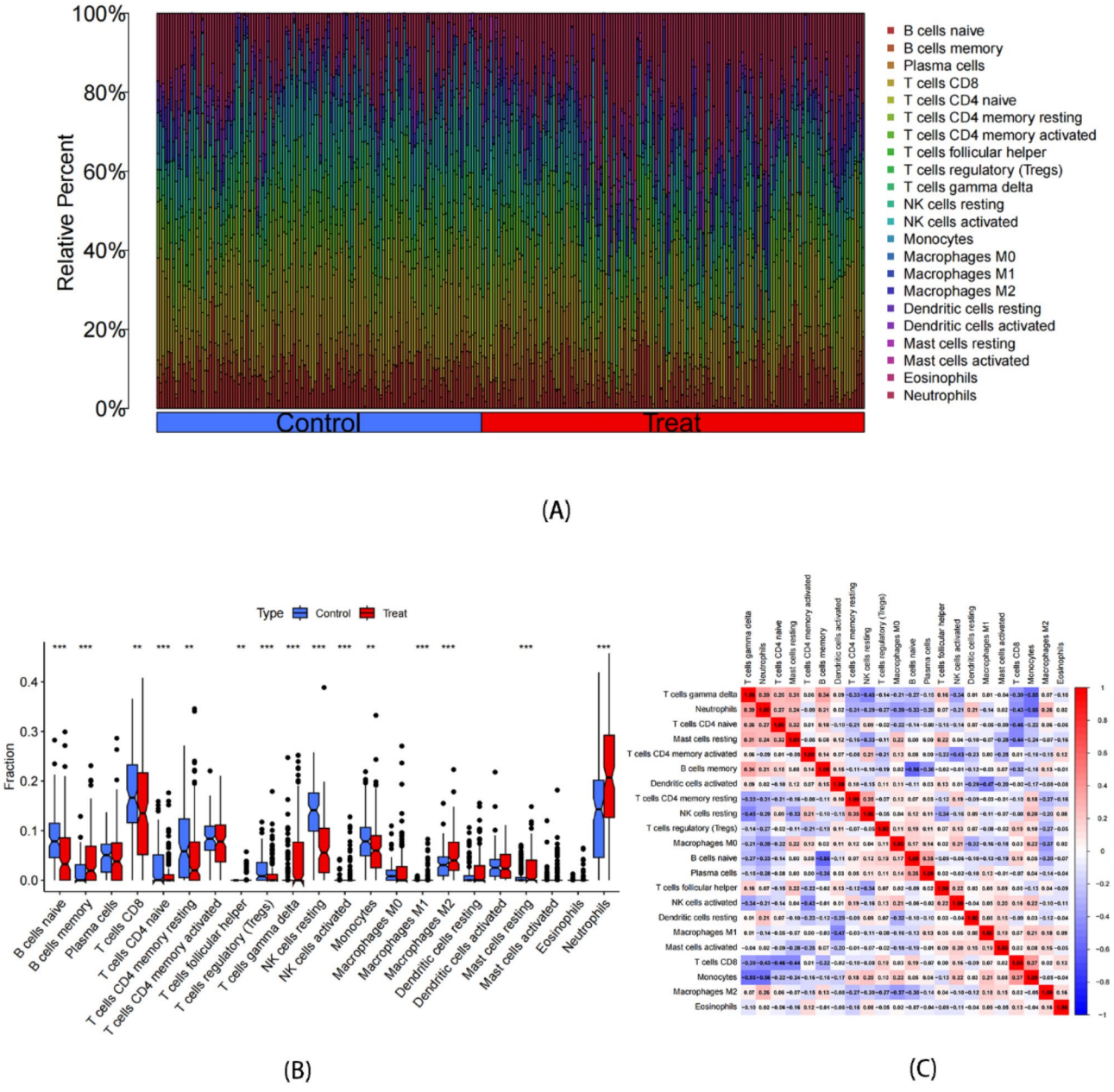

Fig. 2 Immune Infiltration Analysis. (**A**) The bar chart displays immune cell infiltration results of 22 immune cells in two groups. (**B**) The group comparison chart illustrates differences in the abundance of immune cell infiltration in two groups. (**C**) The correlation matrix of immune cell proportions

naïve B lymphocytes, CD8 + T lymphocytes, resting CD4 + memory T lymphocytes, regulatory T lymphocytes, resting natural killer cells, monocytes, and resting mast cells were lower than those in the control group. Furthermore, a more detailed analysis of immune cell infiltration revealed complex correlations among the cells (Fig. 2C). For instance, the correlation between monocytes and γδ T lymphocytes was 0.55, the correlation between monocytes and neutrophils was 0.56, and the correlation between naïve B lymphocytes and memory B lymphocytes was 0.56.

Identification of 60 potential genes associated with M2 macrophage infiltration

We employed the WGCNA method to identify modules associated with M2 macrophages in DLBCL. When the soft threshold β was set to 8, the scale-free topology fitting index R^2 reached 0.9. Subsequently, we used the "dynamic merging" algorithm to obtain three modules (Fig. 3A and B). The analysis results indicated a strong correlation between the turquoise module and M2 macrophages (cor = 0.38; $P < 0.001$; Fig. 3C). Additionally, the characteristic genes of M2 macrophages showed a

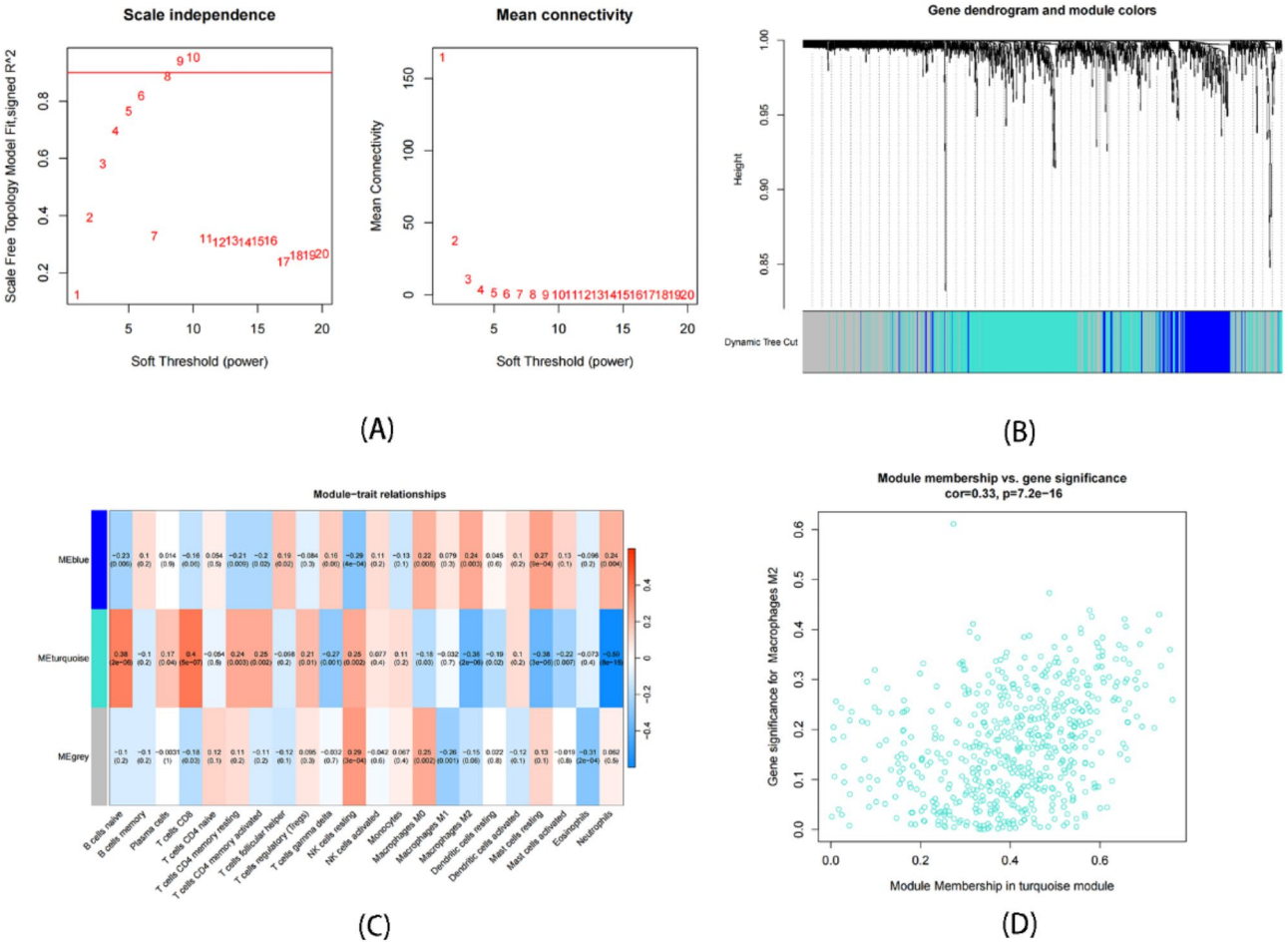

Fig. 3 Identification of related modules. (**A**) Scale Independence and average connectivity in integrated dataset. (**B**) Cluster dendrogram in integrated dataset. (**C**) Heatmap of correlation between modules and important immune cells in integrated dataset. (**D**)Scatter plot showing the relationship between the associated genes of M2 macrophages and the module members of MEturquoise

significant correlation with the modular genes in the turquoise module (cor = 0.33; $P < 0.001$; Fig. 3D). Therefore, for downstream analysis, we selected the turquoise module from the integrated dataset as the key module related to M2 macrophages. Next, we overlapped the differentially expressed genes (DEGs) with the genes of the turquoise module, ultimately identifying 60 potential genes (Fig. 4A).

Functional analysis of potential genes and identification of core network genes

To elucidate the biological functions and pathways associated with the potential genes, we conducted GO and KEGG enrichment analyses. The results of the GO functional enrichment analysis are presented in Fig. 4B. In the biological process category, the potential genes were associated with the positive regulation of leukocyte cell-cell adhesion, cell adhesion, inflammatory response modulation, and T cell activation. In terms of cellular components, the changes were mainly related to the exterior side of the plasma membrane, tertiary granule lumen, specific granule lumen, and replication fork. Regarding molecular functions, the potential genes exhibited activities related to cytokine binding, immune receptor activity, cytokine receptor activity, and G protein-coupled chemoreceptor activity. The results of the KEGG pathway enrichment analysis are shown in Fig. 4C, indicating that the potential genes are linked to transcriptional dysregulation in cancer, hematopoietic cell lineage, p53 signaling pathway, and the cell cycle. Notably, the potential genes were also significantly enriched in various immune-related features. This evidence suggests that the potential genes may play a crucial role in the pathogenesis of DLBCL by participating in cell adhesion, modulating immune cells, and influencing various enzymatic activities. Finally, we performed PPI analysis of the 60 potential genes using the STRING database and Cytoscape software (Fig. 4D). Ten hub genes were identified using the cytoHubba plugin in Cytoscape, as shown in Fig. 4E.

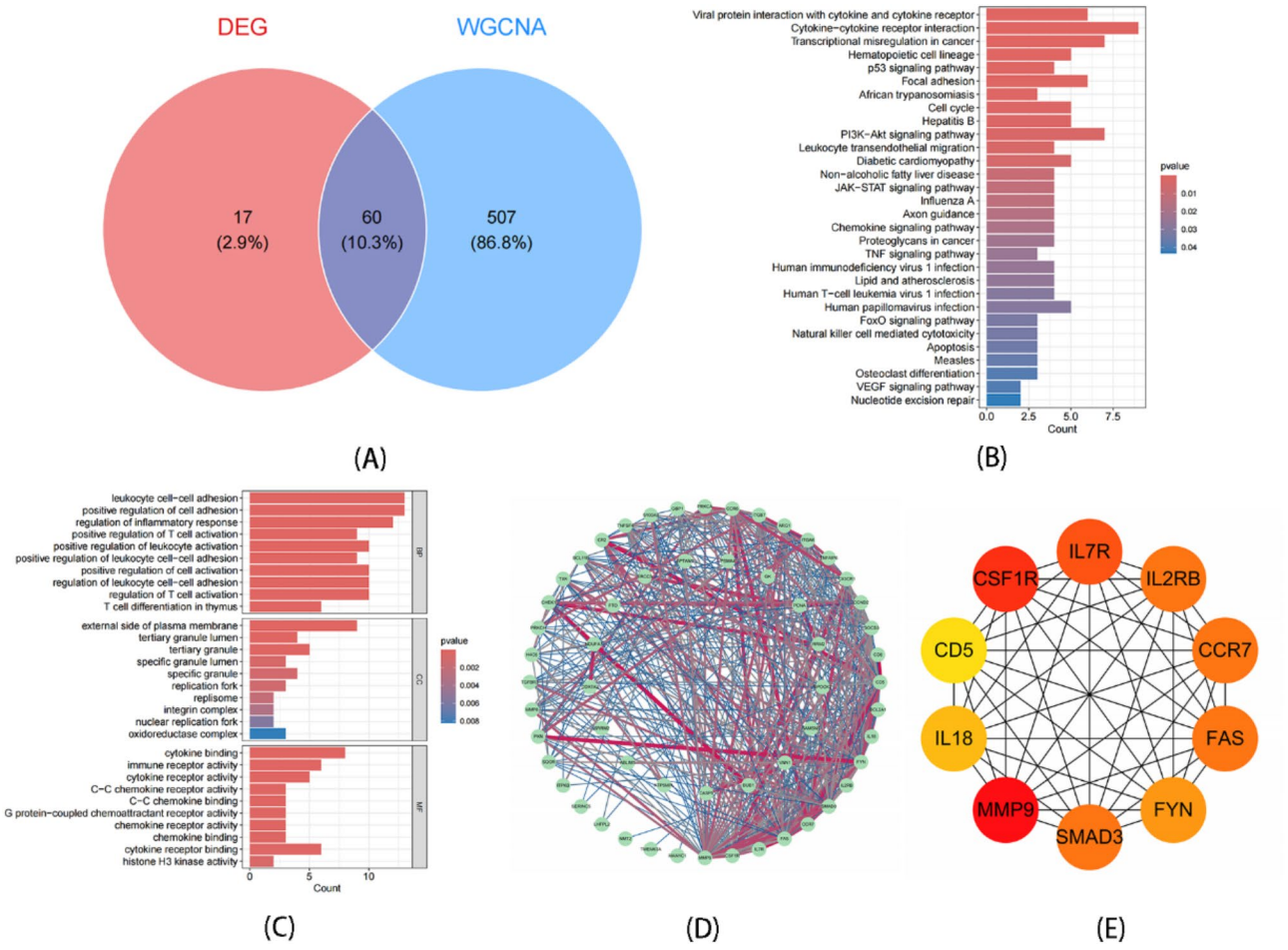

Fig. 4 Identification of ten hub genes. (**A**) Wayne diagram showing the 60 potential genes shared by DEGs and MEturquoise modules. (**B**) Barplot chart of GO analyses of potential genes. (**C**) Barplot chart of KEGG analyses of potential genes. (**D**) Cytoscape visualization showing the network diagram of protein-protein interactions. (**E**) Network diagram of hub gene junctions generated by cytoHubba plugin

Seven differentially expressed genes identified as diagnostic genes for DLBCL

Considering the differences between DLBCL patients and healthy individuals, we aimed to evaluate the diagnostic potential of the DEGs. Subsequently, we implemented three machine learning algorithms—LASSO, SVW-RFE, and RF—to select significant DEGs that could distinguish DLBCL patients from healthy individuals within the integrated dataset. Using the LASSO logistic regression algorithm with 10-fold cross-validation, we identified 9 feature genes associated with DLBCL (Fig. 5A and B). Next, the SVW-RFE algorithm identified 10 genes (maximum accuracy = 0.871, minimum RMSE = 0.129) as the optimal feature genes (Fig. 5C and D). Additionally, the RF algorithm determined 8 genes as the best feature genes (Fig. 5E and F). Finally, we performed a cross-analysis of the marker genes obtained from the three machine learning algorithms, identifying 7 signature genes—SMAD3, IL7R, IL18, FAS, CD5, CCR7, and CSF1R—for further analysis (Fig. 6A).

Subsequently, we annotated the locations of the 7 signature genes on the human chromosomes and visualized them using a pie chart (Fig. 6B). The results showed that IL7R and CSF1R are located on chromosome 5, while IL18 and CD5 are located on chromosome 11. We further explored the expression levels of these 7 signature genes between the control and experimental groups, as depicted in Fig. 6C. Compared to the control group, the expression levels of SMAD3, IL7R, CD5, CCR7, and CSF1R were reduced in the experimental group, whereas the expression levels of IL18 and FAS were increased.

Based on the aforementioned 7 signature genes, we constructed a logistic regression model using the R package glm. Subsequent ROC curve analysis demonstrated that this logistic regression model effectively distinguished DLBCL patients from healthy individuals, with an AUC of 0.921 (Fig. 7A). Furthermore, to evaluate the ability of each individual gene to differentiate between DLBCL and normal samples, we generated ROC curves for each of the 7 signature genes. As shown in Fig. 7B, all

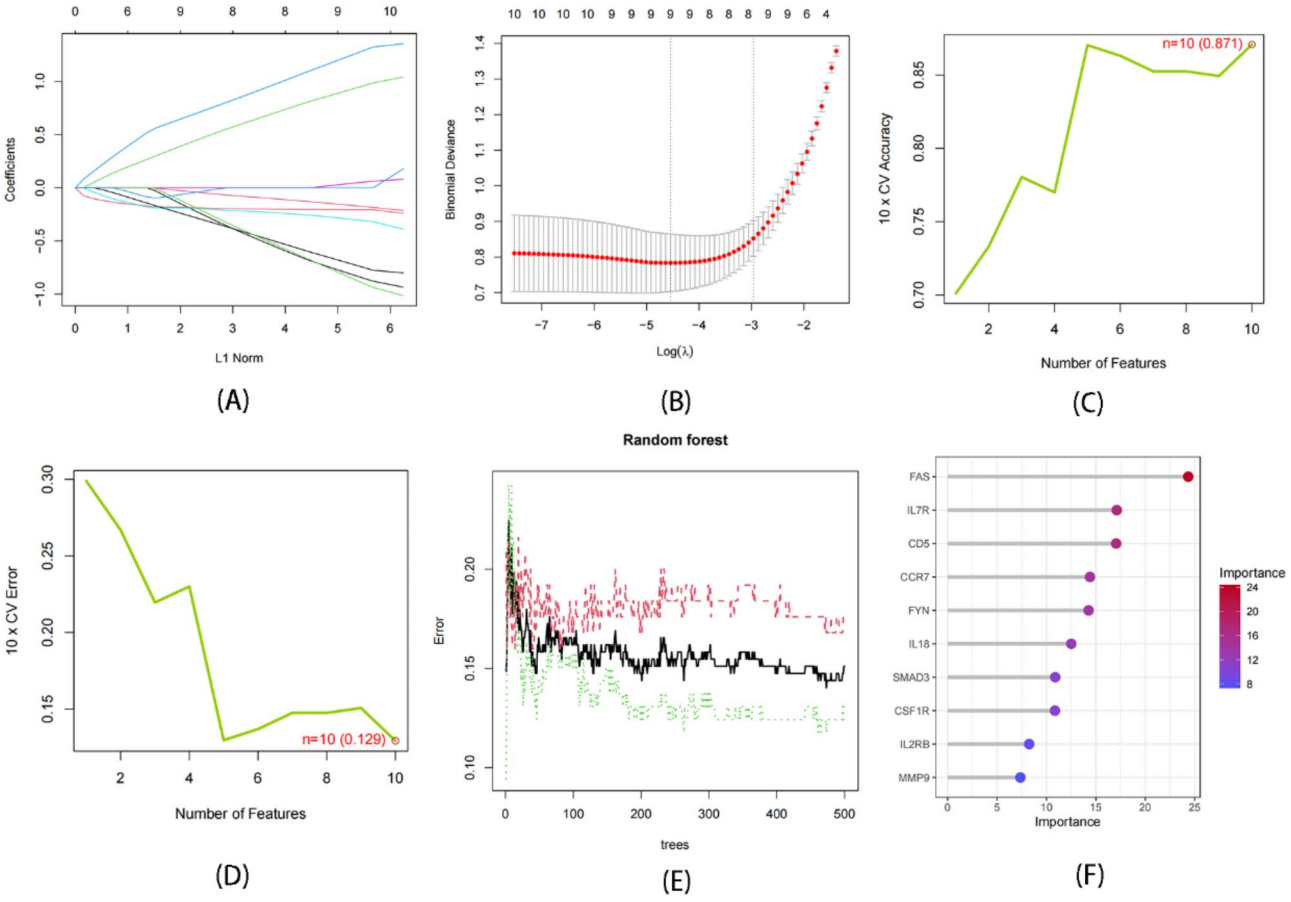

Fig. 5 Identification of diagnostic genes. (**A** and **B**) The LASSO logistic regression algorithm was utilized, with penalty parameter tuning performed through 10-fold cross-validation, leading to the selection of 9 genes associated with DLBCL characteristics. (**C** and **D**) The SVW-RFE algorithm was applied to determine the optimal combination of feature genes and ultimately identifying 10 genes (maximum accuracy = 0.871, minimum RMSE = 0.129) as the optimal feature set. (**E** and **F**) The RF algorithm determined 8 genes as the best feature genes

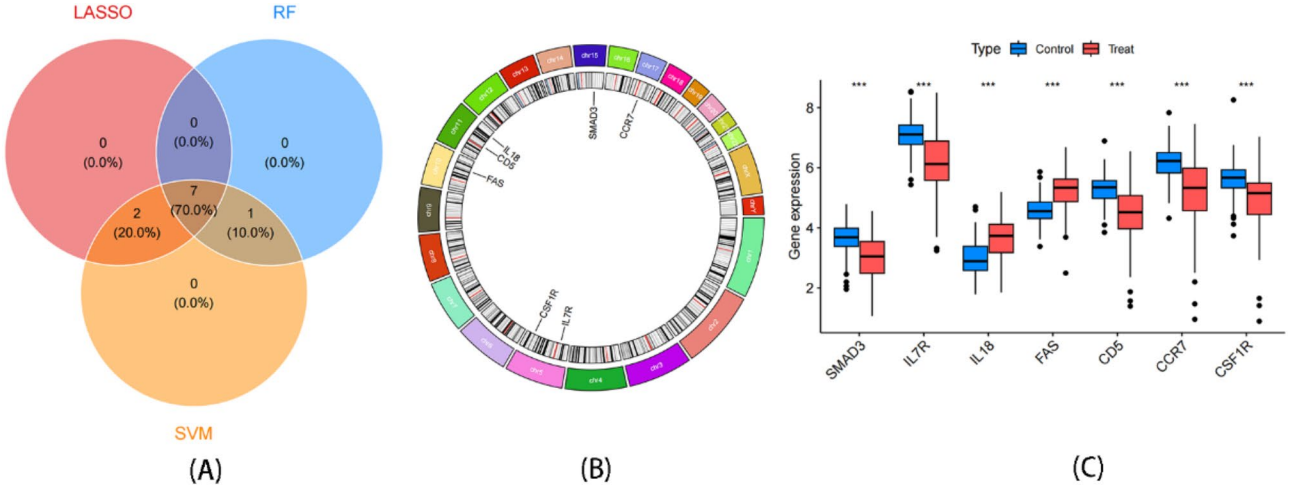

Fig. 6 Expression of the 7 signature genes in DLBCL dataset. (**A**) Wayne diagram showing the 7 signature genes shared by LASSO、SVW-RFE and RF. (**B**) Chromosome location map of the 7 signature genes. (**C**) The expression levels of the 7 signature genes in control and treat samples

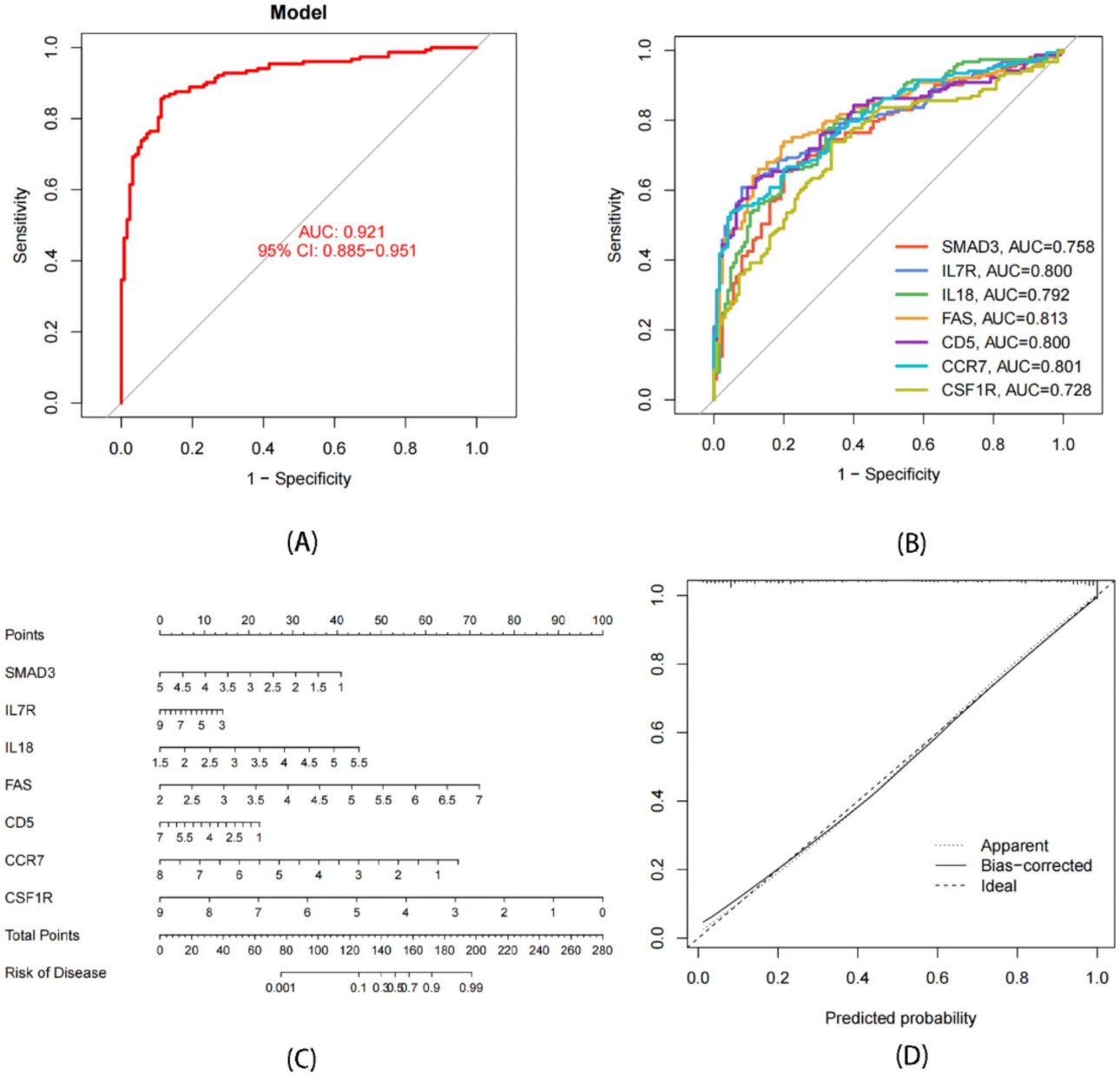

Fig. 7 Logistic regression model and nomogram model of DLBCL patients were constructed based on 7 signature genes. (**A**) The AUC of the logistic regression model for identifying DLBCL samples is shown. (**B**) The ROC curves for the 7 signature genes are displayed. (**C**) A nomogram model combined with based on 7 signature genes was constructed to predict the risk of DLBCL patients. (**D**) The calibration curve of the nomogram tests the predictive performance of the model

genes exhibited AUC values greater than 0.7, indicating good discriminatory power. This evidence suggests that the logistic regression model provides higher accuracy and specificity in differentiating DLBCL samples from normal samples compared to individual signature genes. Subsequently, we developed a nomogram model based on these 7 signature genes to predict the risk of disease in DLBCL patients, as illustrated in Fig. 7C. Additionally, the calibration curve of the nomogram further confirmed the good predictive performance of our model (Fig. 7D).

Survival and prognostic analysis of seven biomarker genes

We analyzed the prognostic value of the signature genes in DLBCL using the GSE181063 dataset. As shown in Fig. 8, patients in the low-expression group had significantly shorter overall survival compared to those in the high-expression group for CD5 ($p < 0.001$; Fig. 8A), FAS ($p = 0.004$; Fig. 8B), IL7R ($p < 0.001$; Fig. 8C), IL18 ($p < 0.001$; Fig. 8D), and SMAD3 ($p = 0.004$; Fig. 8E). In contrast, the expression of CCR7 ($p = 0.882$, Fig.S1A) and CSF1R ($p = 0.056$, Fig.S1B) did not show a significant

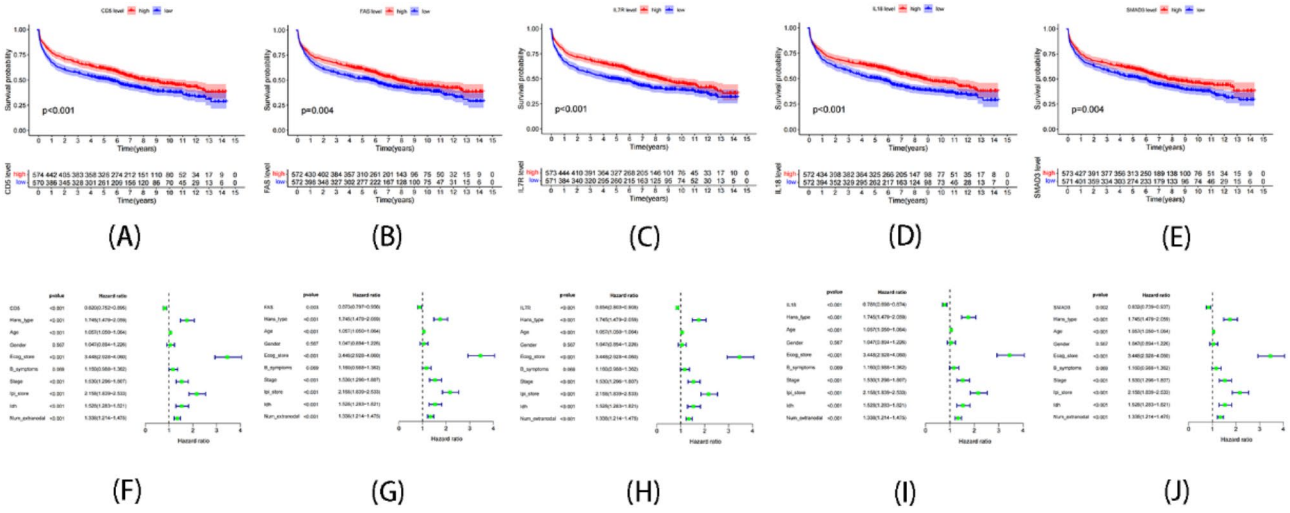

Fig. 8 Survival analysis and independent prognostic analysis for individual genes: Kaplan-Meier curve of CD5 (**A**), FAS (**B**), IL7R (**C**) IL18(**D**) and SMAD3 (**E**). The univariate Cox regression analyses of CD5 (**F**), FAS (**G**), IL7R (**H**), IL18(**I**) and SMAD3 (**J**)

correlation with overall survival. Additionally, we assessed the impact of signature gene expression on the survival of DLBCL patients through univariate Cox regression analysis. The results indicated that low expression of CD5 (HR: 0.820; 95% confidence interval [CI]: 0.752–0.895; $p < 0.001$; Fig. 8F), FAS (HR: 0.873; 95% CI: 0.797–0.956; $p = 0.003$; Fig. 8G), IL7R (HR: 0.854; 95% CI: 0.803–0.908; $p < 0.001$; Fig. 8H), IL18 (HR: 0.781; 95% CI: 0.698–0.874; $p < 0.001$; Fig. 8I), CSF1R (HR: 0.858; 95% CI: 0.764–0.964; $p = 0.010$; Fig.S1D), and SMAD3 (HR: 0.832; 95% CI: 0.739–0.937; $p = 0.002$; Fig. 8J) was associated with poorer survival probabilities. However, the expression of CCR7 ($p = 0.260$, Fig.S1C) was not significantly correlated with patient survival. These results suggest that SMAD3, IL7R, IL18, FAS, and CD5 can serve as biomarkers for predicting the prognosis of DLBCL.

M2 macrophage infiltration is closely associated with various pathways related to DLBCL

To further explore the relationship between M2 macrophage infiltration and DLBCL-related pathways, we conducted GSEA-KEGG pathway analysis, as shown in the figures. The results indicated that the cytosolic DNA sensing pathway, oxidative phosphorylation, and ribosome pathway were primarily enriched in the high-expression group of M2 macrophages (Fig. 9A), while the calcium signaling pathway, hematopoietic cell lineage, melanogenesis, Notch signaling pathway, and tight junctions were mainly enriched in the low-expression group of M2 macrophages (Fig. 9B). Next, we examined the differential activation pathways between the high-expression and low-expression groups based on the expression levels of M2 macrophages, incorporating GSVA. The results showed that the high-expression group of M2 macrophages might activate pathways such as peroxisome, PPAR signaling pathway, pyrimidine metabolism, and the metabolism of exogenous substances via cytochrome p450. In contrast, in the low-expression group of M2 macrophages, pathways related to meiosis, focal adhesion, T cell receptor signaling, and proximal tight junctions might be activated (Fig. 9C).

Discussions

DLBCL is the most common type of lymphoma, characterized by extensive heterogeneity. Despite significant advancements in diagnosis and treatment, particularly with the application of CAR-T therapy, which has notably improved survival rates for DLBCL patients, some patients still face poor treatment outcomes. Relapse and treatment resistance remain significant challenges in the management of DLBCL [27, 28]. Therefore, there is an urgent need for new early diagnostic methods, effective treatment strategies, and accurate prognostic assessment tools, and research on the immune microenvironment of DLBCL has gained increasing attention. A recent transcriptome-based expression clustering analysis established two immune-related epigenetic clusters, termed EC1 and EC2, with EC1 associated with poorer prognosis. Furthermore, EC1 and EC2 exhibit differing sensitivities to various drugs [29]. This new immune-related epigenetic characterization has strong clinical predictive value for DLBCL, particularly in guiding epigenetic therapeutic strategies.

The components of the TME, including tumor-associated macrophages (TAMs), myeloid-derived suppressor cells (MDSCs), and tumor-associated neutrophils (TANs), interact in complex ways with tumor cells and may contribute to treatment failures [30]. Recent studies have identified the significant role of macrophages in DLBCL; however, the specific mechanisms and targets

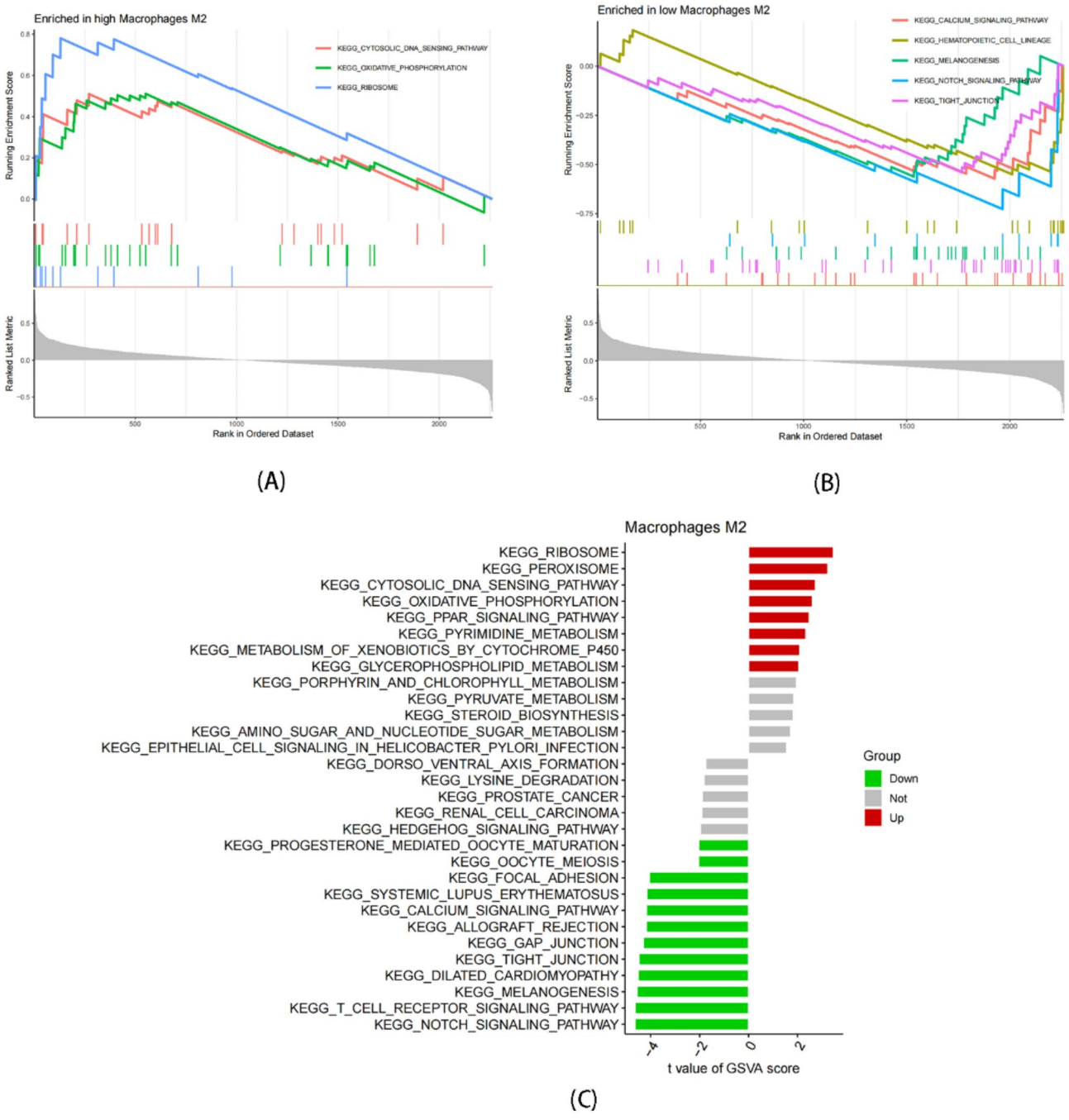

Fig. 9 The relative pathways between M2 macrophage infiltration and DLBCL. (**A**) The pathways enriched in the high-expression group of M2 macrophages. (**B**) The pathways enriched in the low-expression group of M2 macrophages. (**C**) The results of GSVA

of their action have not been thoroughly investigated. Utilizing bioinformatics methods for comprehensive analysis of expression profile data is one of the effective approaches to identify disease pathogenesis, biomarkers, and prognostic features, offering advantages of low cost and high efficiency. Machine learning, as an artificial intelligence approach, applies statistical algorithms to datasets, and is widely used for feature selection in high-throughput data. The combination of bioinformatics methods with machine learning provides a more reliable and effective technique for screening disease-related genes, making it a technological hotspot in omics research.

This study combines bioinformatics and machine learning to explore the mechanisms of immune cell infiltration in DLBCL, particularly focusing on the infiltration of M2 macrophages. Using the CIBERSORT algorithm, we revealed differences in the immune

microenvironment between DLBCL patients and control samples, notably finding a significantly higher proportion of M2 macrophages in the experimental group compared to the control group. This suggests that the infiltration of M2 macrophages may be closely related to the occurrence and development of DLBCL. Subsequently, we identified genes associated with M2 macrophages through WGCNA and intersected them with differentially expressed genes to obtain M2 macrophage-related differentially expressed genes. After performing GO and KEGG analyses on these genes, we discovered that they are associated with positive regulation of cell adhesion, positive regulation of T-cell activation, cytokine receptor activity, and pathways related to DLBCL. This further indicates that M2 macrophages play a crucial role in the pathogenesis of DLBCL.

Next, we utilized Cytoscape to identify key hub genes and further screened SMAD3, IL7R, IL18, FAS, CD5, CCR7, and CSF1R as key signature genes through machine learning algorithms. Survival analysis and independent prognostic analysis of these signature genes revealed a significant association between these genes and the survival rates of DLBCL patients, further supporting the importance of these genes in DLBCL.

Although this study reveals the important role of M2 macrophage-related genes in DLBCL, several limitations should be noted. First, the sample size used in this study is relatively small, which may affect the generalizability and reliability of the results. At the same time, we did not query the subtype information of DLBCL samples in the data set, and we could not determine the proportion of various subtypes. Therefore, the model we established may not be effective in predicting a certain subtype of DLBCL. Secondly, the research primarily relies on bioinformatics analysis and lacks essential in vitro and in vivo experimental validation, limiting a deeper understanding of the functions of the signature genes. Additionally, the expression differences of M2 macrophage-related genes across different subtypes of DLBCL and their clinical significance have not been thoroughly explored. To address these limitations, future research should focus on the following directions: first, laboratory studies should be conducted to validate the functions and mechanisms of the signature genes, particularly through investigations in cellular and animal models to ensure the validity of the results. And the relationship between marker genes and patient outcomes should be validated in our own clinical patient cohort. In addition, exploring the relationship between M2 macrophage-related genes and different subtypes of DLBCL will aid in developing personalized treatment strategies. Finally, the efficacy of targeted drugs against the signature genes in clinical applications should be evaluated to improve the prognosis of DLBCL patients.

Conclusion

In summary, this study provides new insights into the molecular mechanisms of DLBCL, particularly regarding the role of signature genes in the immune microenvironment and the potential therapeutic strategies targeting these genes, all of which warrant further exploration. Future research should focus on validating the functions of these signature genes and elucidating their mechanisms of interaction with immune cells, thereby providing new avenues for the diagnosis and treatment of DLBCL.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13062-025-00649-4.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Acknowledgements

We extend our sincere appreciation to all researchers who generously shared the data utilized in this study. We also express our gratitude to GEO for their invaluable contributions in providing the data to the public.

Author contributions

JZ: Conceptualization, Data curation, Writing – original draft. ZJ: Conceptualization, Data curation, Methodology, Software, Validation, Visualization, Writing – original draft. JZ: Methodology, Data curation, Software, Validation. XM: Conceptualization, Data curation, Writing – original draft. LA: Conceptualization, Methodology, Supervision, Writing–review & editing. All authors reviewed the manuscript.

Funding

The authors received no specific funding for this work.

Data availability

The datasets utilized in this study, GSE9327, GSE23647, GSE32018, GSE181063 and GSE83632, were procured from the GEO database. All datas and code necessary for the analyses are available upon request. For further information, please contact the corresponding author.

Declarations

Competing interests

The authors declare no competing interests.

Received: 22 February 2025 / Accepted: 5 April 2025
Published online: 29 April 2025

References

1. Shao R, Liu C, Xue R, Deng X, Liu L, Song C, et al. Tumor-derived Exosomal ENO2 modulates polarization of Tumor-associated macrophages through reprogramming Glycolysis to promote progression of diffuse large B-cell lymphoma. Int J Biol Sci. 2024;20(3):848–863.
2. Jia Z, Zhang J, Li Z, Ai L. Identification of ferroptosis-related genes associated with diffuse large B-cell lymphoma via bioinformatics and machine learning approaches. Int J Biol Macromol. 2024;282.
3. Feugier P, Van Hoof A, Sebban C, Solal-Celigny P, Bouabdallah R, Fermé C, et al. Long-Term results of the R-CHOP study in the treatment of elderly patients

with diffuse large B-Cell lymphoma: A study by the groupe d'etude des lymphomes de L'Adulte. J Clin Oncol. 2005;23(18):4117–4126.
4. Pfreundschuh M, Trümper L, Österborg A, Pettengell R, Trneny M, Imrie K, et al. CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in young patients with good-prognosis diffuse large-B-cell lymphoma: a randomised controlled trial by the MabThera international trial (MInT) group. Lancet Oncol. 2006;7(5):379–391.
5. Pfreundschuh M, Schubert J, Ziepert M, Schmits R, Mohren M, Lengfelder E, et al. Six versus eight cycles of bi-weekly CHOP-14 with or without rituximab in elderly patients with aggressive CD20 + B-cell lymphomas: a randomised controlled trial (RICOVER-60). Lancet Oncol. 2008;9(2):105–116.
6. Coiffier B, Thieblemont C, Van Den Neste E, Lepeu G, Plantier I, Castaigne S, et al. Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: a study by the groupe d'etudes des lymphomes de L'Adulte. Blood. 2010;116(12):2040–2045.
7. Coiffier B, Sarkozy C. Diffuse large B-cell lymphoma: R-CHOP failure—what to do? Hematology. 2016;2016(1):366–378.
8. Nicholas NS, Apollonio B, Ramsay AG. Tumor microenvironment (TME)-driven immune suppression in B cell malignancy. Biochimica et biophysica acta (BBA) -. Mol Cell Res. 2016;1863(3):471–482.
9. Wang L, Qin W, Huo Y-J, Li X, Shi Q, Rasko JEJ et al. Advances in targeted therapy for malignant lymphoma. Signal Transduct Target Therapy. 2020;5(1).
10. Zhang Q, Sioud M. Tumor-Associated macrophage subsets: shaping polarization and targeting. Int J Mol Sci. 2023;24(8).
11. Li M, Yang Y, Xiong L, Jiang P, Wang J, Li C. Metabolism, metabolites, and macrophages in cancer. J Hematol Oncol. 2023;16(1).
12. Qian B-Z, Pollard JW. Macrophage diversity enhances tumor progression and metastasis. Cell. 2010;141(1):39–51.
13. Li C, Xu X, Wei S, Jiang P, Xue L, Wang J. Tumor-associated macrophages: potential therapeutic strategies and future prospects in cancer. J Immunother Cancer. 2021;9(1).
14. Binnewies M, Roberts EW, Kersten K, Chan V, Fearon DF, Merad M, et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat Med. 2018;24(5):541–550.
15. Beatty GL, Gladney WL. Immune escape mechanisms as a guide for cancer immunotherapy. Clin Cancer Res. 2015;21(4):687–692.
16. Liu S, Zhang H, Li Y, Zhang Y, Bian Y, Zeng Y et al. S100A4 enhances protumor macrophage polarization by control of PPAR-γ-dependent induction of fatty acid oxidation. J Immunother Cancer. 2021;9(6).
17. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–457.
18. Zhang S, Zhang E, Long J, Hu Z, Peng J, Liu L, et al. Immune infiltration in renal cell carcinoma. Cancer Sci. 2019;110(5):1564–1572.
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1).
20. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–2504.
21. Yang C, Ren J, Li B, Jin C, Ma C, Cheng C et al. Identification of gene biomarkers in patients with postmenopausal osteoporosis. Mol Med Rep. 2018.
22. Qiu J, Peng B, Tang Y, Qian Y, Guo P, Li M, et al. CpG methylation signature predicts recurrence in Early-Stage hepatocellular carcinoma: results from a multicenter study. J Clin Oncol. 2017;35(7):734–742.
23. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013;14(1).
24. Keane C, Gill D, Vari F, Cross D, Griffiths L, Gandhi M. CD4 + Tumor infiltrating lymphocytes are prognostic and independent of R-IPI in patients with DLBCL receiving R-CHOP chemo-immunotherapy. Am J Hematol. 2013;88(4):273–276.
25. Matias A, Suvi-Katri L, Oscar B, Satu M, Judit Mészáros J, Marja-Liisa K-L, et al. Immune cell constitution in the tumor microenvironment predicts the outcome in diffuse large B-cell lymphoma. Haematologica. 2020;106(3):718–729.
26. Tabanelli V, Melle F, Motta G, Mazzara S, Fabbri M, Agostinelli C, et al. The identification of TCF1 + progenitor exhausted T cells in THRLBCL May predict a better response to PD-1/PD-L1 Blockade. Blood Adv. 2022;6(15):4634–4644.
27. Hawkins ER, D'Souza RR, Klampatsa A. Armored CAR T-Cells: the next chapter in T-Cell cancer immunotherapy. Biol Targets Ther. 2021;15:95–105.
28. Asmamaw Dejenie T, Tiruneh G, Medhin M, Dessie Terefe G, Tadele Admasu F, Wale Tesega W, Chekol Abebe E. Current updates on generations, approvals, and clinical trials of CAR T-cell therapy. Hum Vaccines Immunotherapeutics. 2022;18(6).
29. Wang X, Hong Y, Meng S, Gong W, Ren T, Zhang T et al. A novel immune-related epigenetic signature based on the transcriptome for predicting the prognosis and therapeutic response of patients with diffuse large B-cell lymphoma. Clin Immunol. 2022;243.
30. Sinkarevs S, Strumfs B, Volkova S, Strumfa I. Tumour microenvironment: the general principles of pathogenesis and implications in diffuse large B cell lymphoma. Cells. 2024;13(12).

Publisher's note