Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Development of a diagnostic model based on glycolysis-related genes and immune infiltration in intervertebral disc degeneration

Jian Gao [a], Liming He [b], Jianguo Zhang [b], Leimin Xi [b], Haoyu Feng [b,*]

[a] Third Hospital of Shanxi Medical University, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital, 030032, Taiyuan, China
[b] Department of Orthopedics, Third Hospital of Shanxi Medical University, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital, 030032, Taiyuan, China

ABSTRACT

Background: The glycolytic pathway and immune response play pivotal roles in the intervertebral disc degeneration (IDD) progression. This study aimed to develop a glycolysis-related diagnostic model and analyze its relationship with the immune response to IDD.
Methods: GSE70362, GSE23130, and GSE15227 datasets were collected and merged from the Gene Expression Omnibus, and differential expression analysis was performed. Glycolysis-related differentially expressed genes (GLRDEGs) were identified, and a machine learning-based diagnostic model was constructed and validated, followed by Gene Set Enrichment Analysis (GSEA). Gene Ontology functional enrichment and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analyses were performed, and mRNA-miRNA and mRNA-transcription factor (TF) interaction networks were constructed. Immune infiltration was analyzed using single-sample GSEA (ssGSEA) and cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm between high- and low-risk groups.
Results: In the combined dataset, samples from 31 patients with IDD and 55 normal controls were analyzed, revealing differential expression of 16 GLRDEGs between the two groups. Using advanced machine learning techniques (LASSO, support vector machine, and random forest algorithms), we identified eight common GLRDEGs (*PXK*, *EIF3D*, *WSB1*, *ZNF185*, *IGFBP3*, *CKAP4*, *RPL15*, and, *SSR1*) and developed a diagnostic model, which demonstrated high accuracy in distinguishing IDD from control samples (area under the curve, 0.935). We identified 42 mRNA-miRNA and 33 mRNA-TF interaction pairs. Using the RiskScore from the diagnostic model, the combined dataset was stratified into high- and low-risk groups. SsGSEA revealed significant differences in the infiltration abundances of the four immune cell types between the groups. The CIBERSORT algorithm revealed the strongest correlation between resting natural killer (NK) cells and *ZNF185* in the low-risk group and between CD8^{+} T cells and *SSR1* in the high-risk group.
Conclusions: Our study reveals a potential interplay between glycolysis-associated genes and immune infiltration in IDD pathogenesis. These findings contribute to our understanding of IDD and may guide development of novel diagnostic markers and therapeutic interventions.

1. Introduction

Lower back pain (LBP) is the leading cause of productivity loss and disability worldwide [1]. Various multifactorial causes and risk factors contribute to the pathogenesis of LBP, with intervertebral disc degeneration (IDD) emerging as the predominant underlying factor [2]. IDD is an increasingly common health problem and poses significant economic and social burdens in countries with rapidly aging populations [3]. Moreover, the clinical management of IDD that in place today is not ideal. The original biology of the disc cannot be fundamentally restored by pharmacological or physiological treatments for early degenerative changes, or by disc resection, fusion, or replacement for advanced degenerative changes [4].

* Corresponding author.
E-mail address: fenghaoyuspine@126.com (H. Feng).

https://doi.org/10.1016/j.heliyon.2024.e36158
Received 4 January 2024; Received in revised form 3 August 2024; Accepted 11 August 2024
Available online 13 August 2024

Table 1
Intervertebral Disc Degeneration Datasets Information list.

	GSE70362	GSE23130	GSE15227
Platform	GPL17810	GPL1352	GPL1352
Species	Homo sapiens	Homo sapiens	Homo sapiens
Tissue	disc tissue - annulus fibrosus and nucleus pulposus	Disc	Disc
Samples in Control group	28	15	12
Samples in IDD group	20	8	3

The intervertebral disc (IVD) is a critical component of the spine and is made up of fibrous cartilage. It is composed of the nucleus pulposus (NP), annulus fibrosus, and cartilaginous endplates. The IVD plays a vital role by providing flexibility, acting as a shock absorber, and ensuring vertebral stability [5]. The pathogenesis of IDD is complex and overlapping, and is associated with the loss of homeostatic balance in the disc environment, leading to a catabolic and hypoxic microenvironment, a senescent cell profile, and consequent immunometabolic alterations [6]. However, the specific mechanism and pathogenesis of disc degeneration remain unclear.

Glycolysis is a fundamental cellular metabolic pathway that efficiently converts glucose into lactate to produce energy [7]. It plays a vital role as a major energy source, particularly in cells under hypoxic conditions. Increased levels of glycolysis promote the proliferation, invasion, and migration of certain cancer cells by activating various signaling pathways and increasing drug resistance [8]. Researchers have found that by intervening in the glycolytic process, the energy production and metabolic activity of chondrocytes can be significantly regulated, slowing down the progression of osteoarthritis (OA) [9]. Other bone and joint disorders, including rheumatoid arthritis [10], osteoporosis [11], and IDD [12], have also shown significant alterations in the glycolytic metabolic processes.

The IVD tissues exhibit unique metabolic characteristics. Owing to its limited blood supply and lack of neural innervation, the IVD primarily receives blood through capillaries in the outer layers of the disc, whereas nutrient exchange in the central NP relies on diffusion from nearby blood vessels in the endplates, facilitated by the porous structure of the cartilaginous endplates [13]. Compared to plasma, less oxygen and glucose are available for use. In such a low-oxygen environment, the glycolytic metabolic pathway serves as a key energy source for the NP [14]. However, disruptions in this pathway can result in a progressive decline in the number of IVD cells and extracellular matrix, along with structural changes in the fibrous ring and accelerated disc degeneration [15].

The IVD is traditionally considered an immunologically privileged organ that shields NP tissue from the host immune system [16]. However, when the blood-NP barrier is breached, an immune response is triggered, which plays a significant role in IVD degeneration and the subsequent pathological processes. Immune cells such as macrophages [17] and T cells [18] are involved in the inflammatory response of the disc tissue and regulate the inflammatory process and cellular metabolism through the release of cytokines such as tumor necrosis factor-α and interleukin-1β [19]. However, few studies have reported on the role of immune infiltration in the development of disc degeneration, and the relationship between disc degeneration and immune infiltration requires further research to be fully understood.

The objective of this study was to identify glycolysis-related biomarkers and potential therapeutic targets for IDD management using bioinformatics. Initially, three distinct disc degeneration datasets retrieved from the Gene Expression Omnibus (GEO) database were amalgamated, resulting in a combined dataset. Subsequently, we conducted differential expression analysis and Gene Set Enrichment Analysis (GSEA) of samples from the combined dataset. This analysis intersected with glycolysis-related genes (GLRGs) to yield glycolysis-related differentially expressed genes (GLRDEGs). We then constructed GLRDEG diagnostic models using Logistic-Least Absolute Shrinkage and Selection Operator (Logistic-LASSO), support vector machine (SVM), and random forest (RF) algorithms to identify common GLRDEGs. Subsequent steps included Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses and construction of mRNA-miRNA and mRNA-transcription factor (mRNA-TF) interaction networks. Finally, based on the RiskScore derived from the diagnostic model, the combined dataset samples were divided into high- and low-risk groups. Differential analysis and GSEA enrichment analysis were performed, followed by single-sample GSEA (ssGSEA) and CIBERSORT immune signature differential analysis. In summary, our study uncovered new perspectives on the interplay between glycolysis-associated genes and immune cell infiltration in IDD pathogenesis.

2. Materials and methods

2.1. Data download

Data retrieval was performed from GEO datasets using the keywords “intervertebral disc degeneration” and “Homo sapiens”. The following selection criteria were used for data retrieval: 1) gene expression profile of human intervertebral disc tissue samples; 2) Samples were divided into normal controls and IDD group; 3) At least 15 samples were included in the dataset. A total of three datasets met the screening criteria. We downloaded the gene expression profile datasets (GSE70362 [20], GSE23130 [21], and GSE15227 [22]) of patients with IVD from the GEO database [23] using the R package GEOquery [24]. The source species for the three GEO datasets was *Homo sapiens*. The annotations of the probe names in the dataset used the chip GPL platform file (see Table 1 for detailed information).

For the analysis, we selected 20 degenerated IVD samples (Grade: IV, V; group: IDD) and 28 normal control samples (Grade: I, II, III; group: Control) from the GSE70362 dataset; a total of 23 samples from the GSE23130 dataset, including 15 normal control samples

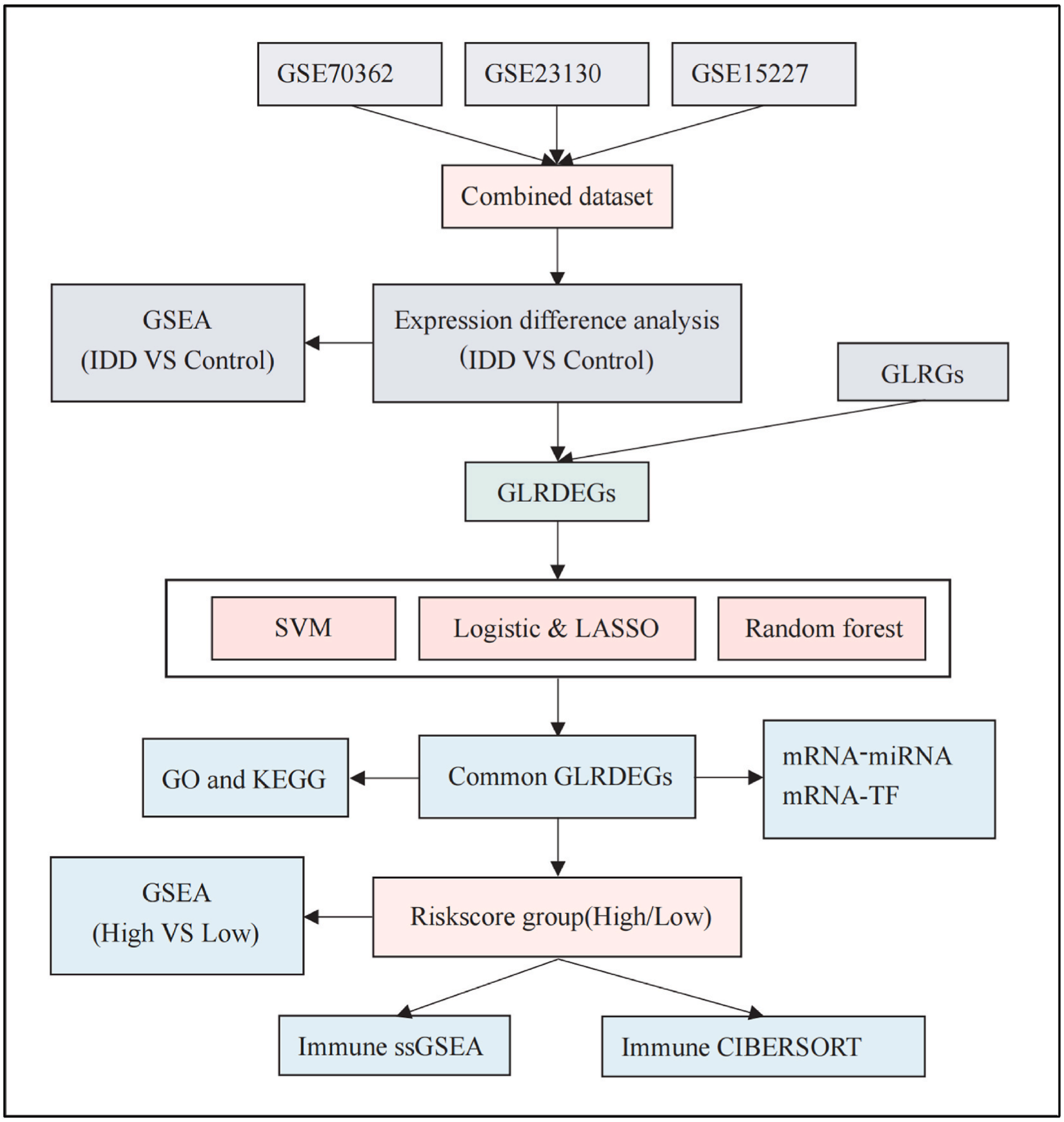


Fig. 1. Technology roadmap.

(Grade: I, II, III; grouping: Control) and eight degenerated IVD samples (Grade: IV, V; grouping: IDD); and a total of 15 samples from the GSE15227 dataset, including 12 normal control samples (Grade: I, II, III; group: Control) and three degenerated IVD samples (Grade: IV; group: IDD).

We collected GLRGs from the GeneCards [25] database (https://www.genecards.org/), which provides comprehensive information on human genes. We used "glycolysis" as the search keyword to obtain the GLRGs in the GeneCards database and obtained a total of 2661 GLRGs. The specific gene names are listed in Supplement Table 1.

2.2. Data Preprocessing and differential expression analysis

We merged the three datasets (GSE70362, GSE23130, and GSE15227), used the ComBat function of R's sva package [26] to batch the data, and then used the ControlizeBetweenArrays function of the limma package [27] to standardize, thus obtaining the combined dataset (31 IDD group samples, 55 control group samples).

Subsequently, we used the limma package in R to conduct expression analysis of all genes between the IDD group samples and control group samples from the combined dataset, identifying differentially expressed genes (DEGs) based on the stringent criteria of $|\log FC| > 0.3$ and P.adj <0.05 for further investigation. A volcano map was generated using the R package ggplot2, and the results are presented. We then intersected the GLRGs and DEGs to obtain GLRDEGs.

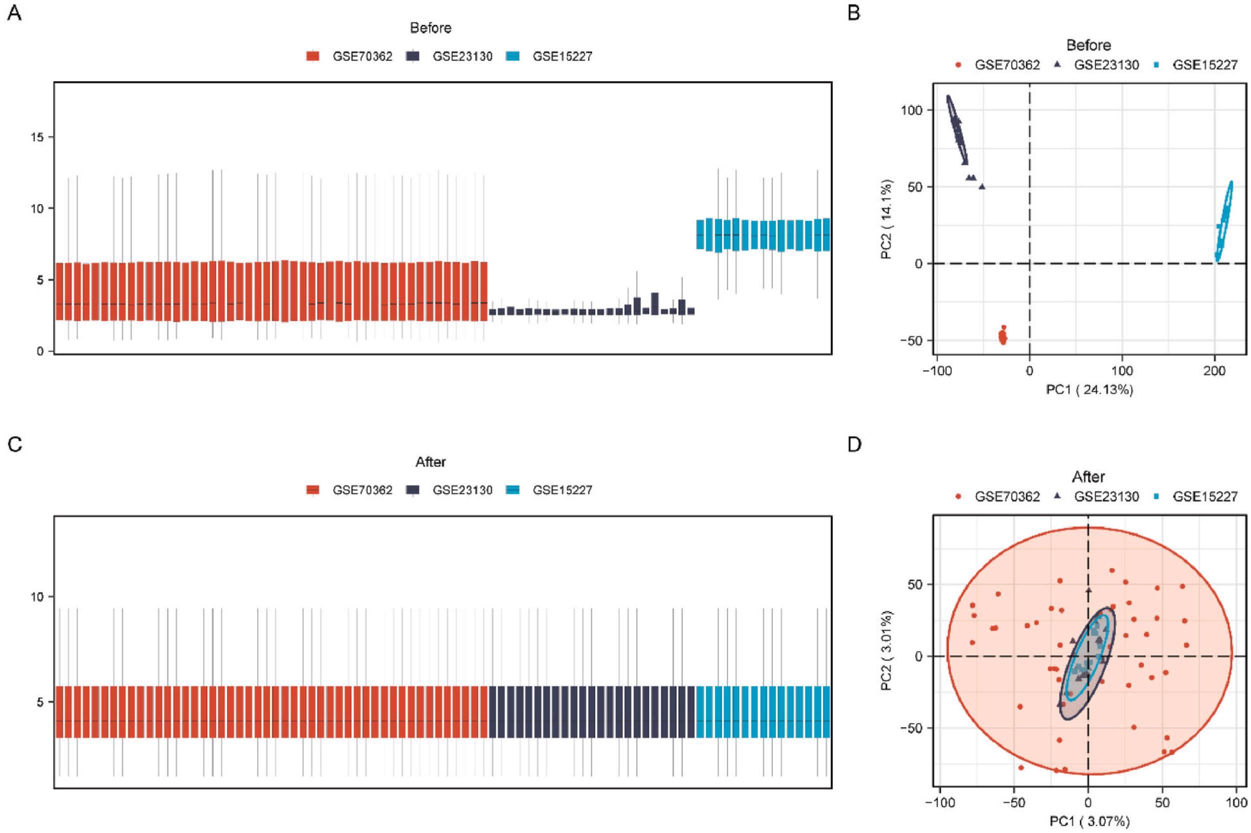


Fig. 2. Dataset standardization processing.
Boxplot (A) and PCA plot (B) of the combined dataset prior to batch effect removal and normalization. Combined dataset removes the batch effect, normalized boxplot plot (C) and PCA plot (D).

2.3. *GSEA enrichment analysis*

GSEA [28] is commonly used to estimate changes in pathway and biological process activities in expression datasets. In this study, based on the positive and negative sorting of the logFC values of the genes, all genes in the different groups of the combined dataset were divided into two groups. The clusterProfiler package was used to analyze positive and negative logFC values in the two groups. For GSEA of all genes, the parameters are as follows: the seed is 2022, the number of calculations is 1000, each gene set contains at least 10 genes, the maximum number of genes is 500, and the p-value correction method is Benjamini-Hochberg (BH). We obtained the "c2. cp.all.v2022.1.Hs.symbols.gmt [All Canonical Pathways] (3050)" gene set from the Molecular Signatures Database (MSigDB) [29]. The screening criteria for significant enrichment were P.adj <0.05 and false discovery rate (FDR) value (q.value) < 0.05.

2.4. *Construction of diagnostic model*

In this study, we employed GLRDEGs to develop an SVM model. This was achieved by applying the SVM algorithm [30] to the expression matrix and grouping the data in the combined dataset. The model was refined to identify genes with the highest accuracy and lowest error rates based on GLRDEGs to ensure optimal gene selection for diagnostic purposes.

The RF [31] is an algorithm that integrates multiple decision trees using ensemble learning. The randomForest package was used to construct a model based on the expression of GLRDEGs in the expression matrix of the combined dataset. The parameters were set.seed (234) and ntree = 1000.

To develop a logistic diagnostic model for the combined dataset, we conducted logistic regression analysis of GLRDEGs, differentiating between the IDD and control groups. GLRDEGs with a p-value <0.05 were selected to construct the model. Molecular expression data for this model were visualized using forest plots. Further, we applied the R package glmnet [32] (with set.seed set to 500 and family to "binomial") for LASSO [33] regression analysis on the GLRDEGs. This process aimed to refine the logistic regression model (termed Logistic-LASSO) and mitigate overfitting, with a run period of 500. LASSO regression, which is an extension of linear regression with a penalty term (lambda times the absolute value of the coefficient), reduces overfitting and enhances the generalizability of the model. The outcomes of the LASSO regression analysis were illustrated using the diagnostic model and variable trajectory diagrams, providing a comprehensive view of the model's performance and the variable selection process.

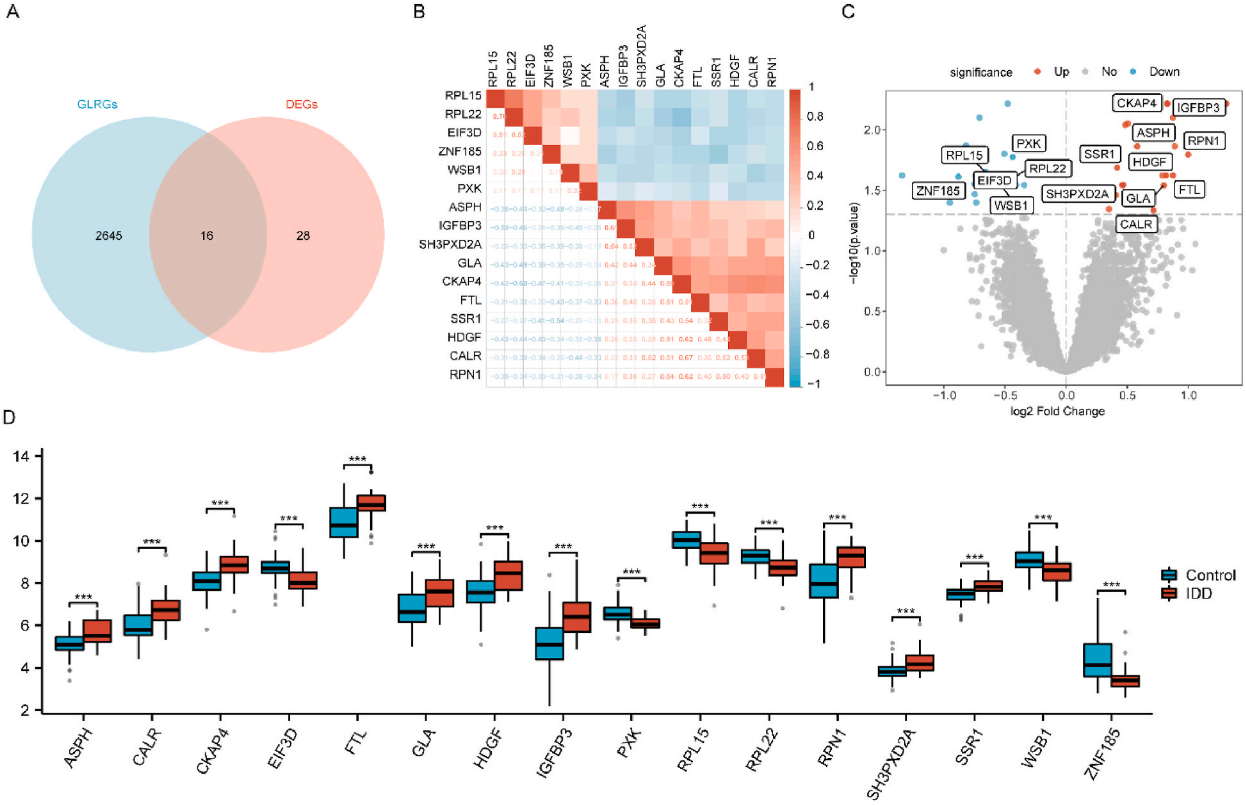


Fig. 3. Expression difference analysis and correlation analysis of GLRDEGs.
A. Venn diagram showing the intersection of differentially expressed genes and glycolysis-related genes between the disease and control groups in the combined dataset. B. Correlation heat map based on the expression matrix of 16 GLRDEGs in the combined dataset. C. Volcano map drawn according to the difference analysis results between the disease and control groups in the combined dataset, and the marked genes are GLRDEGs. D. Group comparison chart of 16 GLRDEGs between disease group and control group in combined dataset. ns is equal to $P \geq 0.05$, which is not statistically significant; *$P < 0.05$, statistically significant; **$P < 0.01$, highly statistically significant; and ***$P < 0.001$, very statistically significant.

We then intersected the GLRDEGs identified in the logistic (Logistic-LASSO) regression, SVM, and RF models. This intersection was visualized using a Venn diagram to identify common GLRDEGs. Subsequently, we combined the coefficients of common GLRDEGs from the Logistic-LASSO regression model with their expression levels in the combined dataset. This approach enabled us to develop a comprehensive GLRDEGs diagnostic model and calculate corresponding risk scores.

In our analysis, we utilized nomograms [34], which employ a series of disjoint line segments in a Cartesian coordinate system to represent the functional relationship between multiple independent variables. Specifically, genes identified in the GLRDEGs diagnostic model were analyzed using the R package ’rms’ based on Logistic-LASSO regression. We performed logistic regression on the expression levels of these genes within the combined dataset and constructed a nomogram to visually represent the results of the GLRDEGs analysis.

Additionally, to assess the accuracy and resolution of our GLRDEGs diagnostic model, we conducted a calibration analysis and generated a calibration curve. Moreover, decision curve analysis (DCA) [35], a method for evaluating clinical prediction models, diagnostic tests, and molecular markers, was employed. We used the R package ‘ggDCA’ to create DCA diagrams specifically for our GLRDEGs, thereby evaluating the accuracy and resolution of the diagnostic model.

2.5. Gene Function enrichment analysis (GO) and pathway enrichment (KEGG) analysis

GO [36] analysis is a common method for large-scale functional enrichment research, including biological process (BP), molecular function (MF), and cellular components (CC). The KEGG [37] is a widely used database that stores information on genomes, biological pathways, diseases, drugs, etc. We used the R package clusterProfiler [38] to perform GO and KEGG annotation analyses of GLRDEGs. The item selection criteria were P.adj <0.05 and FDR value (q.value) < 0.05 is statistically significant, and the BH method was used for p-value correction.

Table 2
GSEA enrichment analysis results of Combined dataset Control-IDD group genes.

ID	enrichmentScore	NES	pvalue	p.adjust	qvalue
KEGG OXIDATIVE PHOSPHORYLATION	0.520952	2.111817	0.001927	0.029484	0.024768
REACTOME HEDGEHOG LIGAND BIOGENESIS	0.547656	2.02681	0.002058	0.029484	0.024768
REACTOME SIGNALING BY NOTCH4	0.467794	1.831013	0.002024	0.029484	0.024768
WP IL9 SIGNALING PATHWAY	0.660696	1.828353	0.002141	0.029484	0.024768
REACTOME MAPK6 MAPK4 SIGNALING	0.473459	1.861385	0.002016	0.029484	0.024768
REACTOME SIGNALING BY NOTCH	0.356717	1.570358	0.001876	0.029484	0.024768

2.6. *Construction of mRNA-miRNA, mRNA-TF interaction network*

The ENCORI [39] database (https://starbase.sysu.edu.cn/) is version 3.0 of the starBase database, and the miRNA-mRNA interactions in the ENCORI database are based on CLIP-seq and degradome sequencing (for plants). Data mining provides a variety of visualization interfaces for exploring miRNA targets. We used the ENCORI database to predict miRNAs interacting with GLRDEGs and then constructed an mRNA-miRNA interaction network in Cytoscape.

The CHIPBase database (version 3.0) [40] (https://rna.sysu.edu.cn/chipbase/) identified thousands of binding motif matrices and their binding sites from the ChIP-seq data of DNA-binding proteins and predicted the transcriptional regulatory relationship between millions of TFs and genes. The hTFtarget [41] database (http://bioinfo.life.hust.edu.cn/hTFtarget) is a comprehensive database containing information on human TFs and their corresponding regulatory targets. We searched for TFs binding to key genes using the CHIPBase (version 3.0) and hTFtarget databases and visualized them using Cytoscape software.

2.7. *Identification and correlation analysis of immune infiltrating cells*

We used the ssGSEA algorithm to quantify the relative abundance of each immune cell infiltrate; mark various infiltrating immune cell types, such as CD8$^+$ T cells, dendritic cells, macrophages, regulatory T cells, and other human immune cell subtypes; and use the enrichment fraction calculated by ssGSEA analysis to represent the relative abundance of each immune cell infiltration in each sample [42,43]. We analyzed the ssGSEA algorithm in the R package GSVA [44] and calculated the enrichment scores of the high- and low-risk grouping samples to represent the infiltration levels of different types of immune cells in each sample. The difference in the infiltration abundance of immune cells between samples from the high- and low-risk groups is displayed in a boxplot. The correlation was calculated using the Spearman algorithm and visualized using a correlation point diagram.

CIBERSORT [45] is an immune infiltration analysis algorithm that deconvolutes a transcriptome expression matrix based on the principle of linear support vector regression, thereby estimating the composition and abundance of immune cells in mixed cells. We uploaded the expression matrix data of different groups of samples in the combined dataset to CIBERSORT, combined with the LM22 characteristic gene matrix; screened out data with immune cell enrichment scores greater than zero; and finally obtained and displayed the specific results of the immune cell infiltration abundance matrix.

The differences in the infiltration abundance of immune cells in the samples between the different groups of the combined dataset are displayed in a stacked histogram. The correlation between different immune cells in the combined dataset was calculated using the Spearman algorithm and visualized using the R package ggplot2. We then combined the gene expression matrix of the combined dataset to calculate the correlation between immune cells and GLRDEGs and drew a correlation dot plot using the R package ggplot2.

2.8. *Statistical analysis*

All data processing and analyses in this study were performed using R software (Version 4.1.2). For the comparison of two groups of continuous variables, the statistical significance of normally distributed variables was estimated using the independent Student's t-test, and the Mann–Whitney *U* test was used (i.e., Wilcoxon rank sum test) to analyze the differences among non-normally distributed variables. If not specified, the results were calculated using Spearman correlation analysis to determine the correlation coefficients between different molecules; all statistical p-values were two-sided, and $P < 0.05$ is considered statistically significant.

3. Results

3.1. *Dataset processing*

The mRNA expression profile of IDD was downloaded from GEO, and the data were analyzed as in the roadmap (Fig. 1). We merged three IDD datasets (GSE70362, GSE23130, and GSE15227) and applied the 'ComBat' function from R's 'sva' package [26] for batch normalization, followed by standardization using the 'ControlizeBetweenArrays' function of the limma package [27]. This resulted in a combined dataset comprising 31 IDD and 55 control samples.

Boxplots and principal component analysis (PCA) diagrams of the combined dataset were generated to illustrate the data distribution before (Fig. 2A and B) and after (Fig. 2C and D) processing, categorized by sample sources. Post-processing results indicated a more uniform expression pattern across samples in the combined dataset, suggesting the successful mitigation of batch effects.

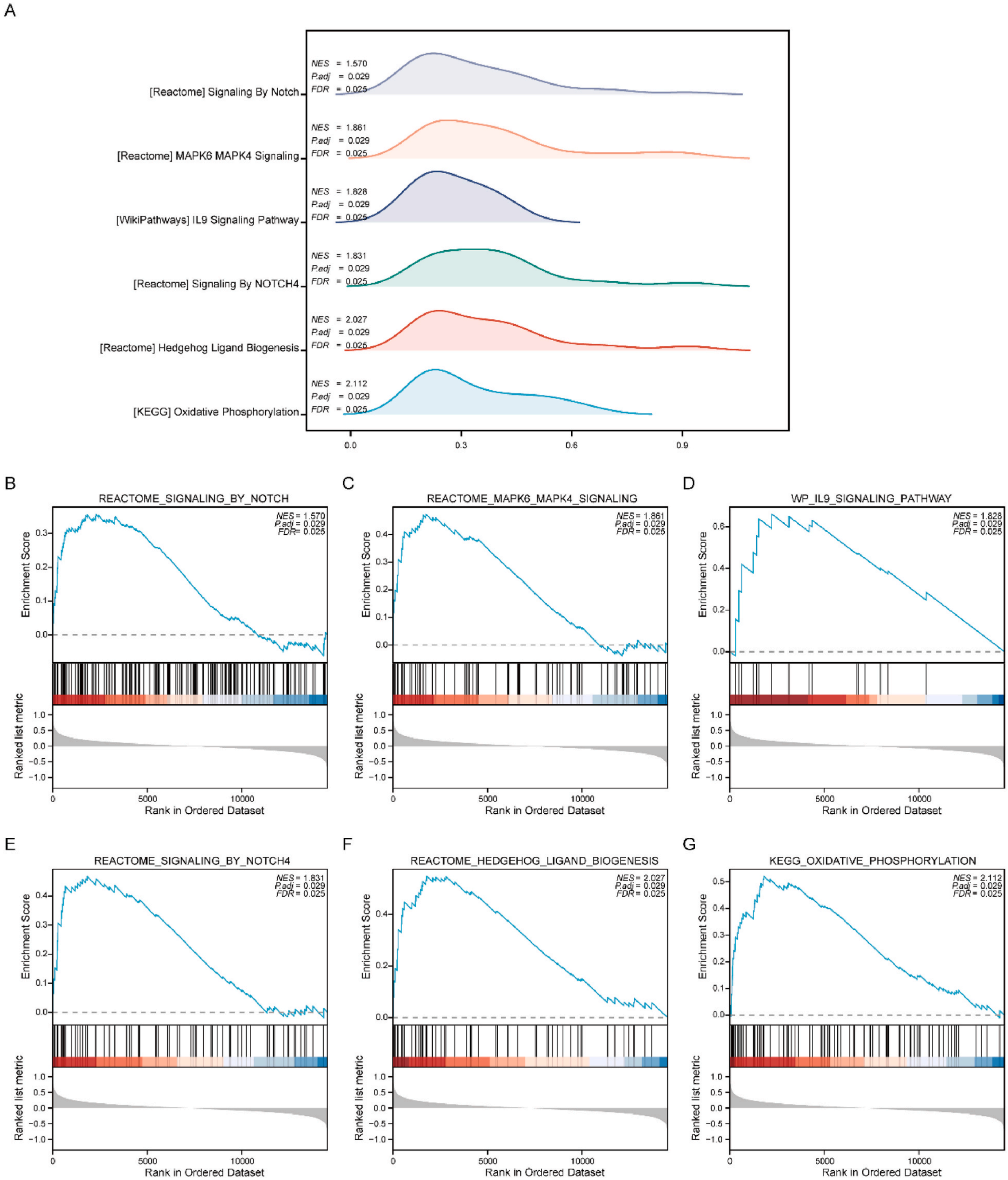


Fig. 4. GSEA of combined dataset.

(A) Six main biological characteristics from the GSEA of genes between different groups (IDD/control) in the combined dataset. B-G. Genes in the IDD dataset were significantly enriched in Notch signaling (B), MAPK6 MAPK4 signaling (C), the IL9 signaling pathway (D), Notch4 signaling (E), Hedgehog ligand biogenesis (F), and oxidative phosphorylation (G). The significant enrichment screening criteria for GSEA enrichment analysis were P. adj <0.05 and FDR value (q. value) < 0.05.

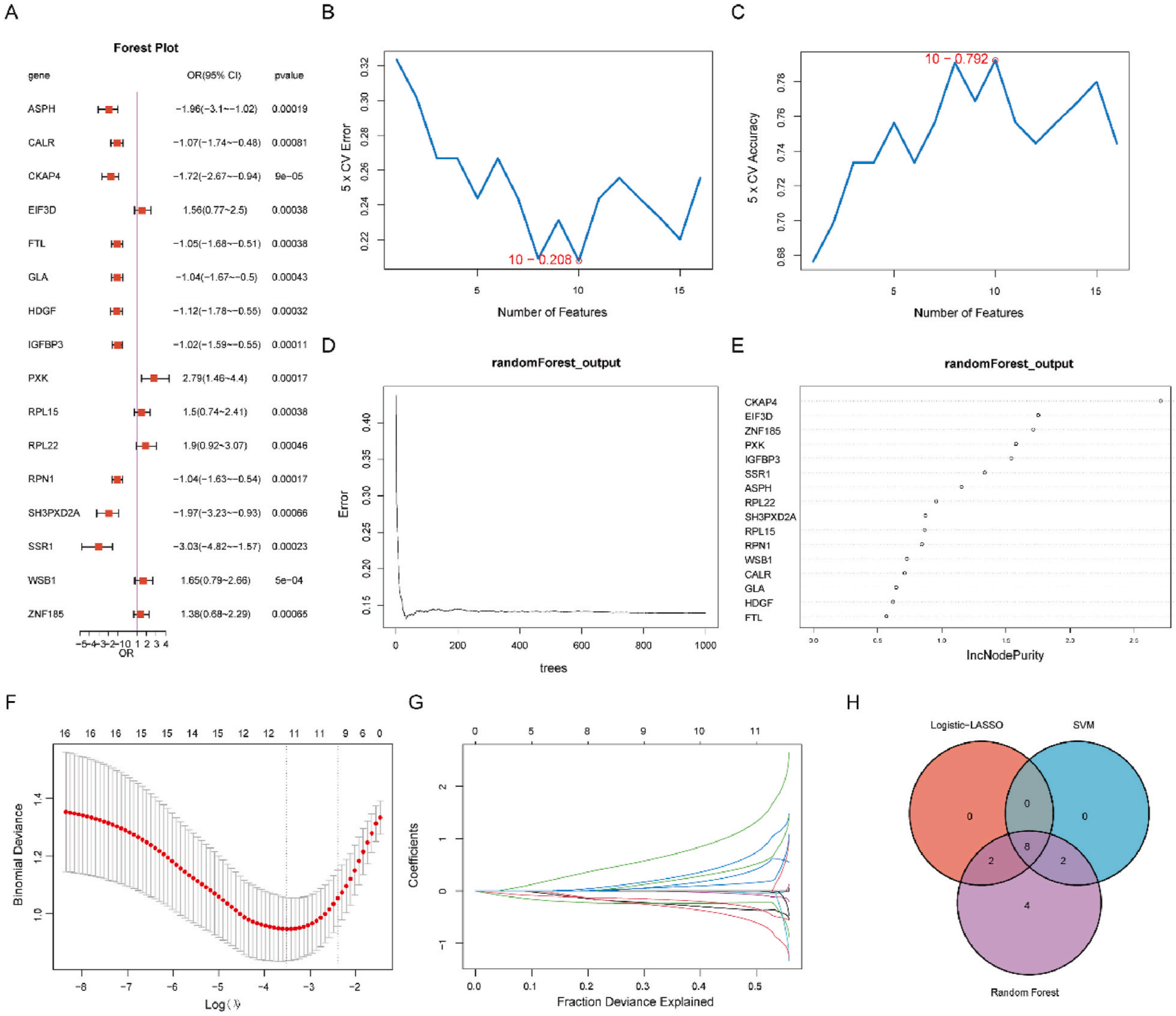


Fig. 5. Construction of diagnostic model.
A. Forest Plot of logistic regression model of GLRDEGs. B. The number of genes with the lowest error rate obtained by the SVM algorithm. C. The number of genes with the highest accuracy obtained by the SVM algorithm. D. Model training error plot for the RF algorithm. E. RF model showing GLRDEGs (in descending order of IncNodePurity). F. Diagnostic model diagram of the LASSO regression model. G. Variable trajectory plot of the LASSO regression model. H. GLRDEGs and SVM models in Logistic-LASSO regression model, GLRDEGs Venn diagram in RF model.

Subsequent analyses were conducted on the batch-effect-corrected dataset.

3.2. Identification of glycolysis-related DEGs

We used the limma package to analyze the differential expression between the IDD group samples and the control group samples in the combined dataset, with 44 genes that met the threshold of |logFC| >0.30 and P.adj <0.05. Under this threshold, in the high RiskScore group, 20 genes showed high expression and 24 genes showed low expression. We intersected these 44 DEGs with 2661 GLRGs; as shown in Fig. 3A, a total of 16 GLRDEGs were obtained, including *CKAP4*, *IGFBP3*, *ASPH*, *RPN1*, *PXK*, *SSR1*, *RPL15*, *FTL*, *HDGF*, *EIF3D*, *ZNF185*, *RPL22*, *WSB1*, *SH3P XD2A*, *GLA*, and *CALR*.

We then generated a correlation heatmap using the expression matrix of the 16 GLRDEGs in the combined dataset (Fig. 3B). The results revealed two distinct correlation patterns among these genes. The first group, comprising *ASPH*, *IGFBP3*, *SH3PXD2A*, *GLA*, *CKAP4*, *FTL*, *SSR1*, *HDGF*, *CALR*, and *RPN1*, exhibited positive correlations. The second group, which included *EIF3D*, *PXK*, *RPL15*, *RPL22*, *WSB1*, and *ZNF185*, also showed positive correlations within its members. However, the second group displayed a negative correlation with the first group of GLRDEGs.

We also used a volcano map (Fig. 3C) to show the results of the difference analysis between the disease and control groups of the

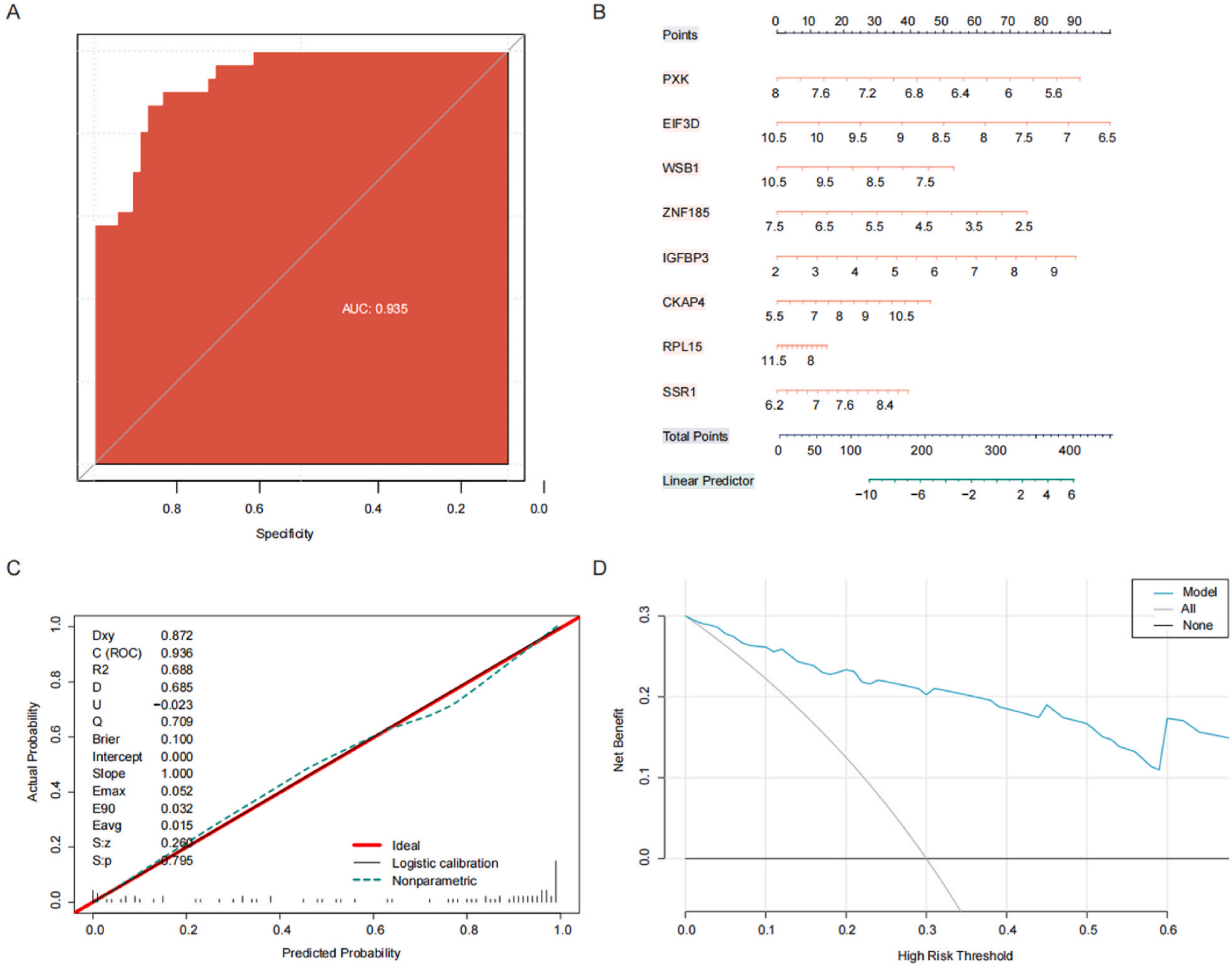
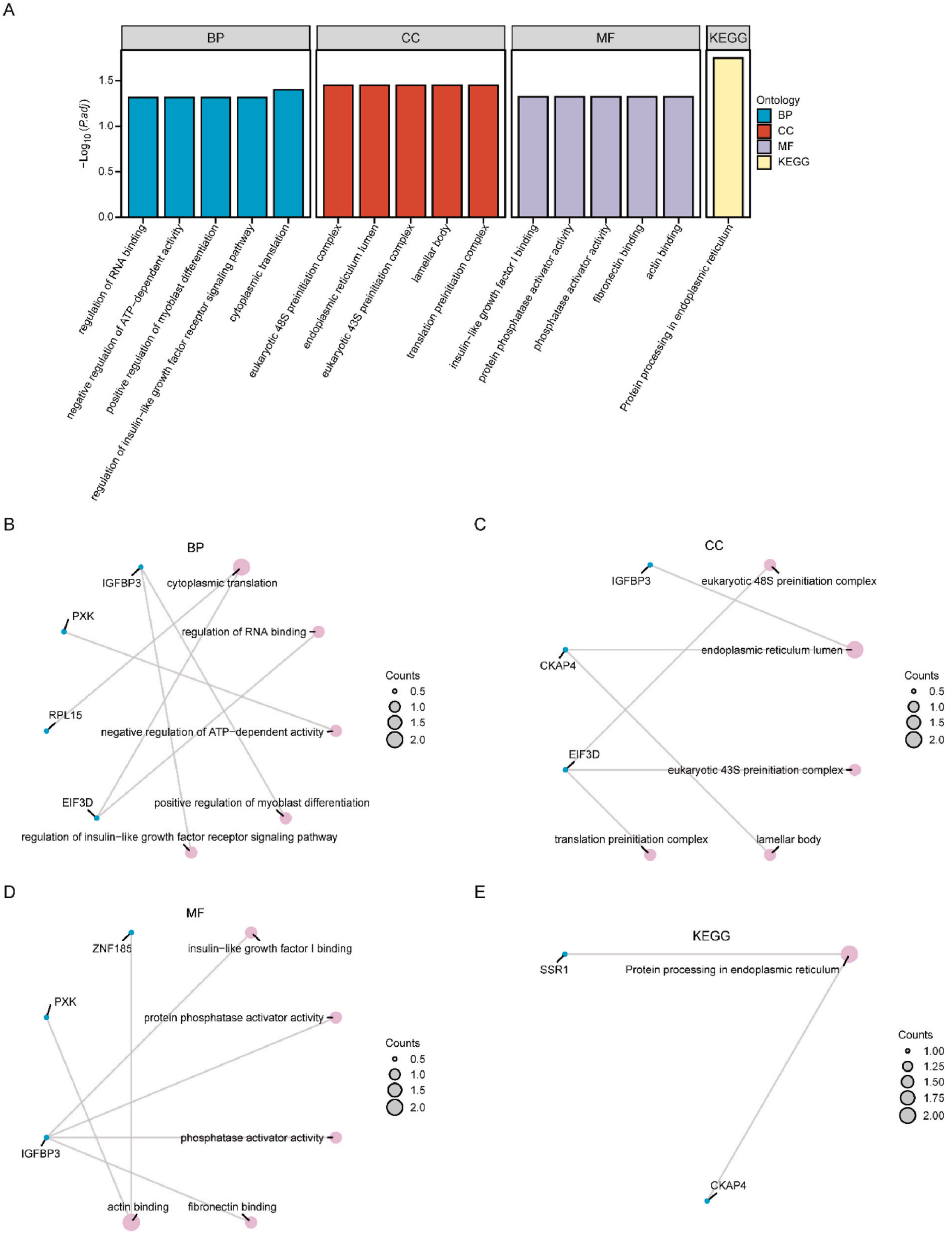


Fig. 6. Verification of diagnostic model.
A. The ROC curve of the GLRDEGs diagnostic model in the combined dataset. B. Nomogram of eight common GLRDEGs in the GLRDEGs logistic regression model. C. Calibration curve of nomogram of GLRDEGs logistic regression model. D. Decision curve (DCA) in GLRDEGs logistic regression model.

Table 3
GO and KEGG enrichment analysis results of GLRDEGs.

ONTOLOGY	ID	Description	pvalue	p.adjust	qvalue
BP	GO:0002181	cytoplasmic translation	0.00088	0.039614	0.020386
BP	GO:1905214	regulation of RNA binding	0.003824	0.048303	0.024858
BP	GO:0032780	negative regulation of ATP-dependent activity	0.007002	0.048303	0.024858
BP	GO:0045663	positive regulation of myoblast differentiation	0.007319	0.048303	0.024858
BP	GO:0043567	regulation of insulin-like growth factor receptor signaling pathway	0.007636	0.048303	0.024858
CC	GO:0033290	eukaryotic 48S preinitiation complex	0.006109	0.035413	0.019281
CC	GO:0005788	endoplasmic reticulum lumen	0.006601	0.035413	0.019281
CC	GO:0016282	eukaryotic 43S preinitiation complex	0.006921	0.035413	0.019281
CC	GO:0042599	lamellar body	0.006921	0.035413	0.019281
CC	GO:0070993	translation preinitiation complex	0.007327	0.035413	0.019281
MF	GO:0031994	insulin-like growth factor I binding	0.005636	0.047455	0.02389
MF	GO:0072542	protein phosphatase activator activity	0.006933	0.047455	0.02389
MF	GO:0019211	phosphatase activator activity	0.009953	0.047455	0.02389
MF	GO:0001968	fibronectin binding	0.013394	0.047455	0.02389
MF	GO:0003779	actin binding	0.014443	0.047455	0.02389
KEGG	has04141	Protein processing in endoplasmic reticulum	0.002546	0.017819	0.013398



(caption on next page)

Fig. 7. Functional enrichment analysis (GO) and pathway enrichment (KEGG) analysis of GLRDEGs.
A. Histogram display of GO enrichment analysis and KEGG pathway enrichment analysis results of GLRDEGs. B-E. The circular network diagram display of BP (B), CC (C), MF (D), and KEGG pathways (E) in the GO functional enrichment analysis of GLRDEGs and the circular network diagram of KEGG enrichment results exhibit (E). The abscissa in the histogram (A) is GO terms, and the height of the bar indicates the Padj value of GO or KEGG terms. The blue dots in the network diagram (B, C, D, E) represent specific genes, and the pink dots represent specific pathways.

combined dataset and marked the positions of the 16 GLRDEGs in the volcano map. A group comparison chart was generated to illustrate the differences in expression of the 16 GLRDEGs between the IDD and control groups (Fig. 3D). The chart revealed significant differential expression of *ASPH*, *IGFBP3*, *SH3PXD2A*, *GLA*, *CKAP4*, *FTL*, *SSR1*, *HDGF*, *CALR*, and *RPN1*, which were markedly upregulated in the IDD group, whereas *EIF3D*, *PXK*, *RPL15*, *RPL22*, *WSB1*, and *ZNF185* were significantly downregulated in the IDD group compared to those in the controls.

3.3. GSEA enrichment analysis between IDD and control groups based on combined dataset

To assess the impact of differential gene expression between IDD and control groups in the combined dataset, we conducted GSEA. We set the significance threshold for enrichment screening with an adjusted p-value (P. adj) and FDR, q. value) below 0.05. Our analysis revealed that in the combined dataset, genes from the different groups (IDD/control) were significantly enriched in pathways such as oxidative phosphorylation, hedgehog ligand biogenesis, and Notch4 signaling, as detailed in Table 2 (Fig. 4B–G). The GSEA results are visually summarized in a mountain plot (Fig. 4A).

3.4. Construction of diagnostic model

To identify 16 GLRDEGs in the combined dataset, we performed logistic regression using the expression levels of these genes and grouping information (IDD/control). A logistic regression model was constructed, selecting genes with $P < 0.05$ as the inclusion criterion. This model encompasses all 16 GLRDEGs. Subsequently, we visualized their expression levels using a Forest Plot (Fig. 5A).

We then constructed the SVM model based on 16 GLRDEGs and the SVM algorithm and obtained the number of genes with the lowest error rate (Fig. 5B) and the highest accuracy rate (Fig. 5C). The results showed that the accuracy of the SVM model was the highest when the number of genes was 10.

We applied the RF algorithm to analyze the expression levels of the 16 GLRDEGs in the combined dataset (Fig. 5D). 'IncNodePurity' was used to indicate the increase in node purity, with higher values signifying fewer impurities (i.e., a lower Gini coefficient). Setting an IncNodePurity threshold >0.5, we identified 16 diagnostic markers from the GLRDEGs using the RF algorithm (Fig. 5E). These markers included CKAP4, *IGFBP3*, *ASPH*, *RPN1*, *PXK*, *SSR1*, *RPL15*, *FTL*, *HDGF*, *EIF3D*, *ZNF185*, *RPL22*, *WSB1*, *SH3PXD2A*, *GLA*, and CALR.

The outcomes of the LASSO regression analysis are depicted in a LASSO regression model diagram (Fig. 5F) and a LASSO variable trajectory diagram (Fig. 5G). This process led to the identification of 10 GLRDEGs in the model: *ASPH*, *CKAP4*, *EIF3D*, *IGFBP3*, *PXK*, *RPL15*, *RPN1*, *SSR1*, *WSB1*, and *ZNF185*.

To identify commonly diagnosed GLRDEGs (common GLRDEGs), we intersected the GLRDEGs from the Logistic-LASSO regression, SVM, and RF models. This intersection revealed eight common GLRDEGs, which are visualized in a Venn diagram (Fig. 5H). These genes include *PXK*, *EIF3D*, *WSB1*, *ZNF185*, *IGFBP3*, *CKAP4*, *RPL15*, and *SSR1*.

Next, we developed a new diagnostic model based on the expression levels of eight common GLRDEGs in the combined dataset in conjunction with the coefficients of these genes in the diagnostic model constructed through LASSO regression analysis.

To further validate the value of the GLRDEGs diagnostic model, we plotted the receiver operating characteristic curve based on the RiskScore of the GLRDEGs diagnostic model and the group information (IDD/control) of the combined dataset (Fig. 6A). As shown in Fig. 6A, the GLRDEGs diagnostic model exhibited high accuracy in diagnosing the two groups, with an area under the curve (AUC) of 0.935.

We also constructed a GLRDEGs logistic regression model and drew a nomogram to show the contribution of eight common GLRDEGs to the GLRDEGs logistic regression model (Fig. 6B). The results showed that the expression levels of *EIF3D*, *PXK*, and *IGFBP3* had significantly higher effects on the GLRDEGs logistic regression model than other variables.

Calibration analysis was performed, and a calibration curve was drawn (Fig. 6C). The optimal theoretical probability (solid line) and model prediction under different conditions were determined. The fit of the probabilities (dashed line) indicated how well the model predicted the outcome (Fig. 6C). In addition, we used DCA to evaluate the role of the GLRDEGs diagnostic model in terms of clinical utility. The results are presented in Fig. 6D. In the DCA diagram, when the line of the model is stably higher than all positive and all negative values within a certain range, the larger the range, the higher the net income and the better the model effect. From the results, we can see that the constructed model has a higher diagnostic value for the occurrence of IDD.

3.5. Functional enrichment analysis (GO) and pathway enrichment (KEGG) analysis

To analyze the relationships of the eight common GLRDEGs, we conducted GO functional enrichment analysis and KEGG pathway enrichment analysis. The results show that the eight common GLRDEGs were significantly enriched in various BPs such as cytoplasmic translation, regulation of RNA binding, negative regulation of ATP-dependent activity, and positive regulation of myoblast

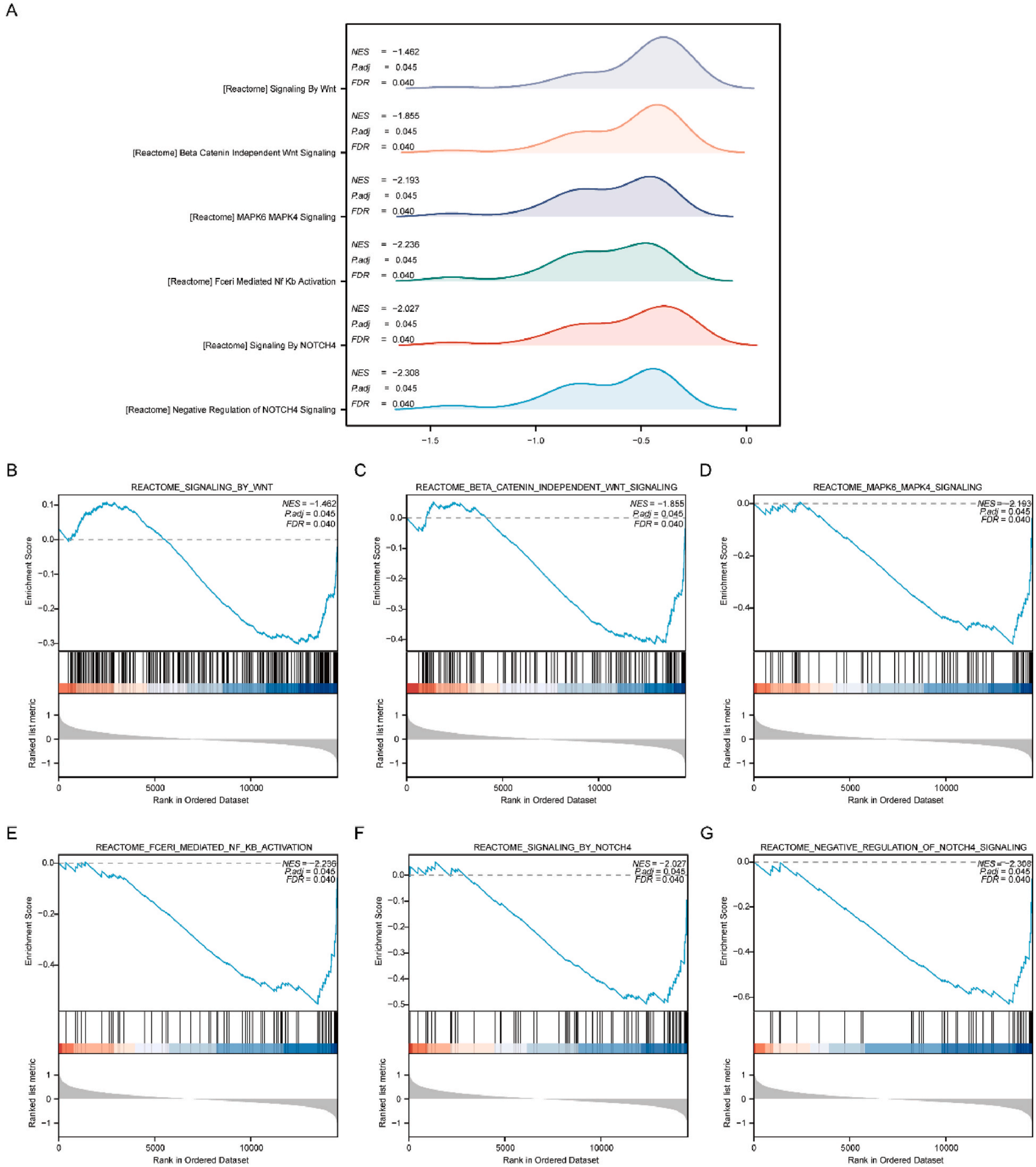


Fig. 8. GSEA enrichment analysis of combined dataset.
A. Combined dataset of different groups (high/low RiskScore group) GSEA of genes among the main six biological characteristics. B-G. REACTOME_SIGNALING_BY_WNT (B), REACTOME_BETA_CATENIN_INDEPENDENT_WNT_SIGNALING (C), REACTOME_MAPK6_MAPK4_SIGNALING (D), REACTOME_FCERI_MEDIATED_NF_KB_ACTIVATION (E), REACTOME_SIGNALING_BY_NOTCH 4 (F), and REACTOME_NEGATIVE_R-EGULATION_OF_NOTCH4_SIGNALING (G). The significant enrichment screening criteria for GSEA were P.adj <0.05 and FDR value (p-value) < 0.05.

differentiation, including the insulin-like growth factor receptor signaling pathway. Additionally, these genes were associated with CCs such as the eukaryotic 48S and 43S pre-initiation complexes, endoplasmic reticulum lumen, lamellar body, and translation pre-initiation complex. In terms of MFs, enrichment was noted for insulin-like growth factor I binding, protein phosphatase activator

Table 4
GSEA enrichment analysis results of Combined dataset Low-High RiskScore group genes.

ID	enrichmentScore	NES	pvalue	p.adjust	qvalue
REACTOME_NEGATIVE_REGULATION_OF_NOTCH4_SIGNALING	−0.63081	−2.30816	0.002075	0.044564	0.040247
REACTOME_SIGNALING_BY_NOTCH4	−0.49719	−2.02722	0.002179	0.044564	0.040247
REACTOME_FCERI_MEDIATED_NF_KB_ACTIVATION	−0.55319	−2.23626	0.002128	0.044564	0.040247
REACTOME_MAPK6_MAPK4_SIGNALING	−0.53807	−2.19342	0.002183	0.044564	0.040247
REACTOME_BETA_CATENIN_INDEPENDENT_WNT_SIGNALING	−0.41536	−1.85468	0.002364	0.044564	0.040247
REACTOME_SIGNALING_BY_WNT	−0.30205	−1.46246	0.002506	0.044564	0.040247

activity, fibronectin binding, and actin binding. Moreover, these genes were implicated in the KEGG pathway for protein processing in the endoplasmic reticulum (see Table 3 for pathway details).

The results of GO functional enrichment analysis and KEGG pathway enrichment analysis are displayed using histograms (Fig. 7A). In addition, we displayed the BP (Fig. 7B), CC (Fig. 7C), MF (Fig. 7D), and KEGG (Fig. 7E) pathway enrichment results of the GO gene functional enrichment analysis in the form of a circular network diagram.

3.6. GSEA enrichment analysis between high- and low-risk groups based on combined dataset

Using the RiskScore from the previously established GLRDEGs diagnostic model, we divided the samples into high- and low-risk groups using the combined dataset. Differential analysis between these groups was conducted using the ‘limma’ package. We assessed the expression of all genes across these groups, focusing on their involvement in BPs, impact on CCs, and exertion of MFs. For significant enrichment, we set stringent criteria with P.adj <0.05 and an FDR (q.value) < 0.05. We presented the results of the GSEA of genes between different groups using a ridge plot (Fig. 8A). The results show that the different groups in the combined dataset data (high/low RiskScore group) were significantly enriched in Wnt signaling (Fig. 8B), beta catenin independent Wnt signaling (Fig. 8C), MAPK6 MAPK4 signaling (Fig. 8D), fceri-mediated NF-kB activation (Fig. 8E), NOTCH4 signaling (Fig. 8F), negative regulation of NOTCH4 signaling (Fig. 8G), and other pathways (See Table 4 for details).

3.7. Construction of mRNA-miRNA, mRNA-TF interaction network

Using the mRNA-miRNA data from the miRDB database, we predicted the miRNAs interacting with the eight identified GLRDEGs, retaining only those interactions that were reported in at least five different sources. Subsequently, we used the Cytoscape software to construct and visualize the mRNA-miRNA interaction network (Fig. 9A). Within this network, mRNAs are represented by red nodes and miRNAs are represented by blue nodes. Analysis of the mRNA-miRNA interaction network revealed that our network comprised seven hub genes (*PXK*, *WSB1*, *ZNF185*, *IGFBP3*, *CKAP4*, *RPL15*, and *SSR1*) and 38 miRNA molecules, collectively forming 42 mRNA-miRNA interaction pairs. Detailed information on these mRNA-miRNA interactions is provided in Table 5.

We identified TFs that bind to eight GLRDEGs (*PXK*, *EIF3D*, *WSB1*, *ZNF185*, *IGFBP3*, *CKAP4*, *RPL15*, and *SSR1*) using the CHIPBase (version 3.0) and hTFtarget databases. We filtered the results with the criteria “Number of samples found (upstream) $>$ 0” and “Number of samples found (downstream) $>$ 0”. Finally, we obtained data for 33 interaction pairs involving four hub genes (*PXK*, *RPL15*, *WSB1*, and *ZNF185*) and 29 TFs. These interactions were visualized using the Cytoscape software (Fig. 9B), where red nodes represent mRNAs and blue nodes represent TFs. Details of the mRNA-TF interactions are provided in Table 6. We also conducted a functional similarity analysis on the eight datasets and presented the results using cloud and rain diagrams. Using the R package GOSemSim, we calculated the semantic similarity between GO terms, sets of GO terms, gene products, and gene clusters. During this process, we retained only those genes that were annotated to pathways in terms of BPs, MFs, and CCs for similarity analysis. Ultimately, we obtained functional similarity analysis results for six GLRDEGs (*PXK*, *EIF3D*, *WSB1*, *IGFBP3*, *RPL15*, and *SSR1*) and visualized them using a raincloud diagram (Fig. 9C). The results demonstrated that *WSB1* had the highest functional similarity value with other GLRDEGs (Fig. 9C; the x-axis represents the similarity score, where a larger value denotes higher functional similarity with other genes).

3.8. Difference analysis of ssGSEA immune signature between high- and low-risk groups

We used the median RiskScore of the GLRDEGs diagnostic model to divide IDD samples from the combined dataset into low- and high-risk groups. To explore the differences in immune infiltration between the high- and low-risk groups, we used the ssGSEA algorithm to calculate the infiltration abundance of 28 types of immune cells in the high- and low-risk group samples. The Mann–Whitney *U* test was used to analyze the degree of difference in the infiltration of 28 types of immune cells between the low- and high-risk groups. The results are displayed in the group comparison chart (Fig. 10A). The results showed that in the combined dataset data, there were statistically significant differences ($P < 0.05$) in the infiltration abundance of four types of immune cells between the low- and high-risk groups: activated dendritic cells, monocytes, type 2 T helper cells, and activated CD8 T cells.

We further calculated the correlation between the infiltration abundances of the four types of immune cells with group statistical differences in the samples of the low- and high-risk groups and displayed the results (Fig. 10B–C). In the low-risk group, the infiltration abundances of the four types of immune cells were mostly positively correlated, among which the relationship between activated CD8

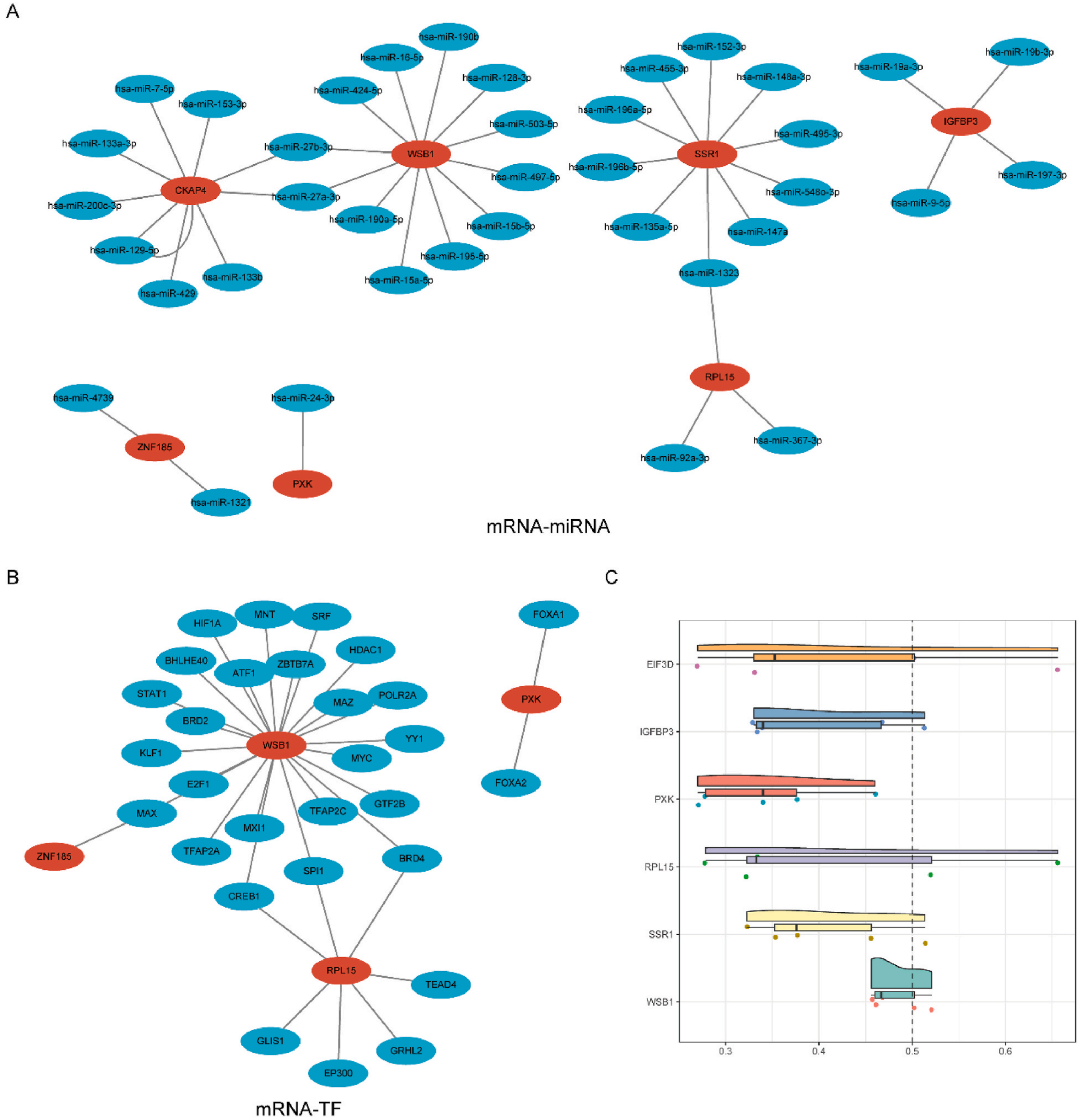


Fig. 9. mRNA-miRNA and mRNA-TF interaction networks and functional similarity analysis diagram of GLRDEGs.
A-D. mRNA-miRNA (A) and mRNA-TF (B) interaction networks of eight GLRDEGs. C. Functional similarity analysis between eight GLRDEGs. The red circles in the mRNA-miRNA (A) interaction network are mRNAs, and the blue circles are miRNAs. (B) The red circles in the mRNA-TF interaction network are mRNAs, and the blue circles are the TFs.

T cells and type 2 T helper cells was the most significant (Fig. 10B), whereas in the high-risk group, except type 2 T helper cells and monocytes, type 2 T helper cells and activated CD8 T cells were negatively correlated, and all the other cells were positively correlated. Among all correlations, the strongest were between type 2 T helper cells and activated CD8 T cells (negative correlation) and activated CD8 T cells and monocytes (positive correlation) (Fig. 10C).

At the same time, we also calculated the infiltration abundance of four types of immune cells and eight GLRDEGs (correlation between the expression levels of *PXK*, *EIF3D*, *WSB1*, *ZNF185*, *IGFBP3*, *CKAP4*, *RPL15*, and *SSR1*) (Fig. 10D–E). Most of the genes were positively correlated, among which the correlation between activated CD8 T cells and *EIF3D* was the strongest (positive correlation). In the high-risk group of the combined dataset data (Fig. 10 E), four types of cells were positively correlated with eight genes and among

Table 5
mRNA-miRNA interaction network nodes.

node1	node2	node1	node2
hsa-miR-27a-3p	CKAP4	hsa-miR-135a-5p	SSR1
hsa-miR-129-5p	CKAP4	hsa-miR-152-3p	SSR1
hsa-miR-129-5p	CKAP4	hsa-miR-196b-5p	SSR1
hsa-miR-7-5p	CKAP4	hsa-miR-495-3p	SSR1
hsa-miR-27b-3p	CKAP4	hsa-miR-455-3p	SSR1
hsa-miR-133a-3p	CKAP4	hsa-miR-1323	SSR1
hsa-miR-153-3p	CKAP4	hsa-miR-548o-3p	SSR1
hsa-miR-200c-3p	CKAP4	hsa-miR-15a-5p	WSB1
hsa-miR-133b	CKAP4	hsa-miR-16-5p	WSB1
hsa-miR-429	CKAP4	hsa-miR-27a-3p	WSB1
hsa-miR-19a-3p	IGFBP3	hsa-miR-15b-5p	WSB1
hsa-miR-19b-3p	IGFBP3	hsa-miR-27b-3p	WSB1
hsa-miR-197-3p	IGFBP3	hsa-miR-128-3p	WSB1
hsa-miR-9-5p	IGFBP3	hsa-miR-190a-5p	WSB1
hsa-miR-24-3p	PXK	hsa-miR-195-5p	WSB1
hsa-miR-92a-3p	RPL15	hsa-miR-424-5p	WSB1
hsa-miR-367-3p	RPL15	hsa-miR-497-5p	WSB1
hsa-miR-1323	RPL15	hsa-miR-503-5p	WSB1
hsa-miR-196a-5p	SSR1	hsa-miR-190b	WSB1
hsa-miR-148a-3p	SSR1	hsa-miR-1321	ZNF185
hsa-miR-147a	SSR1	hsa-miR-4739	ZNF185

Table 6
mRNA-TF interaction network nodes.

mRNA	Transcription factor	mRNA	Transcription factor
PXK	FOXA1	WSB1	HIF1A
PXK	FOXA2	WSB1	KLF1
RPL15	EP300	WSB1	MAX
RPL15	GLIS1	WSB1	MAZ
RPL15	GRHL2	WSB1	MNT
RPL15	BRD4	WSB1	MXI1
RPL15	SPI1	WSB1	MYC
RPL15	TEAD4	WSB1	POLR2A
RPL15	CREB1	WSB1	SPI1
WSB1	ATF1	WSB1	SRF
WSB1	BHLHE40	WSB1	STAT1
WSB1	BRD2	WSB1	TFAP2A
WSB1	BRD4	WSB1	TFAP2C
WSB1	CREB1	WSB1	YY1
WSB1	E2F1	WSB1	ZBTB7A
WSB1	GTF2B	ZNF185	MAX
WSB1	HDAC1		

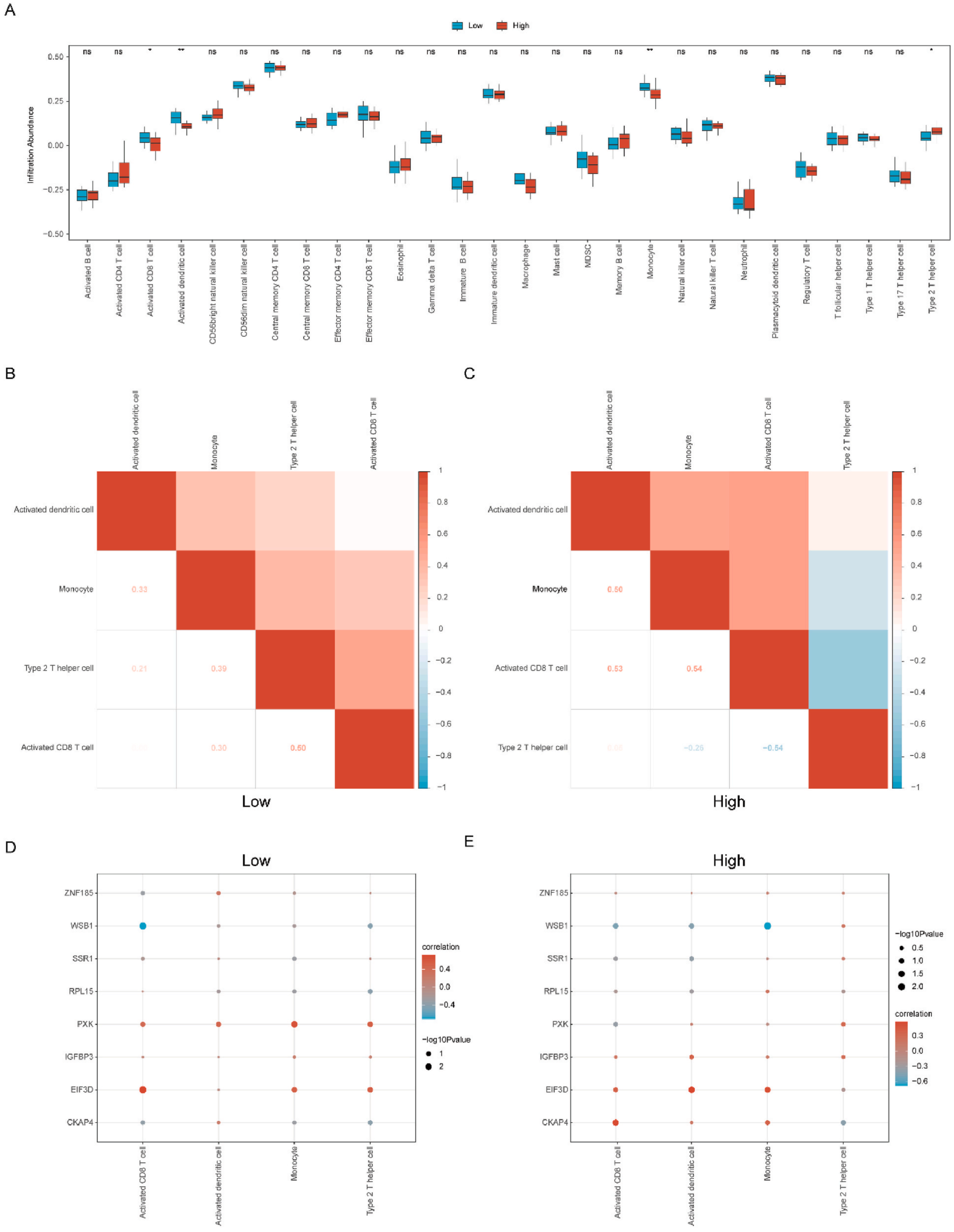
the negative correlations, monocytes showed the strongest correlation (negative correlation) with *WSB1*.

3.9. 3.9CIBERSORT immune signature difference analysis between high-risk group and low-risk group of combined dataset

Subsequently, we employed the CIBERSORT algorithm to calculate the infiltration abundance of 22 types of immune cells in the high- and low-risk groups. A proportional representation of the immune cell infiltration in the dataset samples is depicted in stacked bar charts (Fig. 11A). The results showed that of the 22 types of immune cells, 21 had a non-zero infiltration abundance in the combined dataset samples. These 21 immune cell types are naïve B cells, memory B cells, plasma cells, CD8 T cells, CD4 memory resting T cells, CD4 memory activated T cells, follicular helper T cells, regulatory T cells (Tregs), gamma delta T cells, resting NK cells, activated NK cells, monocytes, M0 macrophages, M1 macrophages, M2 macrophages, resting dendritic cells, activated dendritic cells, resting mast cells, activated mast cells, eosinophils, and neutrophils.

We then used the "spearman" algorithm to calculate the correlation between the infiltration abundance of 21 types of immune cells in the samples of the combined dataset data for high- and low-risk groups (low/high RiskScore group) whose infiltration abundance is not all 0 (Fig. 11B and C). The results showed that in the combined dataset data, the low-risk group (low RiskScore group, infiltration abundance of M2 macrophages in the low-risk group is all 0, leaving 20 types of cells) (Fig. 11C) and the high-risk group (high RiskScore group) (Fig. 11D), the expression levels of immune cells (Fig. 11D) were mostly negatively correlated.

We also calculated the low-risk group (low RiskScore group) (Fig. 11D) and high-risk group (high RiskScore group) (Fig. 11E) in the patient samples of 21 types of immune cell infiltration abundance and 8 GLRDEGs (correlation between the expression levels of *PXK*,



(caption on next page)

Fig. 10. Difference analysis of ssGSEA immune characteristics between different groups (low/high RiskScore groups) of combined dataset data. A. Combined dataset data: high- and low-risk group (low/high RiskScore group). SsGSEA immune infiltration analysis results in group comparison chart display. The results of correlation analysis of immune cell infiltration abundance in the low RiskScore group (B) and high RiskScore group (C) of BC. Combined dataset data are shown. D, E. Combined dataset data of low-risk group (D) and high-risk group (E). Correlation dot plot of immune cells and 27 GLRDEGs. ns is equal to $P \geq 0.05$, which is not statistically significant; $^{*}P < 0.05$, statistically significant; $^{**}P < 0.01$, highly statistically significant; and $^{***}P < 0.001$, very statistically significant.

EIF3D, *WSB1*, *ZNF185*, *IGFBP3*, *CKAP4*, *RPL15*, *SSR1*) (Fig. 11D–E), were screened with $P < 0.05$ as the standard and the results were displayed by the correlation dot plot. The results showed that there were significant correlations between four types of cells (eosinophils, resting NK cells, plasma cells, and focal helper T cells) and five genes (*CKAP4*, *IGFBP3*, *PXK*, *WSB1*, and *ZNF185*) in the low-risk group of combined dataset data (Fig. 11D). Among them, the correlation between resting NK cells and *ZNF185* was the strongest (positive correlation) in the high-risk group of combined dataset data (Fig. 11E). There was a significant correlation between CD4 memory activated T cells, CD8 T cells, focal helper T cells, regulatory T cells, and six genes (*CKAP4*, *EIF3D*, *PXK*, *RPL15*, *SSR1*, and *ZNF185*), among which T cell CD8 and SSR1 showed the strongest correlation (negative correlation).

4. Discussion

IDD is a prevalent cause of LBP, resulting in significant social and economic burdens [1]. The high prevalence of LBP makes it a common reason for seeking medical attention. Considering the increasing incidence of IDD in the aging population, there is an urgent need to uncover its underlying causes and identify effective therapies. Currently, IDD diagnosis primarily relies on symptoms and imaging, which hampers early detection and timely intervention [46]. Therefore, the identification of potential biomarkers for the prediction of IDD is crucial. In recent years, advances in machine learning techniques and the availability of gene expression data in public databases have provided new approaches for identifying biomarkers for disease detection [47]. Despite extensive research efforts, the precise mechanisms underlying IDD remain elusive, and effective treatment options are still lacking. Therefore, our study aimed to explore potential glycolysis-associated gene signatures associated with IDD and investigate their correlation with immune cell infiltration using a comprehensive analytical approach.

In this study, we comprehensively investigated the association among GLRGs, immune cells, and IDD using various machine learning techniques and the CIBERSORT algorithm. Our study established a strong link between these elements, providing a novel perspective for understanding the molecular mechanisms underlying IDD and identifying potential therapeutic targets.

We identified eight GLRDEGs—*PXK*, *EIF3D*, *WSB1*, *ZNF185*, *IGFBP3*, *CKAP4*, *RPL15*, and *SSR1*—by Logistic-LASSO regression, SVM, and RF models. These GLRDEGs not only provide insight into the pathogenesis of IDD but also constitute a promising diagnostic model for IDD. Previous studies have implicated some of these genes in different physiological and pathological contexts, partially confirming our findings.

For instance, IGFBP3, also known as insulin-like growth factor binding protein 3 in yeast, suppresses apoptosis and improves cell survival in several cell systems when calorie intake is restricted [48–50]. IGFBP3, which is essential for cell differentiation, proliferation, apoptosis, and cell aging, deacetylates several apoptosis-associated nonhistone proteins [49,51]. Patients with disc degeneration show higher levels of IGFBP3 expression, which may protect against NP cell apoptosis [20]. EIF3D (eukaryotic translation initiation factor 3, subunit D) is an mRNA cap-binding protein required for specialized translation initiation [52]. EIF3D is associated with tumor development and is widely expressed in a variety of tumor tissues, regulating the cycle of tumor cells and dysregulating apoptotic and anti-apoptotic signaling by increasing apoptosis [52–54]. WSB1 is a member of the SOCS box family and plays a key role in mediating the degradation of substrate proteins via the ubiquitin-proteasome pathway as a central component of the E3 ubiquitin ligase complex of ECS (Elongin B/C-Cullin 2/5-SOCS box protein) [55]. Recent research has indicated that WSB1 regulates thyroid homeostasis, immunological response, glycolysis, and hypoxia, and possibly influences the onset or progression of cancer. It has been shown that WSB1 affects cell invasion, survival, and proliferation [56]. However, further experimental and clinical studies are required to verify whether EIF3D and WSB1 regulate the proliferation of NP cells during IDD.

Enrichment analyses (GO and KEGG) provided deeper insight into the functional roles and pathways related to the identified GLRDEGs. These analyses revealed that the identified genes are significantly involved in the glycolysis pathway, which has been extensively implicated in several disorders including IDD [57–59]. Therefore, it is plausible to assume that modulation of the glycolysis pathway may provide new opportunities for IDD therapeutic strategies.

Our constructed mRNA-miRNA and mRNA-TF interaction networks provide a comprehensive map of the potential regulatory relationships among the identified genes, miRNAs, and TFs. This regulatory network aids in understanding the intricacies of the molecular mechanisms underlying IDD and their contributions to the disease pathology.

Finally, RiskScore based on the diagnostic model divided the samples in the combined dataset into two groups of high/low risk for differential analysis and GSEA, further ssGSEA, and CIBERSORT immune signature differential analysis between the groups. To the best of our knowledge, studies on GLRDEGs and immune responses related to IDD are scarce, and our study examining the cross-talk between hub GLRGs and immune cells may provide new insights into the etiology of IDD.

Considering the physical and biological barriers, such as an avascular microenvironment, high proteoglycan concentration, intense physical pressure, notochordal cells, and apoptosis inducers such as Fas ligand, the IVD is perceived as an immunologically privileged site [16]. Nonetheless, during IDD, immune cells within the IVD have unique properties. Certain immune cells secrete substances that promote angiogenesis and inflammatory responses in the disc, thus playing a vital role in the progression of IDD. Macrophages exhibit

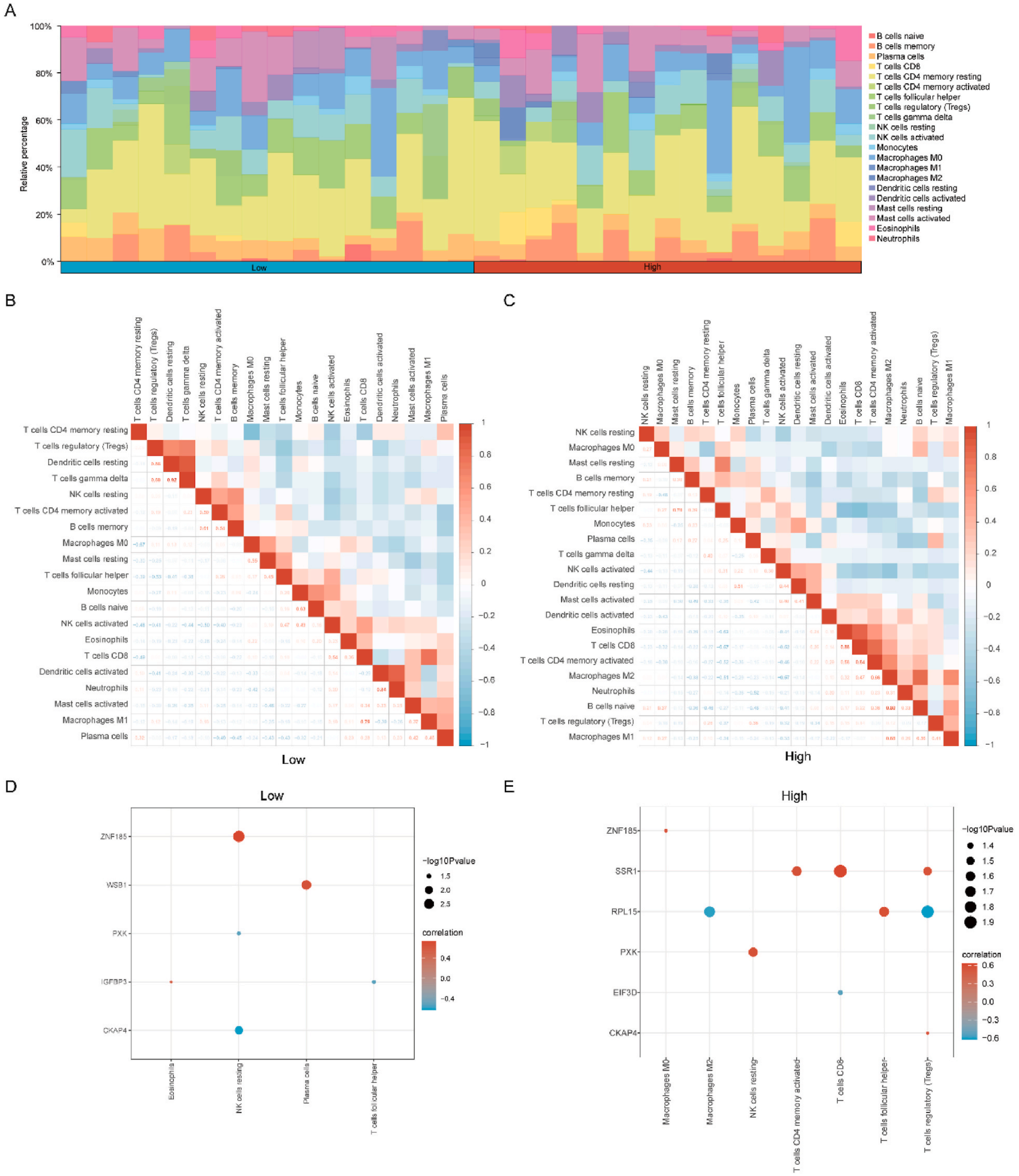


Fig. 11. Difference analysis of CIBERSORT immune signature among different risk groups of combined dataset data.
A. Stacked histogram display of CIBERSORT immune infiltration analysis results between different risk groups of the combined dataset data. B–C. Combined dataset data: low-risk group (B) and high-risk group (C). Correlation heat map between immune cells with infiltrating abundance not 0. D-E. Combined dataset data: low-risk group (D) and high-risk group (E). Correlation dot plot of immune cells with infiltration abundance not zero and GLRDEGs. ns is equal to $P \geq 0.05$, which is not statistically significant; *$P < 0.05$, statistically significant; **$P < 0.01$, highly statistically significant; and ***$P < 0.001$, very statistically significant.

a higher pro-inflammatory profile in degenerate IVDs, as evidenced by the elevated expression of pro-inflammatory markers (such as CCR7, IL-6, CD86, and iNOS) [60]. A novel lineage of Th17 lymphocytes expressing substantial amounts of IL-17 has been discovered in surgical tissues obtained from patients with deteriorated IVDs [61]. IL-17 not only promotes chemokine synthesis, but also orchestrates the migration of neutrophils and monocytes to the inflammation site, thereby instigating chemokine production and attracting immune cells [62].

However, this study has some limitations. First, the database used had a very small sample size, which calls for additional research using larger samples. Second, there were insufficient experiments to validate these claims. Therefore, extensive biological research on immune cell infiltration and glycolysis is necessary.

5. Conclusions

Our study provides novel insights into the role of glycolysis and the immune response in IDD. The identified GLRGs and immune cells can potentially serve as biomarkers or therapeutic targets for IDD. Despite the need for further experimental validation, these findings pave the way for future research on the molecular mechanisms underlying IDD and have potential implications for improving diagnostic accuracy and developing new therapeutic strategies.

Funding statement

This work was supported by the grant from General Project of Natural Science Research Project of Shanxi Provincial Health Commission (No. 2020TD13) and Hepcidin expression changes and related effect mechanisms in intervertebral disc nucleus pulposus degeneration (No. 202203021211074).

Data availability statement

The datasets presented in this study can be found in the GEO database including GSE70362, GSE23130 and GSE15227, which is global and public online repositories.

CRediT authorship contribution statement

Jian Gao: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Liming He:** Funding acquisition. **Jianguo Zhang:** Formal analysis, Data curation. **Leimin Xi:** Project administration, Methodology. **Haoyu Feng:** Visualization, Validation, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36158.

References

[1] N.N. Knezevic, K.D. Candido, J.W.S. Vlaeyen, J. Van Zundert, and S.P. Cohen, Low back pain. Lancet 398 78-92.
[2] R. Chou, Low back pain. Ann. Intern. Med. 174 ITC113-ITC128.
[3] T. Oichi, Y. Taniguchi, Y. Oshima, S. Tanaka, T. Saito, Pathomechanism of intervertebral disc degeneration, JOR spine 3 (2020) e1076.
[4] J. Xin, Y. Wang, Z. Zheng, S. Wang, S. Na, S. Zhang, Treatment of intervertebral disc degeneration, Orthop. Surg. 14 (2022) 1271–1280.
[5] I.L. Mohd Isa, S.L. Teoh, N.H. Mohd Nor, S.A. Mokhtar, Discogenic low back pain: anatomy, pathophysiology and treatments of intervertebral disc degeneration, Int. J. Mol. Sci. 24 (2022) 208.
[6] V. Francisco, J. Pino, M.Á. González-Gay, F. Lago, J. Karppinen, O. Tervonen, A. Mobasheri, and O. Gualillo, A new immunometabolic perspective of intervertebral disc degeneration. Nat. Rev. Rheumatol. 18 47-60.
[7] N.S. Chandel, Glycolysis. Cold Spring Harbor Perspect. Biol. 13 a040535.
[8] Z. Abbaszadeh, S. Çeşmeli, Ç.B. Avcı, Crucial players in glycolysis: cancer progress, Gene 726 (2020) 144158.
[9] C. Tan, L. Li, J. Han, K. Xu, X. Liu, A new strategy for osteoarthritis therapy: inhibition of glycolysis, Front. Pharmacol. 13 (2022) 1057229.
[10] J. Zuo, J. Tang, M. Lu, Z. Zhou, Y. Li, H. Tian, E. Liu, B. Gao, T. Liu, and P. Shao, Glycolysis rate-limiting enzymes: novel potential regulators of rheuma toid arthritis pathogenesis. Front. Immunol. 12 779787.
[11] W.-C. Lee, A.R. Guntur, F. Long, and C.J. Rosen, Energy metabolism of the osteoblast: implications for osteoporosis. Endocr. Rev. 38 255-266.
[12] L. Wu, J. Shen, X. Zhang, Z. Hu, LDHA-mediated glycolytic metabolism in nucleus pulposus cells is a potential therapeutic target for intervertebral disc degeneration, BioMed Res. Int. 2021 (2021).
[13] J.-W. Kim, N. Jeon, D.-E. Shin, S.-Y. Lee, M. Kim, D.H. Han, J.Y. Shin, and S. Lee, Regeneration in spinal disease: therapeutic role of hypoxia-inducible factor-1 alpha in regeneration of degenerative intervertebral disc. Int. J. Mol. Sci. 22 5281.
[14] Y. Song, S. Lu, W. Geng, X. Feng, R. Luo, G. Li, and C. Yang, Mitochondrial quality control in intervertebral disc degeneration. Exp. Mol. Med. 53 1124-1133.

[15] S.N. Johnston, E.S. Silagi, V. Madhu, D.H. Nguyen, I.M. Shapiro, and M.V. Risbud, GLUT1 is redundant in hypoxic and glycolytic nucleus pulposus cells of the intervertebral disc. JCI Insight 8 e164883.

[16] Z. Sun, B. Liu, Z.-J. Luo, The immune privilege of the intervertebral disc: implications for intervertebral disc degeneration treatment, Int. J. Med. Sci. 17 (2020) 685.

[17] Z. Ling, Y. Liu, Z. Wang, Z. Zhang, B. Chen, J. Yang, B. Zeng, Y. Gao, C. Jiang, Y. Huang, X. Zou, X. Wang, and F. Wei, Single-cell RNA-seq analysis reveals macrophage involved in the progre ssion of human intervertebral disc degeneration. Front. Cell Dev. Biol. 9 833420.

[18] H. Kedong, D. Wang, M. Sagaram, H.S. An, and A. Chee, Anti-inflammatory effects of interleukin-4 on intervertebral disc cell s. Spine J. 20 60-68.

[19] L. Wang, T. He, J. Liu, J. Tai, B. Wang, L. Zhang, and Z. Quan, Revealing the immune infiltration landscape and identifying diagnostic biomarkers for lumbar disc herniation. Front. Immunol. 12 666355.

[20] Z. Kazezian, R. Gawri, L. Haglund, J. Ouellet, F. Mwale, F. Tarrant, P. O'Gaora, A. Pandit, M. Alini, S. Grad, Gene expression profiling identifies interferon signalling molecules and IGFBP3 in human degenerative annulus fibrosus, Sci. Rep. 5 (2015) 1–13.

[21] H.E. Gruber, G.L. Hoelscher, J.A. Ingram, and E.N. Hanley Jr, Genome-wide analysis of pain-, nerve-and neurotrophin-related gene expression in the degenerating human annulus. Mol. Pain 8 (2012) 1744-8069-8-63.

[22] H.E. Gruber, G. Hoelscher, B. Loeffler, Y. Chow, J.A. Ingram, W. Halligan, E.N. Hanley Jr., Prostaglandin E1 and misoprostol increase epidermal growth factor production in 3D-cultured human annulus cells, Spine J. 9 (2009) 760–766.

[23] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, R. Edgar, NCBI GEO: mining tens of millions of expression profiles—database and tools update, Nucleic Acids Res. 35 (2007) D760–D765.

[24] S. Davis, P.S. Meltzer, GEOquery: a bridge between the gene expression Omnibus (GEO) and BioConductor, Bioinformatics 23 (2007) 1846–1847.

[25] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T.I. Stein, R. Nudel, I. Lieder, and Y. Mazor, The GeneCards suite: from gene data mining to disease genome sequence analyses. Current protocols in bioinformatics 54 (2016) 1.30. 1-1.30. 33.

[26] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (2012) 882–883.

[27] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015) e47.

[28] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (2005) 15545–15550.

[29] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. Mesirov, Molecular signatures database (MSigDB) 3.0, Bioinformatics 27 (12) (2011 Jun 15) 1739–1740 [Internet].

[30] H. Sanz, C. Valim, E. Vegas, J.M. Oller, F. Reverter, SVM-RFE: selection and visualization of the most relevant features through non-linear kernels, BMC Bioinf. 19 (2018) 1–18.

[31] Y. Liu, H. Zhao, Variable importance-weighted random forests, Quantitative Biology 5 (2017) 338–351.

[32] S. Engebretsen, J. Bohlin, Statistical predictions with glmnet, Clin. Epigenet. 11 (1) (2019) 123.

[33] W. Cai, M. van der Laan, Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (LASSO) estimator, Int. J. Biostat. 16 (2020).

[34] S.Y. Park, Nomogram: an analogue tool to deliver digital knowledge, J. Thorac. Cardiovasc. Surg. 155 (2018) 1793.

[35] T. Tataranni, C. Piccoli, Dichloroacetate (DCA) and cancer: an overview towards clinical applications, Oxid. Med. Cell. Longev. 2019 (2019).

[36] G. Yu, Gene ontology semantic similarity analysis using GOSemSim, Stem Cell Transcriptional Networks: Methods and Protocols (2020) 207–215.

[37] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.

[38] G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS A J. Integr. Biol. 16 (2012) 284–287.

[39] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, J.-H. Yang, starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data, Nucleic Acids Res. 42 (2014) D92–D97.

[40] J. Huang, W. Zheng, P. Zhang, Q. Lin, Z. Chen, J. Xuan, C. Liu, D. Wu, Q. Huang, L. Zheng, ChIPBase v3. 0: the encyclopedia of transcriptional regulations of non-coding RNAs and protein-coding genes, Nucleic Acids Res. 51 (2023) D46–D56.

[41] Q. Zhang, W. Liu, H.-M. Zhang, G.-Y. Xie, Y.-R. Miao, M. Xia, A.-Y. Guo, hTFtarget: a comprehensive database for regulations of human transcription factors and their targets, Dev. Reprod. Biol. 18 (2020) 120–128.

[42] P. Charoentong, F. Finotello, M. Angelova, C. Mayer, M. Efremova, D. Rieder, H. Hackl, Z. Trajanoski, Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade, Cell Rep. 18 (2017) 248–262.

[43] D.A. Barbie, P. Tamayo, J.S. Boehm, S.Y. Kim, S.E. Moody, I.F. Dunn, A.C. Schinzel, P. Sandy, E. Meylan, C. Scholl, Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1, Nature 462 (2009) 108–112.

[44] S. Hänzelmann, R. Castelo, J. Guinney, GSVA: gene set variation analysis for microarray and RNA-seq data, BMC Bioinf. 14 (2013) 1–15.

[45] B. Chen, M.S. Khodadoust, C.L. Liu, A.M. Newman, A.A. Alizadeh, Profiling tumor infiltrating immune cells with CIBERSORT, Cancer Systems Biology: Methods and Protocols (2018) 243–259.

[46] A. Kamali, R. Ziadlou, G. Lang, J. Pfannkuche, S. Cui, Z. Li, R.G. Richards, M. Alini, and S. Grad, Small molecule-based treatment approaches for intervertebral disc dege neration: current options and future directions. Theranostics 11 27-47.

[47] O. Aromolaran, D. Aromolaran, I. Isewon, and J. Oyelade, Machine learning approach to gene essentiality prediction: a review. Brief Bioinform 22 bbab128.

[48] F. D'Addio, A. Maestroni, E. Assi, M. Ben Nasr, G. Amabile, V. Usuelli, C. Loretelli, F. Bertuzzi, B. Antonioli, F. Cardarelli, B. El Essawy, A. Solini, I.C. Gerling, C. Bianchi, G. Becchi, S. Mazzucchelli, D. Corradi, G.P. Fadini, D. Foschi, J.F. Markmann, E. Orsi, J. Škrha, Jr., M.G. Camboni, R. Abdi, A.M. James Shapiro, F. Folli, J. Ludvigsson, S. Del Prato, G. Zuccotti, and P. Fiorina, The IGFBP3/TMEM219 pathway regulates beta cell homeostasis. Nat. Commun. 13 684.

[49] D. Hu, Y. Ge, Y. Cui, K. Li, J. Chen, C. Zhang, Q. Liu, L. He, W. Chen, J. Chen, C. Hu, H. Xiao, Upregulated IGFBP3 with aging is involved in modulating apoptosis, oxi dative stress, and fibrosis: a target of age-related erectile dysfunct ion, Oxid. Med. Cell. Longev. (2022) 6831779.

[50] M. Li, W. Wu, S. Deng, Z. Shao, and X. Jin, TRAIP modulates the IGFBP3/AKT pathway to enhance the invasion and pro liferation of osteosarcoma by promoting KANK1 degradation. Cell Death Dis. 12 767.

[51] Y. Liu, H. Lv, X. Li, J. Liu, S. Chen, Y. Chen, Y. Jin, R. An, S. Yu, and Z. Wang, Cyclovirobuxine inhibits the progression of clear cell renal cell carc inoma by suppressing the IGFBP3-AKT/STAT3/MAPK-Snail signalling pathwa y. Int. J. Biol. Sci. 17 3522-3537.

[52] A.M. Lamper, R.H. Fleming, K.M. Ladd, and A.S.Y. Lee, A phosphorylation-regulated eIF3d translation switch mediates cellular adaptation to metabolic stress. Science 370 853-856.

[53] H. Huang, Y. Gao, A. Liu, X. Yang, F. Huang, L. Xu, X. Danfeng, and L. Chen, EIF3D promotes sunitinib resistance of renal cell carcinoma by interac ting with GRP78 and inhibiting its degradation. EBioMedicine 49 189-201.

[54] C. Li, K. Lu, C. Yang, W. Du, and Z. Liang, EIF3D promotes resistance to 5-fluorouracil in colorectal cancer throu gh upregulating RUVBL1. J. Clin. Lab. Anal. 37 e24825.

[55] M. Haque, J.K. Kendal, R.M. MacIsaac, and D.J. Demetrick, WSB1: from homeostasis to hypoxia. J. Biomed. Sci. 23 61.

[56] J.J. Kim, S.B. Lee, J. Jang, S.-Y. Yi, S.-H. Kim, S.-A. Han, J.-M. Lee, S.-Y. Tong, N.D. Vincelette, B. Gao, P. Yin, D. Evans, D.W. Choi, B. Qin, T. Liu, H. Zhang, M. Deng, J. Jen, J. Zhang, L. Wang, and Z. Lou, WSB1 promotes tumor metastasis by inducing pVHL degradation. Genes Dev. 29 2244-2257.

[57] L. Zhang, Z. Zhang, and Z. Yu, Identification of a novel glycolysis-related gene signature for predic ting metastasis and survival in patients with lung adenocarcinoma. J. Transl. Med. 17 423.

[58] Q. Xu, D. Miao, X. Song, Z. Chen, L. Zeng, L. Zhao, J. Xu, Z. Lin, and F. Yu, Glycolysis-related gene signature can predict survival and immune stat us of hepatocellular carcinoma. Ann. Surg Oncol. 29 3963-3976.

[59] Z. Liu, Z. Liu, X. Zhou, Y. Lu, Y. Yao, W. Wang, S. Lu, B. Wang, F. Li, and W. Fu, A glycolysis-related two-gene risk model that can effectively predict the prognosis of patients with rectal cancer. Hum Genomics 16 5.

[60] X.-C. Li, S.-J. Luo, W. Fan, T.-L. Zhou, D.-Q. Tan, R.-X. Tan, Q.-Z. Xian, J. Li, C.-M. Huang, and M.-S. Wang, Macrophage polarization regulates intervertebral disc degeneration by modulating cell proliferation, inflammation mediator secretion, and ex tracellular matrix metabolism. Front. Immunol. 13 922173.

[61] L. Cheng, W. Fan, B. Liu, X. Wang, and L. Nie, Th17 lymphocyte levels are higher in patients with ruptured than non-r uptured lumbar discs, and are correlated with pain intensity. Injury 44 1805-1810.

[62] W. Li, P. Chen, Y. Zhao, M. Cao, W. Hu, L. Pan, H. Sun, D. Huang, H. Wu, Z. Song, H. Zhong, L. Mou, S. Luan, X. Chen, and H. Gao, Human IL-17 and TNF-α additively or synergistically regulate the expre ssion of proinflammatory genes, coagulation-related genes, and tight J unction genes in porcine aortic endothelial cells. Front. Immunol. 13 857311.