scientific reports

Check for updates

OPEN

Unraveling shared diagnostic genes and cellular microenvironmental changes in endometriosis and recurrent implantation failure through multi-omics analysis

Dongxu Qin[1,3]✉, Yongquan Zheng[1,3]✉, Libo Wang[2,3], Zhenyi Lin[1], Yao Yao[1], Weidong Fei[1] & Caihong Zheng[1]

Endometriosis and Recurrent Implantation Failure (RIF) are both pivotal clinical issues within the realm of reproductive medicine, sharing significant overlap in their pathophysiological mechanisms. However, research exploring the commonalities between these two conditions remains relatively scarce, and reliable shared diagnostic biomarkers have yet to be identified. In this study, we integrated transcriptomic and single-cell sequencing data from the Gene Expression Omnibus (GEO) database to identify shared diagnostic genes and alterations in the cellular microenvironment between EMs and RIF. Differential expression analysis and weighted gene co-expression network analysis (WGCNA) were employed to identify key genes. Machine learning algorithms, including Random Forest (RF) and XGBoost, were utilized to screen for shared diagnostic genes, which were subsequently validated through receiver operating characteristic (ROC) analysis and clinical prediction models. Single-cell analysis was conducted to investigate the expression patterns of these diagnostic genes across various cellular subpopulations. Additionally, gene set enrichment analysis (GSEA) and competing endogenous RNA (ceRNA) network analysis were employed to further elucidate the biological functions and regulatory mechanisms of these genes. A total of 16 key genes were identified, which were predominantly expressed in fibroblasts. Through machine learning, the optimal model combining RF and XGBoost was selected to identify the shared diagnostic genes PDIA4 and PGBD5. Single-cell analysis revealed significant differences in the expression of these diagnostic genes in fibroblasts between normal and disease states. ROC analysis showed that the Area Under the Curve (AUC) values for individual genes in disease diagnosis were all above 0.7. The constructed clinical prediction model demonstrated robust predictive capacity for the disease. Immune infiltration analysis indicated that M2 macrophages and γδ T cells play important roles in the pathogenesis of EMs and RIF. GSEA revealed that these genes are involved in immune responses, vascular function, and hormone regulation, and are regulated by miR-3121-3p. This study provides comprehensive insights into the shared cellular microenvironmental alterations and molecular mechanisms underlying EMs and RIF. The identification of PDIA4 and PGBD5 as shared diagnostic biomarkers offers new avenues for early diagnosis and targeted treatment of EMs-related RIF. Future work will focus on validating these findings in larger cohorts and exploring their therapeutic potential.

Keywords Endometriosis, Recurrent implantation failure, Machine learning, PDIA4, PGBD5

Recurrent Implantation Failure (RIF) refers to the condition where a woman fails to achieve clinical pregnancy after undergoing at least three cycles of high-quality embryo transfer during In Vitro Fertilization (IVF)[1]. RIF

[1]Department of Pharmacy, Women's Hospital, Zhejiang University School of Medicine, Hangzhou 310006, China. [2]Department of pharmacy, Affiliated Xianju's Hospital, XianJu People's Hospital, Zhejiang Southeast Campus of Zhejiang Provincial People's Hospital, Hangzhou Medical College, Xianju 317300, Zhejiang, China. [3]Dongxu Qin, Yongquan Zheng and Libo Wang: These authors have contributed equally to this work and share first authorship. ✉email: qdxyxt@zju.edu.cn; 5515058@zju.edu.cn

occurs in approximately 10% of IVF patients globally, imposing significant economic and psychological burdens on patients[2,3]. Endometriosis (EMs) is one of the significant causes of RIF, affecting the endometrial receptivity and the embryo implantation process, leading to infertility and pregnancy failure[4]. However, there is still a lack of in-depth research on diagnostic markers and therapeutic targets related to EMs-associated RIF, which limits the effectiveness of clinical interventions. Therefore, identifying molecular markers associated with EMs-related RIF is of great significance for improving the success rate of IVF and improving patient prognosis.

Although studies have revealed the association between EMs and RIF[5,6], current diagnostic methods still have limitations. Traditional diagnosis relies on single markers or imaging examinations, which fail to comprehensively reflect the complexity of the disease. Moreover, the etiology of RIF is complex, involving multiple factors such as immune abnormalities[7], poor endometrial receptivity[8], and endocrine disorders[9], posing challenges to precise diagnosis and individualized treatment. For example, preimplantation genetic testing (PGT) can identify chromosomal abnormalities in embryos, but not all RIF patients have such issues[10]. Therefore, developing new diagnostic tools to identify the risk of EMs-related RIF at an early stage is an urgent problem to be solved.

Gene markers play an indispensable role in the prevention and diagnosis of diseases, offering novel perspectives for understanding the molecular mechanisms underlying these conditions and enabling the development of targeted therapeutics. For instance, in the early risk screening for hereditary breast and ovarian cancers, assessing the expression levels of BRCA genes has been shown to significantly improve clinical outcomes[11]. Consequently, identifying molecular diagnostic markers associated with EMs-related RIF may allow for the early prediction of IVF outcomes in EMs patients and inform targeted treatment strategies, thereby substantially reducing the incidence of RIF[12].

Traditional single data analysis does not effectively reveal the commonalities among diseases. This study aims to identify diagnostic genes associated with EMs-related RIF through machine learning and single-cell analysis technologies, and to construct a clinical prediction model. By integrating transcriptomic and single-cell sequencing data, combined with differential analysis, weighted gene co-expression network analysis (WGCNA), and various machine learning algorithms, we successfully screened out PDIA4 and PGBD5 as potential diagnostic markers. This research objective not only helps in the early prediction of IVF outcomes for EMs patients but also provides a precise diagnostic tool for the clinic, thereby significantly reducing the incidence of RIF. Furthermore, the significance of this study lies in filling the gap in the research of molecular markers related to EMs-associated RIF, providing new diagnostic and therapeutic targets for the clinic. By early identification of the RIF risk in EMs patients, treatment plans can be optimized, unnecessary interventions reduced, and the success rate of IVF improved.

Methods

Date collection and preparation

The Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) was utilized to identify pertinent datasets for disease research. The inclusion criteria for the datasets are as follows: (1) Each dataset must include data from both the normal and diseased groups, ensuring that there is no confounding factors present in either set; (2) The samples in the datasets must be derived from human subjects. Table 1 presents a summary of all the included datasets. For the EMs cohort, datasets GSE23339 and GSE11691were subjected to standardization and merging to constitute the training set, with GSE7305 earmarked for validation purposes. Following a similar protocol, datasets GSE111974 and GSE26787 were standardized and combined to form the RIF training set, with GSE106602 designated as the validation set. Utilizing the platform annotation file as a foundation, we conducted the transformation of gene symbols within the Perl programming framework. Subsequently, the expression matrices of genes with congruent symbols were averaged, and genes with zero expression across all samples were eliminated. Within the R computational framework, the "limma" script (normalize Between Arrays function) was applied to normalize the raw data from diverse datasets[13]. Subsequently, the training set was processed for batch correction using the "sva" package to consolidate the distinct datasets[14]. After the normalization and consolidation steps, stringent filtering criteria were implemented to identify differentially expressed genes (DEGs) between the healthy and disease groups. Specifically, genes with a fold change (|FC|) of ≥ 1.5 and an adjusted P-value < 0.05 were deemed as DEGs[15,16]. The overall workflow is illustrated in Fig. 1.

Functional enrichment analysis

In order to identify the biological functions and signalling pathways of genes, we employed the "Cluster Finder" package for gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis[17,18]. The resulting bar graphs demonstrate significant functional enrichment results at adjust $P < 0.05$.

Weighted gene co-expression network analysis (WGCNA)

we utilized the "WGCNA" R package to analyze gene co-expression networks, following a series of specific parameters and steps. For data preprocessing, we employed the goodSamplesGenes function to identify and remove outlier samples and genes, ensuring the integrity of our dataset by excluding those with missing data. We then performed hierarchical clustering using the hclust function with the average linkage method to visualize the relationships among samples and to identify potential outliers, based on the Euclidean distance matrix of the expression data. For the selection of the soft threshold, we used the pickSoftThreshold function to determine the optimal soft-power threshold. We tested a range of powers from 1 to 30 and selected the one that best fit the scale-free topology model, guided by the highest R-squared value and the lowest mean connectivity. In constructing the network, we first built the adjacency matrix using the chosen soft-power threshold. This adjacency matrix was subsequently converted into a topological overlap matrix (TOM) using the TOMsimilarity function. We then conducted hierarchical clustering on the TOM-based dissimilarity matrix using the hclust function with the average linkage method. For module detection, we applied the cutreeDynamic function with

Disease	ID	Platform	Type	Sample
Endometriosis	GSE11691	GPL96	Bulk RNA-Seq	9 Control vs. 9 Disease
	GSE23339	GPL6102	Bulk RNA-Seq	9 Control vs. 10 Disease
	GSE7305	GPL570	Bulk RNA-Seq	10 Control vs. 10 Disease
	GSE214411	GPL24676	ScRNA-Seq	6 Control vs. 4 Disease
Recurrent Implantation Failure	GSE111974	GPL17077	Bulk RNA-Seq	24 Control vs. 24 Disease
	GSE26787	GPL570	Bulk RNA-Seq	5 Control vs. 5 Disease
	GSE106602	GPL16791	Bulk RNA-Seq	16 Control vs. 19 Disease
	GSE183837	GPL24676	ScRNA-Seq	3 Control vs. 6 Disease

Table 1. Summary of dataset information downloaded from GEO.

the dynamic tree-cut algorithm to identify distinct gene modules. We set the minimum module size to 60 and adjusted the deep split parameter to 2 to enhance the sensitivity of module detection. Modules with similar characteristics were merged based on a defined cut height of 0.25, utilizing the mergeCloseModules function to streamline the modules. To calculate the module eigengenes (MEs), we used the moduleEigengenes function. The MEs, representing the first principal component of each module, were crucial for summarizing the expression patterns of the genes within each module. Finally, to assess the correlation with clinical traits, we used Pearson's correlation coefficient to evaluate the relationship between the module eigengenes and clinical traits. This step was essential in identifying modules that were significantly associated with the clinical phenotypes of interest.

Identification of key differential genes

The differentially expressed genes that exhibit the same trend between EMs and RIF were intersected with the key module genes identified by WGCNA to obtain the critical differential genes. The chromosomal locations of these critical differential genes were extracted using Perl, and their positional expressions were visualized using the "RCircos" package.

Single cell sequencing analysis

The single-cell datasets GSE214411 (EMs) and GSE183837 (RIF) were downloaded from the GEO database for the analysis of key cellular subpopulations. In the initial phase of data processing, the Seurat package (version 4.3.0) was employed to load and preprocess the raw data[19,20], thereby constructing a Seurat object. Subsequently, a stringent quality control (QC) procedure was implemented. Cells with a minimum of 200 expressed genes, and genes expressed in at least three cells, were retained. Additionally, the proportion of mitochondrial genes (prefixed with "MT-") and ribosomal protein genes (prefixed with "RP") in each cell was calculated using the PercentageFeatureSet function to determine percent.mt and percent.rb, respectively. The final QC criteria for cell selection were as follows: total UMI count (nCount_RNA) $\geq$ 1000, gene count (nFeature_RNA) between 200 and 10,000, mitochondrial gene proportion (percent.mt) $\leq$ 20%, and ribosomal protein gene proportion (percent.rb) $\leq$ 20%. Following QC, the data were normalized using the NormalizeData function with the LogNormalize method and scaled to a magnitude of 10,000[21]. To identify informative features for subsequent analysis, the FindVariableFeatures function was utilized with the vst method to select highly variable genes, resulting in the selection of 3,000 such genes. For dimensionality reduction, principal component analysis (PCA) was performed using the RunPCA function, and the first 20 principal components were chosen for further analysis. To correct for batch effects, the harmony package (version 1.2.0) was employed, with the RunHarmony function using sample type (Type) as a covariate[22]. The corrected data were then utilized for clustering and visualization. Based on the reduced-dimensional data, cell clustering was conducted using the FindNeighbors and FindClusters functions, with the resolution parameter adjusted to delineate distinct clusters. The SingleR package was employed to annotate cell types by comparing the clustered data with the Human Primary Cell Atlas Data, using the SingleR function to predict the cell type of each cluster, which was subsequently used as the annotation. Furthermore, the DimPlot function was used to generate a UMAP plot to visualize the clustering results and cell type distribution. To investigate the differential expression of key genes among different cell types or states, the FindMarkers function was employed with an adjusted P-value threshold of < 0.05 to identify differentially expressed genes.

Machine Learning-assisted diagnostic gene selection

The best classification model is constructed and screened by systematic machine learning methods. Firstly, the common genes in training and test data are screened out and the data are standardized. Subsequently, a variety of feature selection methods were adopted, and different machine learning algorithms were called through the RunML function, including elastic network (Enet), Lasso regression, ridge regression, stepwise logistic regression (Stepglm), support vector machine (SVM), linear discriminant analysis (LDA), lifting method (glmBoost), partial least squares regression generalized linear model (plsRglm), random forest (RF), XGBoost and Naive Bayes (NaiveBayes). In the model training stage, according to the results of feature selection, the corresponding machine learning algorithm is used to construct the model. For example, for the elastic network model, the optimal regularization parameter (alpha) is determined by cross-validation, and the model training is performed using the glmnet package; for the random forest model, use the randomForestSRC package and optimize parameters such as node size. The training process of each model includes the fitting of training data and the evaluation of the importance of features. The AUC values of each model on the test data are calculated

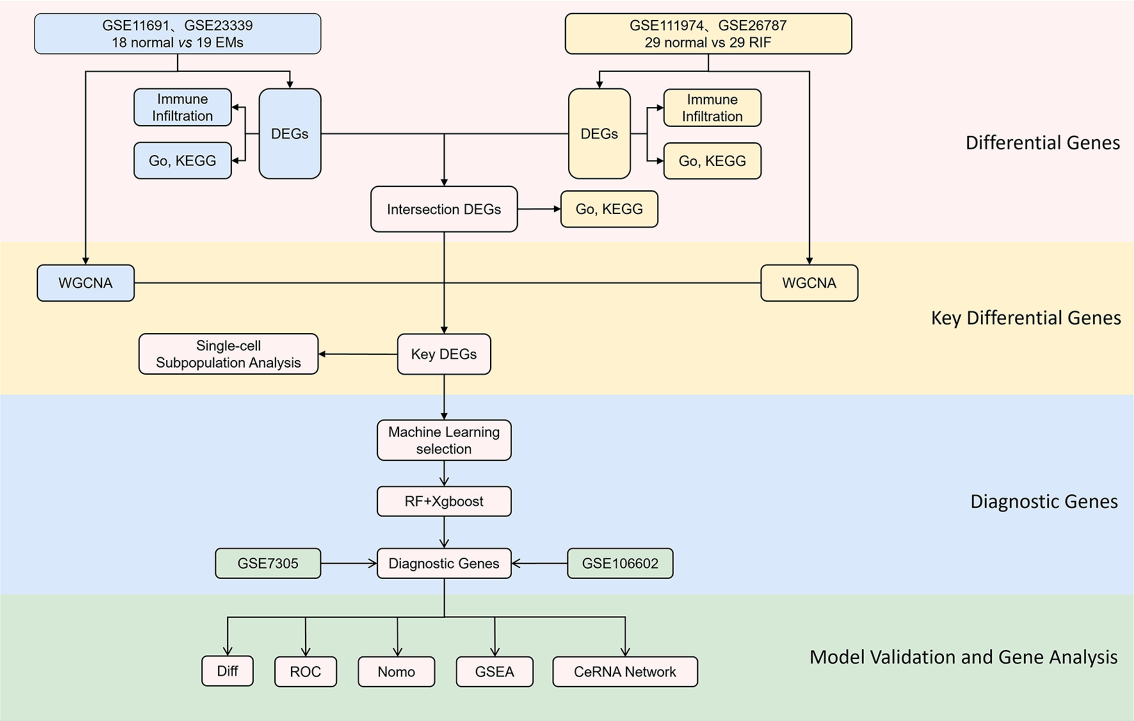


Fig. 1. Flow chart of this study design. EMs, endometriosis; RIF, recurrent implantation failure; DEGs, differentially expressed genes; WGCNA, Weighted Gene Co-expression Network Analysis; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; RF, Random Forest; Diff, Differential expression; ROC, receiver operating characteristic; GSEA, Gene Set Enrichment Analysis.

by the RunEval function, and these values are saved as the AUC matrix. Finally, the model is sorted according to the average of the AUC values and the first 50 models are displayed. The intersection of a single model with an AUC value greater than 0.8 and an average AUC value ranking in the top 5 was used to identify the intersection genes of EMs and RIF. These intersection genes are considered to be shared diagnostic genes for EMs and RIF.

Immune cell abundance

The relative abundance of immune cells was assessed through the application of CIBERSORT analysis on all disease samples. This algorithm employs gene expression data for deconvolution to infer the composition of immune cells. Referring to the CIBERSORT official website (http://cibersort.stanford.edu/), we utilized the LM22, which includes 22 marker gene sets. By applying the CIBERSORT algorithm with 1000 iterations, we performed a quantitative analysis of 22 immune cell types. Subsequently, we selected samples with CIBERSORT P-values < 0.05 for in-depth analysis. For each sample, we normalized the estimates obtained from CIBERSORT analysis so that the sum equals 1, facilitating comparisons across immune cell types and different datasets. We utilized the R packages "vioplot," and "ggplot2" to visualize the analysis results.

Assessment of diagnostic marker prediction model and the construction of nomogram

The diagnostic efficacy of genes for diseases was evaluated using ROC curves, and a clinical risk prediction model was constructed based on the identified diagnostic genes by plotting nomograms with the "rms" package. The calibration of the model was assessed by comparing the predicted probabilities with the actual incidence rates using calibration curves; the predictive accuracy and impact on clinical decision-making of the model were quantified using decision curve analysis (DCA) to evaluate the clinical utility of the model.

Gene set enrichment analysis (GSEA)

GSEA was employed to identify the biological pathways and processes associated with diagnostic genes in EMs and RIF. The "GSEA" package was utilized to perform the GSEA analysis, and the latest pathway-related gene set files (c2.cp.kegg_legacy.v2024.1.Hs.symbols) were downloaded from the Molecular Signatures Database (MSigDB). The GSEA algorithm was run with default settings, with the number of permutations set to 1000 to generate empirical p-values and false discovery rates (FDR) for each gene set. Gene sets with an FDR-adjusted p-value less than 0.05 were considered to be significantly enriched. Finally, the pathview package was utilized to plot the enrichment scores and the rank-ordered list of genes for each significant pathway.

Construction of CeRNA networks

Utilizing three miRNA databases: miRTarBase 6.0, miRDB, and TargetScan 7.0, we identified miRNAs that interact with diagnostic genes, with the criterion that the associated miRNAs must be present in all three databases. Subsequently, we employed the StarBase database to identify lncRNAs associated with these miRNAs. The ceRNA network constructed from the lncRNA-miRNA-mRNA triads was visualized using Cytoscape v3.10.1 software.

Statistical analysis

In this study, statistical differences between groups were considered significant with an adjusted *P*-value < 0.05. For the evaluation of diagnostic capability, an AUC value > 0.7 was used as the standard.

Result

Acquisition of differentially expressed genes

Following data processing and integration, the EMs training set comprised 18 normal samples and 19 disease samples, while the RIF training set consisted of 29 normal samples and 29 disease samples. Principal component analysis (PCA) revealed that the distribution of data across different datasets became more uniform after batch correction (Fig. 2A and C). In the EMs training set, a total of 1,264 differentially expressed genes were identified, including 757 upregulated genes and 507 downregulated genes. The top 10 differentially expressed genes were visualized using a heatmap (Fig. 2B). For the RIF training set, 1,260 differentially expressed genes were identified, comprising 597 upregulated genes and 663 downregulated genes. Similarly, the top 10 differentially expressed genes were visualized using a heatmap (Fig. 2D).

Enrichment analysis of differentially expressed genes

GO analysis indicates that DEGs in EMs are predominantly involved in immune-related processes, including leukocyte migration, chemotaxis, and phagocytosis. KEGG analysis reveals that these DEGs are primarily enriched in pathways associated with the complement and coagulation cascades, suggesting the presence of inflammatory responses and tissue remodeling in the pathogenesis of EMs (Fig. 3A). In RIF, DEGs are enriched in processes related to cellular structure and metabolism, such as connective tissue development and extracellular matrix organization. KEGG analysis shows that DEGs are mainly enriched in phagosome-related processes and folate-mediated one-carbon pool, implying that cellular metabolism and immune evasion may be crucial for the occurrence of RIF (Fig. 3B).

By intersecting the expression trends of DEGs between EMs and RIF, a total of 44 DEGs were identified (Fig. 3C), which are considered to be implicated in the pathogenesis of both conditions. KEGG analysis demonstrates enrichment in pathways related to antigen processing and presentation, as well as natural killer cell-mediated cytotoxicity, further confirming the significance of immune regulation in the development of EMs and RIF (Fig. 3D). Enrichment analysis of molecular functions and biological processes further emphasizes the importance of signaling receptor binding, symporter activity, cytoskeletal dynamics, and ion homeostasis in EMs and RIF (Fig. 3E, F and G).

Identifying key differentially expressed genes

In the EMs training set, when the signed $R^2 = 0.9$, the optimal soft - threshold value is 5, and a total of 11 gene modules are identified. Among them, the turquoise and red modules contain 1,826 genes in total, which are closely associated with the occurrence of EMs (Fig. 4A). In the RIF training set, when the signed $R^2 = 0.9$, the optimal soft - threshold value is 11, and 7 gene modules are identified in total. Among them, the magenta and yellow modules contain 1,759 genes in total, which are closely related to RIF (Fig. 4B). By taking the intersection of the key module genes of EMs and RIF, 316 genes are obtained (Fig. 4C). By intersecting the differentially expressed genes, we ultimately identified 16 key differential genes for subsequent selection in our machine learning model (Fig. 4D). Figure 4E illustrates the chromosomal locations of these critical genes.

Investigating the optimal model and shared diagnostic genes through ensemble machine learning models

Through the evaluation of all machine learning models based on the AUC values, the optimal model combination was selected. The top 50 machine learning model scores for EMs and RIF are shown in Fig. 5A and B. Table 2 presents the AUC values for each of the top 5 models ranked by their performance. Among the top 5 models with the highest average AUC scores, the RF + xgboost model was the only one that appeared more than once, with AUC values greater than 0.7 in both the training and validation sets, thus it was used as the optimal model for screening diagnostic genes.

The top 5 genes scored by individual algorithms were intersected to screen for disease diagnostic genes. The diagnostic genes for EMs are shown in Fig. 6A. The Xgboost algorithm and RF algorithm together identified 4 diagnostic genes, including PDIA4, PGBD5, TSPAN1, TAF7L; the RIF identified three diagnostic genes, including PCSK1N, PDIA4, PGBD5 (Fig. 6B). By intersecting the diagnostic genes from EMs and RIF, the shared diagnostic genes PDIA4 and PGBD5 were selected (Fig. 6C). The differential expression of the shared diagnostic genes was examined in the training and validation sets, as shown in Fig. 6D. Compared to the normal group, the expression of PDIA4 and PGBD5 genes was decreased in both EMs and RIF, and this decrease was statistically significant.

Analysis of key cellular subpopulations associated with diagnostic genes in EMs and RIF

To further elucidate the cellular subpopulations and expression distribution characteristics of diagnostic genes at the single-cell level, we conducted analyses based on single-cell sequencing data from EMs and RIF. In the EMs dataset (GSE214411), we extracted 6 healthy samples and 4 disease samples for analysis. After quality control and normalization (Fig. 7A and B), we performed dimensionality reduction analysis based on 3,000 highly variable genes and identified an optimal harmony value of 9 (Fig. 7C and D). Subsequently, we identified nine distinct

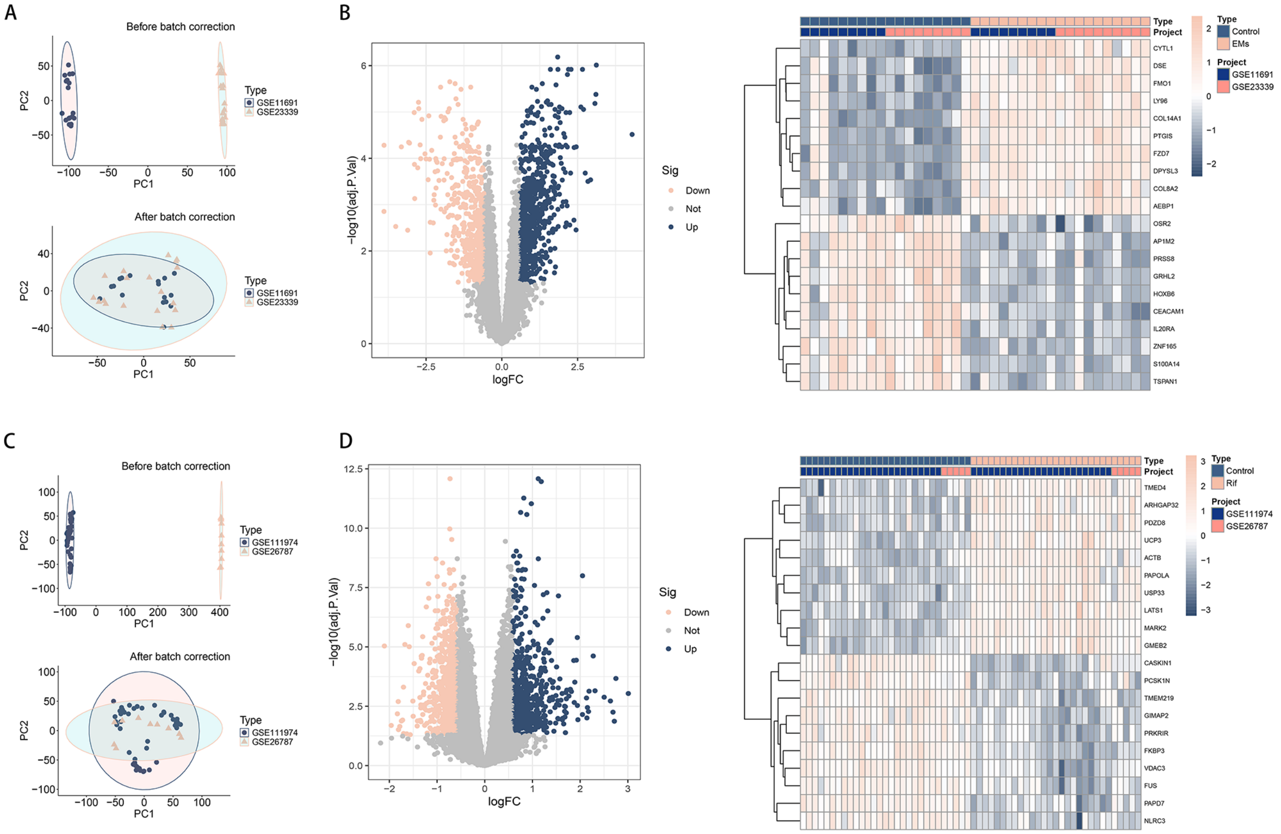


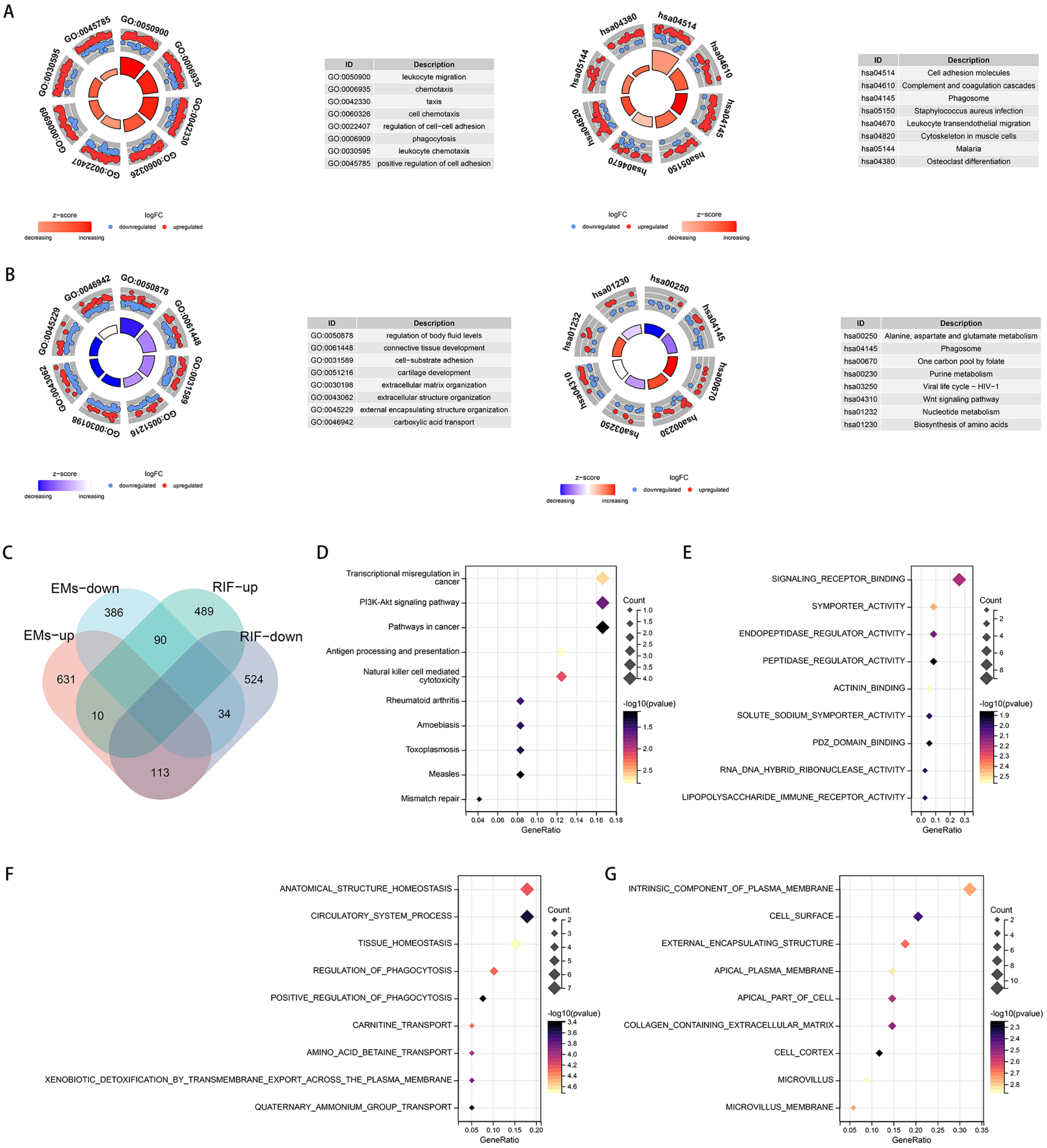
Fig. 2. The integration of EMs and RIF datasets and differential expression analysis of the integrated EMs and RIF dataset. (**A**) PCA of original EMs datasets before batcheffect correction and PCA of the integrated EMs dataset after batcheffect correction. (**B**) The volcano plot and heatmap representing EMs DEGs in the integrated EMs dataset. (**C**) PCA of original RIF datasets before batcheffect correction and PCA of the integrated RIF dataset after batcheffect correction. (**D**) The volcano plot and heatmap representing RIF DEGs in the integrated RIF dataset. EMs, endometriosis; RIF, recurrent implantation failure; PCA, Principle-component Analysis; DEGs, differentially expressed genes.

cell types and further annotated eight cellular subpopulations using the singleR algorithm, including natural killer cells (NK cells), fibroblasts, epithelial cells, endothelial cells, smooth muscle cells, monocytes, neutrophils, and tissue stem cells (Fig. 7E). Through UMAP plots, we visualized the distribution of diagnostic genes across different cellular subpopulations and quantified their expression levels using violin plots (Fig. 7F-H). Results showed that fibroblasts were the predominant subpopulation for the distribution of diagnostic genes (Fig. 7I). We then conducted differential expression analysis of diagnostic genes across various cellular subpopulations between different cohorts. Compared to the healthy group, PIDA4 expression was significantly reduced in all cellular subpopulations except epithelial cells and NK cells in the disease group (Fig. 7J). Additionally, PGBD5 expression was significantly decreased in five cellular subpopulations, including fibroblasts (Fig. 7K).

The same analytical approach was applied to the RIF single-cell dataset (GSE183837), which included 3 healthy samples and 6 disease samples. After quality control and normalization (Fig. 8A and B), we performed dimensionality reduction analysis based on 3,000 highly variable genes and identified an optimal harmony value of 11(Fig. 8C and D). Subsequently, we identified 12 distinct cell types and annotated eight cell types using the singleR algorithm, including smooth muscle cells, fibroblasts, tissue stem cells, mesenchymal stem cells (MSC), NK cells, epithelial cells, endothelial cells, and monocytes (Fig. 8E). UMAP and violin plots were used to visualize the distribution of diagnostic genes across cellular subpopulations. Results showed that diagnostic genes were primarily distributed in fibroblasts (Fig. 8F-H). Differential analysis revealed that (Fig. 8I), compared to the healthy group, PIDA4 expression was significantly reduced in fibroblasts and smooth muscle cells in the RIF cohort (Fig. 8J). In contrast, PGBD5 expression was significantly decreased in fibroblasts and MSC but increased in tissue stem cells and endothelial cells (Fig. 8K).

Construction of diagnostic models for EMs and RIF

To further validate the ability of diagnostic genes to distinguish between normal and disease states, we conducted a ROC analysis on diagnostic genes in the training sets, and constructed a diagnostic model in the training set to assess its clinical utility. In the EMs and RIF, the AUC values for individual diagnostic genes were all greater than 0.7 (Fig. 9A and D), indicating that these genes have a high degree of accuracy in disease identification. The clinical diagnostic model for EMs constructed using these diagnostic genes is detailed in Fig. 9B. The calibration



ID	Description
GO:0050900	leukocyte migration
GO:0006935	chemotaxis
GO:0042330	taxis
GO:0060326	cell chemotaxis
GO:0022407	regulation of cell−cell adhesion
GO:0006909	phagocytosis
GO:0030595	leukocyte chemotaxis
GO:0045785	positive regulation of cell adhesion

ID	Description
hsa04514	Cell adhesion molecules
hsa04610	Complement and coagulation cascades
hsa04145	Phagosome
hsa05150	Staphylococcus aureus infection
hsa04670	Leukocyte transendothelial migration
hsa04820	Cytoskeleton in muscle cells
hsa05144	Malaria
hsa04380	Osteoclast differentiation

ID	Description
GO:0050878	regulation of body fluid levels
GO:0061448	connective tissue development
GO:0031589	cell−substrate adhesion
GO:0051216	cartilage development
GO:0030198	extracellular matrix organization
GO:0043062	extracellular structure organization
GO:0045229	external encapsulating structure organization
GO:0046942	carboxylic acid transport

ID	Description
hsa00250	Alanine, aspartate and glutamate metabolism
hsa04145	Phagosome
hsa00670	One carbon pool by folate
hsa00230	Purine metabolism
hsa03250	Viral life cycle − HIV−1
hsa04310	Wnt signaling pathway
hsa01232	Nucleotide metabolism
hsa01230	Biosynthesis of amino acids

Fig. 3. DEGs Functional Enrichment Analysis. (**A**) GO and KEGG Analysis of DEGs in EMs. (**B**) GO and KEGG Analysis of DEGs in RIF. (**C**) Shared-DEGs with the same expression trend between EMs and RIF. (**D**) KEGG Analysis of Shared-DEGs. (**E**) MF Analysis of Shared-DEGs. (**F**) BP Analysis of Shared-DEGs. (**G**) CC Analysis of Shared-DEGs. EMs, endometriosis; RIF, recurrent implantation failure; Go, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes. CC, Cellular Component; BP, Biological Process; MF, Molecular Function.

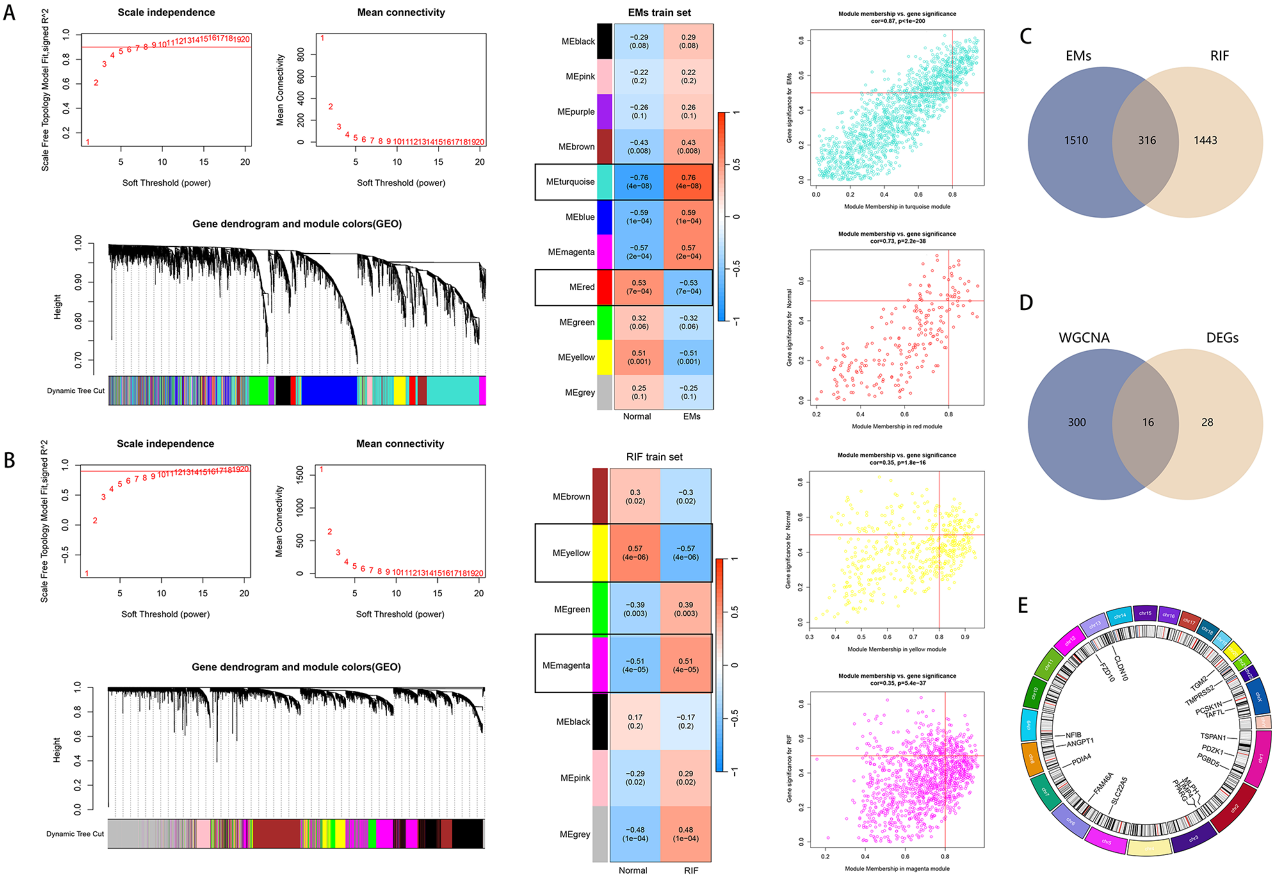


Fig. 4. Key DEGs Screening. (**A**) WGCNA Screening of Key Module Genes in EMs. (**B**) WGCNA Screening of Key Module Genes in RIF. (**C**) Intersection of Key Genes. (**D**) Key DEGs. E: Chromosomal Distribution Map of Key DEGs. EMs, endometriosis; RIF, recurrent implantation failure; DEGs, differentially expressed genes; WGCNA, Weighted Gene Co-expression Network Analysis.

curve of the diagnostic model typically coincides with the ideal curve, and the AUC value of the predictive model reached 0.91. Decision curve analysis (DCA) demonstrate that these diagnostic genes have significant practical value in disease discrimination (Fig. 9C). In the clinical prediction model for RIF, a similar trend in diagnostic effect is also exhibited (Fig. 9E and F).

Gene expression patterns and functional pathway enrichment of PGBD5 and PDIA4 in EMs and RIF

GSEA was utilized to conduct an in-depth analysis of the expression patterns of PGBD5 and PDIA4 in EMs and RIF. In EMs, PGBD5 in the low-expression group was significantly associated with pathways related to vascular function, such as vascular smooth muscle contraction and calcium signaling pathways, while in the high-expression group, it was associated with immune-related pathways, including cytokine-cytokine receptor interaction and leukocyte transendothelial migration (Fig. 10A). PDIA4 in the low-expression group was associated with pathways such as vascular smooth muscle contraction and muscle contraction, while in the high-expression group, it was related to cancer-related pathways, including basal cell carcinoma and Hedgehog signaling pathways (Fig. 10B). In RIF, PGBD5 in the low-expression group was associated with the enrichment of pathways related to steroid hormone biosynthesis and starch and sucrose metabolism, while in the high-expression group, it was associated with oxidative phosphorylation and peroxisome-related pathways (Fig. 10C). PDIA4 in the low-expression group was associated with the enrichment of drug metabolism-cytoochrome P450 and starch and sucrose metabolism-related pathways, which may be related to drug metabolism and carbohydrate metabolism. In the high-expression group, PDIA4 was associated with the enrichment of cell cycle and spliceosome-related pathways, suggesting a potential role in cell proliferation and RNA splicing processes (Fig. 10D).

Immune infiltration profiles in EMs and RIF

To further elucidate the immune pathogenesis of the disease, we conducted an investigation into the immune cell infiltration patterns within the EMs and RIF training cohorts. Utilizing a sophisticated cellular classification algorithm, we ascertained the infiltration profiles for a spectrum of immune cells associated with the disease in question. Figure 11A illustrates the immune cell infiltration landscape in EMs, highlighting a marked increase

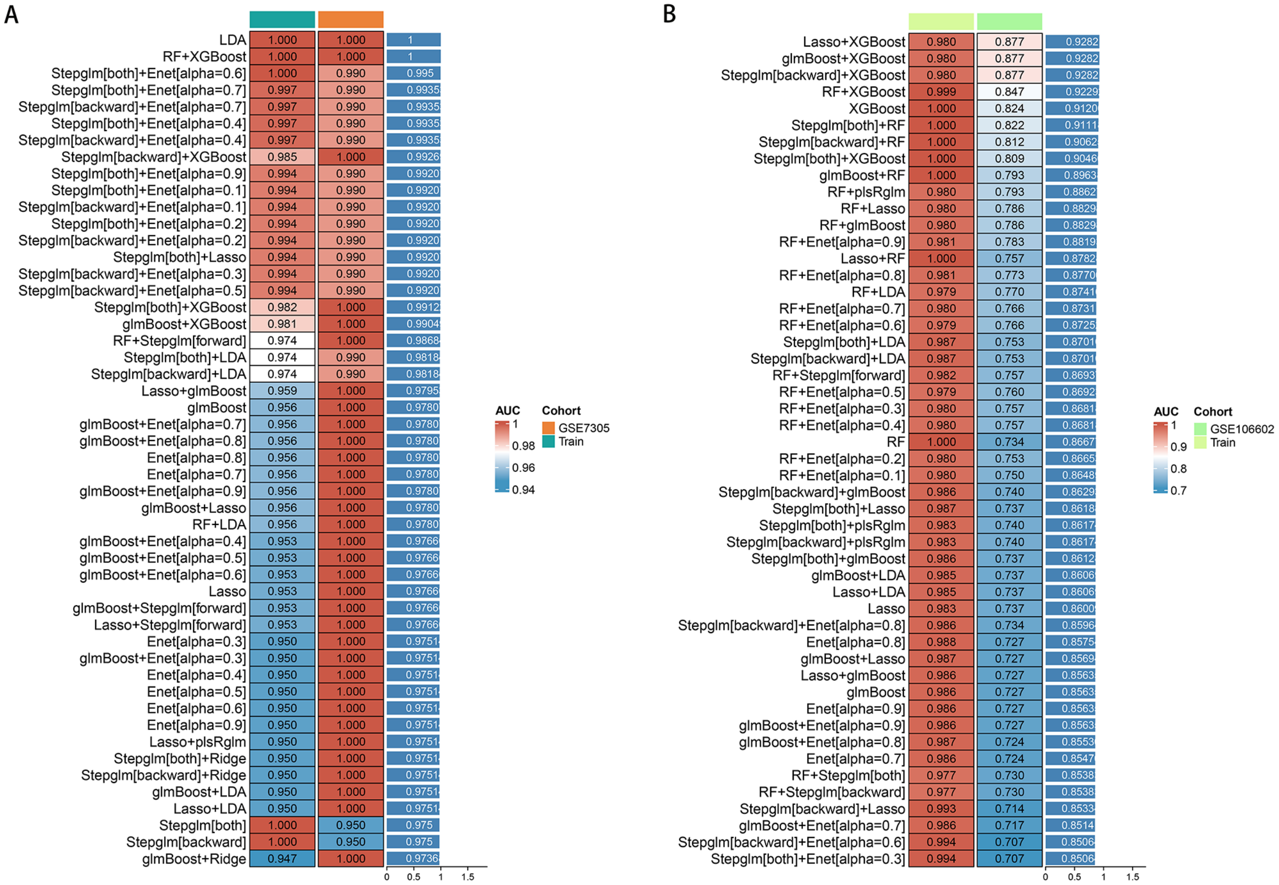


Fig. 5. Constructing the Optimal Model by Combining Multiple Machine Learning Methods. (**A**) Model Construction for EMs. (**B**) Model Construction for RIF. EMs, endometriosis; RIF, recurrent implantation failure.

Disease	Method	AUC-Train	AUC-Validation
EMs	LDA	1	1
	RF + XGBoost	1	1
	Stepglm[both] + Enet[alpha = 0.6]	1	0.99
	Stepglm[both] + Enet[alpha = 0.7]	0.997	0.99
	Stepglm[backward] + Enet[alpha = 0.7]	0.997	0.99
RIF	Lasso + XGBoost	0.98	0.877
	glmBoost + XGBoost	0.98	0.877
	Stepglm[backward] + XGBoost	0.98	0.877
	RF + XGBoost	0.999	0.847
	XGBoost	1	0.824

Table 2. AUC values of the top 5 machine learning models.

in the prevalence of M2 macrophages, centocytes, and select T cell populations compared to the control group, while a notable reduction in the frequency of CD8 T cells, Treg cells, and NK cells was observed. In the RIF cohort, a significant elevations in the abundance of M0 macrophages were detected within the diseased group in contrast to the control, alongside a pronounced decrease in the presence of γδT cells, M2 macrophages, and dendritic cells (Fig. 11B).

The correlation between diagnostic genes and immune cells suggests that in EMs, both PIDA4 and PGBD5 are associated with the suppression of B cells and the activation of M2 macrophages (Fig. 11C). In RIF, however, PDIA4 is correlated with the suppression of B cells, while PGBD5 is associated with the suppression of T cells (Fig. 11D).

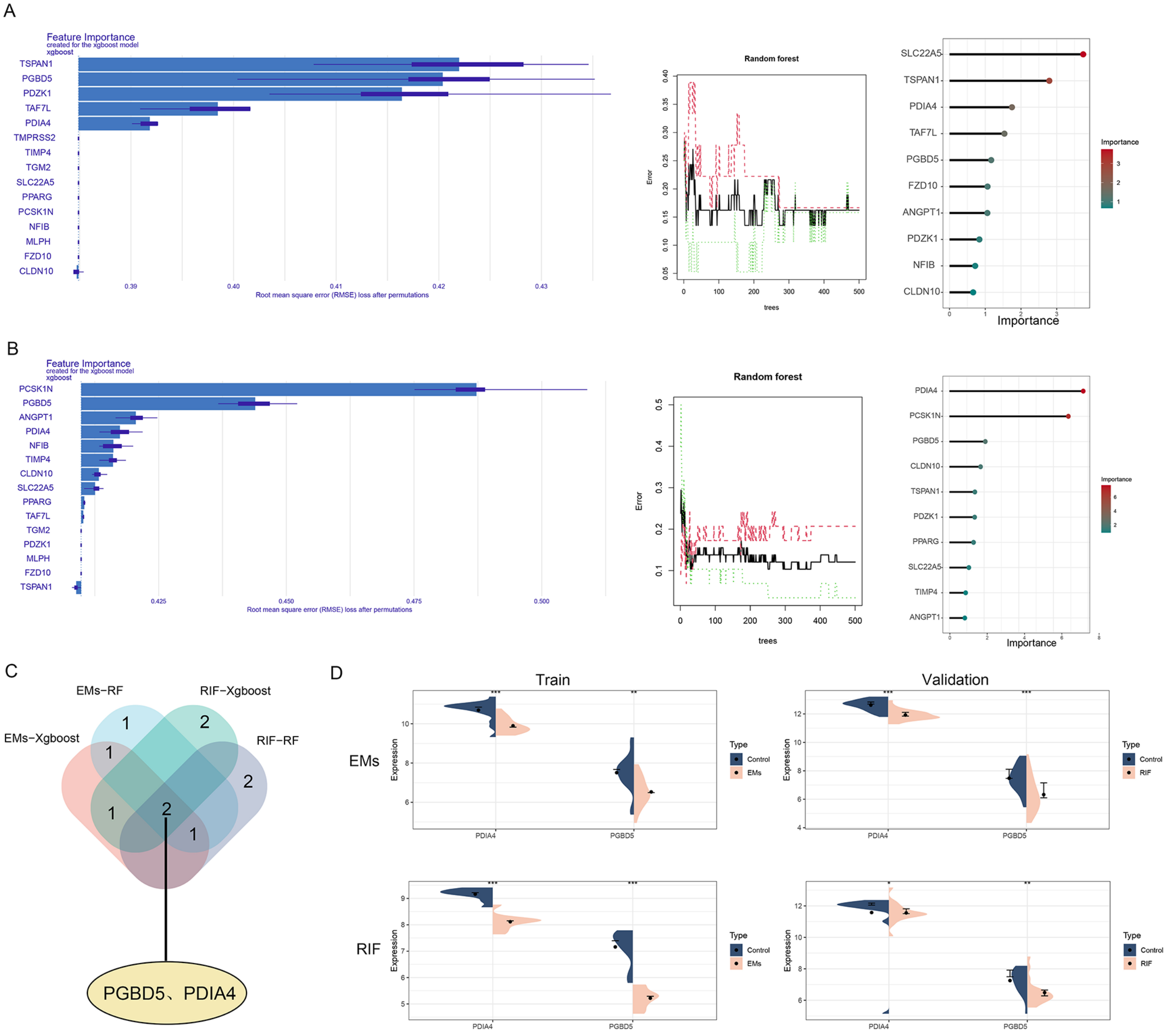


Fig. 6. Shared Diagnostic Gene Screening. (**A**) Diagnostic Gene Screening for EMs Using RF + XGBoost Algorithm. (**B**) Diagnostic Gene Screening for RIF Using RF + XGBoost Algorithm. (**C**) Shared Diagnostic Genes between EMs and RIF. (**D**) Differential Expression of Shared Diagnostic Genes in EMs and RIF. EMs, endometriosis; RIF, recurrent implantation failure; RF, Random Forest.

Construction of the CeRNA network of diagnostic genes

The lncRNA-miRNA-mRNA ceRNA network constructed based on diagnostic genes is shown in Fig. 12. The network comprises a total of 36 miRNA molecules and 45 lncRNA molecules. Among them, miR-3121-3p is connected to two diagnostic genes, suggesting that this miRNA may be involved in the regulation of the occurrence of EMs and RIF by targeting these diagnostic genes.

Discussion

RIF associated with EMs has long been a challenging issue in the field of reproductive medicine. This study aims to identify shared diagnostic genes and changes in the cellular microenvironment between EMs and RIF through multi-omics analysis and machine learning approaches. By integrating transcriptomic data and single-cell sequencing data, combined with differential expression analysis, WGCNA, and various machine learning algorithms, we successfully identified PDIA4 and PGBD5 as shared diagnostic genes between the two diseases and constructed clinically predictive models with high accuracy. During the analysis, we first obtained relevant datasets from the GEO database and identified differentially expressed genes in EMs and RIF through a rigorous screening and normalization process. WGCNA analysis further filtered out gene modules closely related to disease occurrence, and machine learning algorithms (including random forest and XGBoost) were used to select genes with diagnostic potential. Ultimately, ROC analysis validated the diagnostic efficacy of PDIA4 and PGBD5, with AUC values both exceeding 0.7, indicating high accuracy in disease identification. Additionally, single-

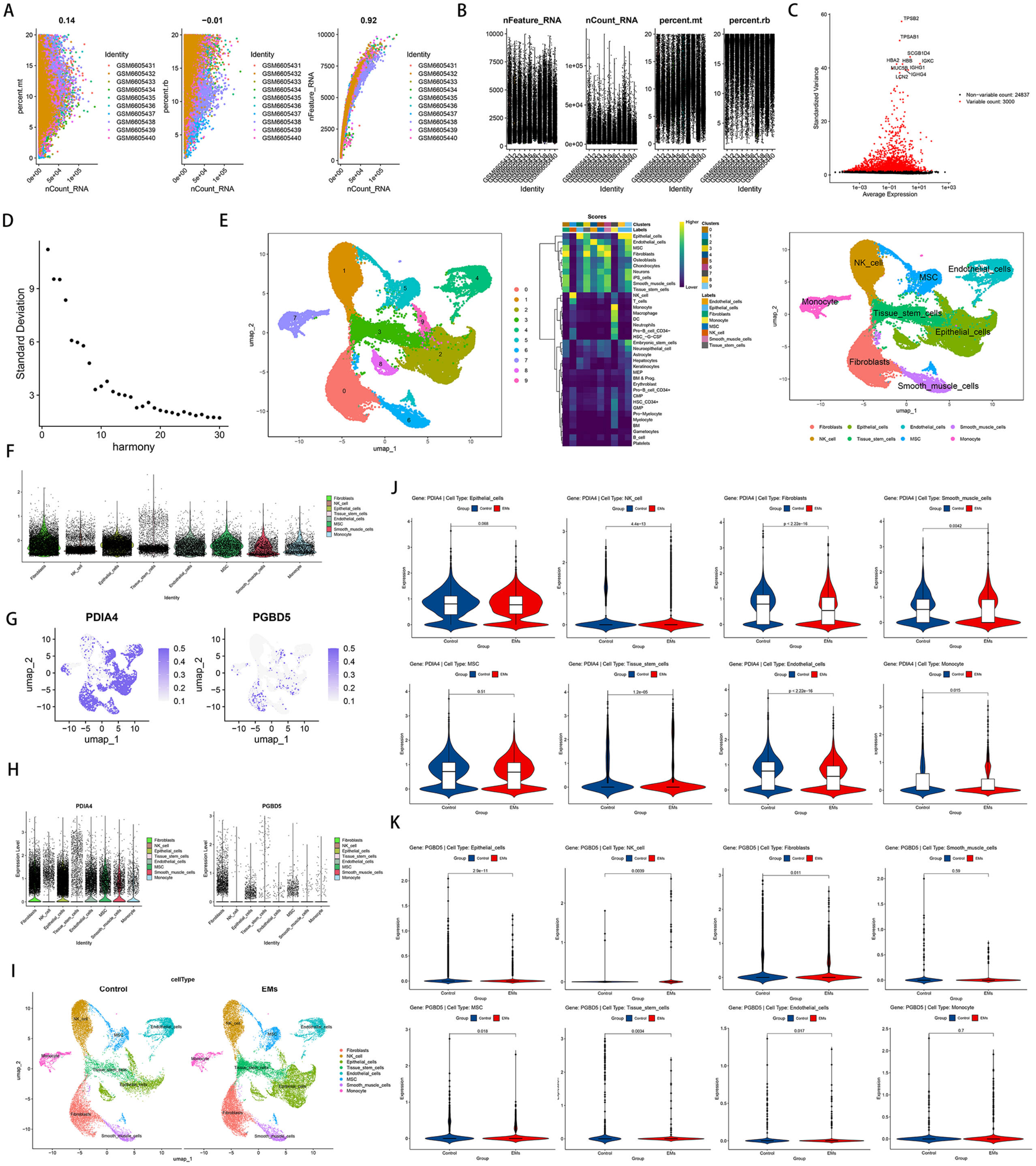


Fig. 7. Single-cell data analysis of EMs. (**A**) Data Quality Control. (**B**) Gene expression quality control level. (**C**) The top 3000 highly variable genes. (**D**) Analysis of the Correlation between Harmony and Standard Deviation. (**E**) Cell clustering analysis and annotation. (**F**) Distribution of Genes in Cellular Subpopulations. (**G**) UMAP-Based Feature Plots of Cellular Gene Expression. (**H**) The distribution of PDIA4 and PGBD5 in different cell subpopulations. (**I**) The distribution of different cell subpopulations in healthy and diseased populations. (**J**) The differential levels of PDIA4 in different cell subpopulations between healthy and diseased populations. **K**: The differential levels of PGBD5 in different cell subpopulations between healthy and diseased populations.

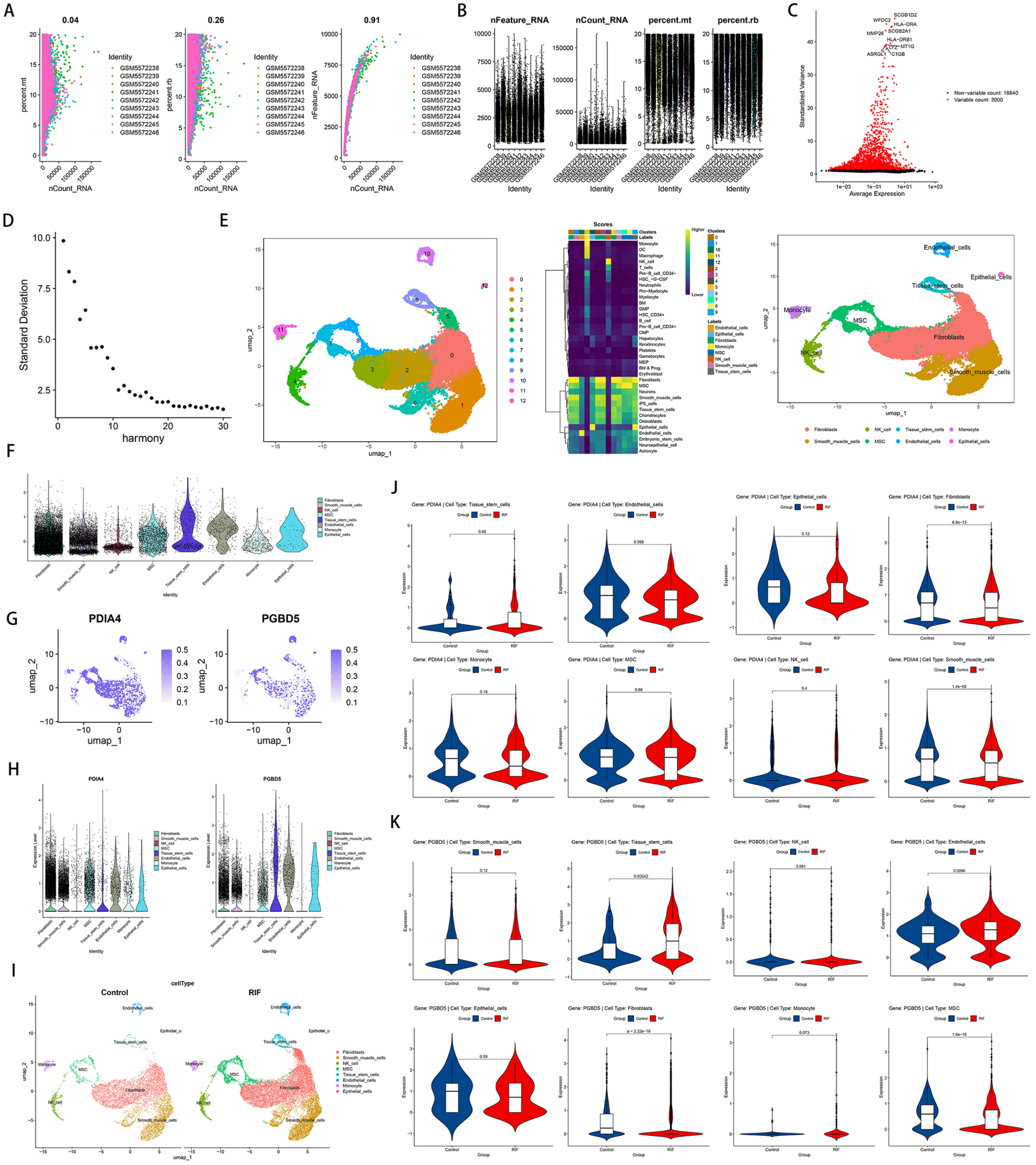


Fig. 8. Single-cell data analysis of RIF. (**A**) Data Quality Control. (**B**) Gene expression quality control level. (**C**) The top 3000 highly variable genes. (**D**) Analysis of the Correlation between Harmony and Standard Deviation. (**E**) Cell clustering analysis and annotation. (**F**) Distribution of Genes in Cellular Subpopulations. (**G**) UMAP-Based Feature Plots of Cellular Gene Expression. (**H**) The distribution of PDIA4 and PGBD5 in different cell subpopulations. (**I**) The distribution of different cell subpopulations in healthy and diseased populations. (**J**) The differential levels of PDIA4 in different cell subpopulations between healthy and diseased populations. **K**: The differential levels of PGBD5 in different cell subpopulations between healthy and diseased populations.

cell analysis revealed significant expression differences of these diagnostic genes in fibroblast subpopulations, suggesting that fibroblasts may play a key role in the pathogenesis of EMs and RIF. GSEA further indicated that these genes are primarily involved in immune response, vascular function, and hormone regulation, and may be regulated by miR-3121-3p. Immune infiltration analysis also revealed the important roles of M2 macrophages and γδ T cells in the pathogenesis of both diseases. In summary, this study not only successfully identified PDIA4 and PGBD5 as shared diagnostic genes for EMs and RIF but also further elucidated their potential mechanisms of action in the diseases through single-cell analysis and bioinformatics methods. These findings provide new directions for the early diagnosis and targeted treatment of EMs-related RIF and offer important evidence for the development of future clinical applications and therapeutic strategies.

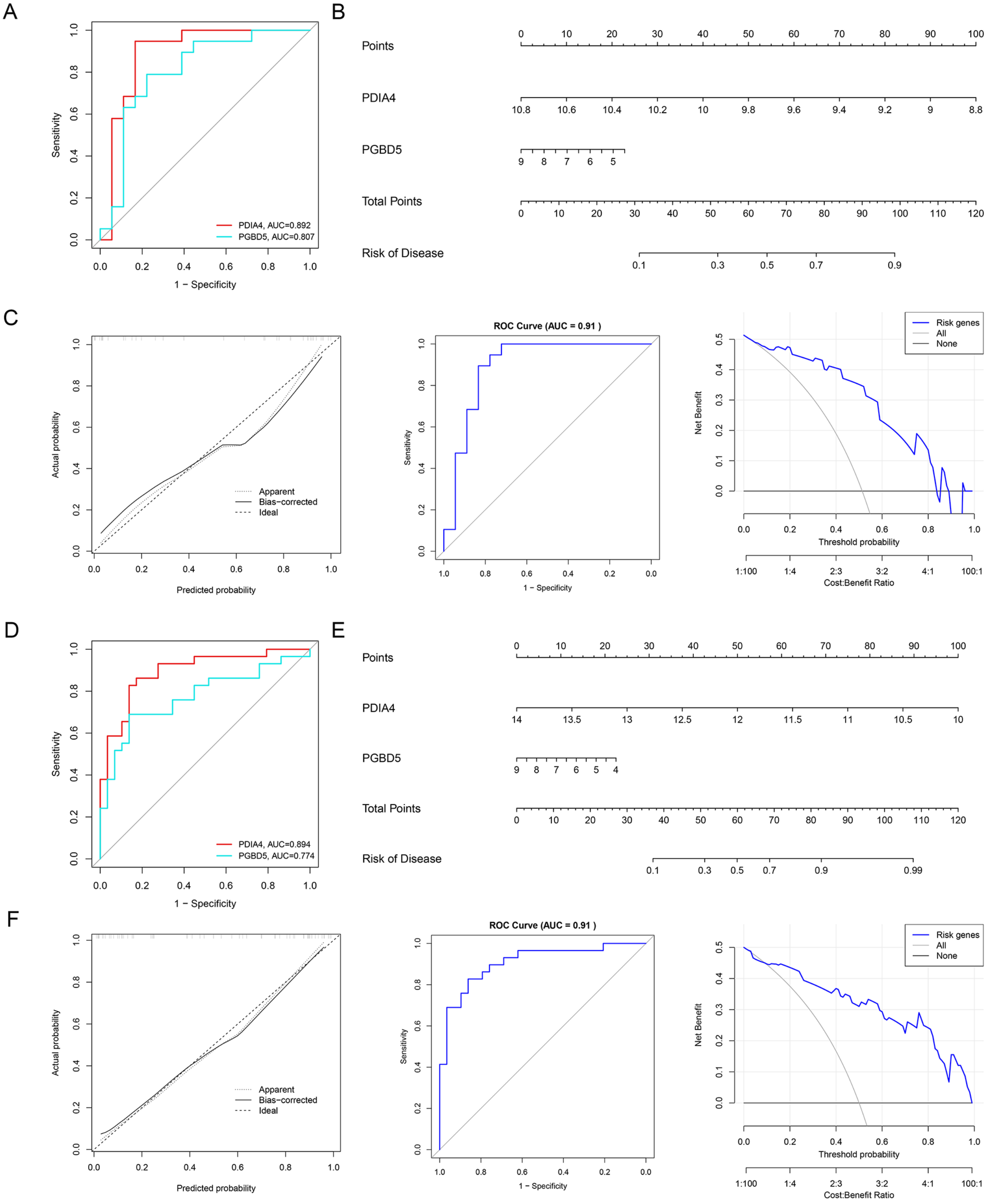


Fig. 9. Clinical Model Validation of Shared Diagnostic Genes. (**A**) ROC Analysis to Assess the Diagnostic Ability of Genes in EMs. (**B**) Constructing EMs Clinical Model. (**C**) Identify the Discriminative Ability of Shared Diagnostic Genes for EMs. (**D**) ROC Analysis to Assess the Diagnostic Ability of Genes in RIF. (**E**) Constructing RIF Clinical Model. (**F**) Identify the Discriminative Ability of Shared Diagnostic Genes for RIF. EMs, endometriosis; RIF, recurrent implantation failure; ROC, receiver operating characteristic.

Consistent with previous studies, our analysis also indicates significant changes in the immunological environment of the endometrium in these diseases[23,24], and we further identified significant alterations in M2 macrophages and γδT cells in both EMs and RIF. M2 macrophages, due to their properties of promoting immune tolerance and angiogenesis, play a pivotal role in the initiation and progression of EMs[25]. Conversely, in RIF, the reduction in M2 macrophage infiltration suggests compromised endometrial receptivity, potentially disrupting the inflammatory environment necessary for embryo implantation[24,26]. As a unique subset of T cells, γδT cells are involved in immune surveillance and inflammatory responses, and are implicated in the pathogenesis of both EMs and RIF[24,27,28]. In EMs, γδT cells may participate in local immune modulation and inflammatory responses[28], while their reduction in RIF may adversely affect embryo implantation[29]. The mechanistic links between these immune cells and the development of EMs and RIF are further supported by research, indicating that the accumulation of M2 macrophages in the peritoneal cavity of women with EMs is due to the secretion of local chemokines[30], and their impaired clearance mechanisms contribute to the establishment of pelvic EMs[31]. Furthermore, γδT cells are believed to influence the initiation and progression of EMs and RIF by modulating the local immune environment and cytokine secretion[32]. In summary, our analysis further highlights the significant roles of M2 macrophages and γδT cells in the pathogenesis of EMs and RIF, providing new insights into diagnostic and therapeutic strategies in reproductive immunology.

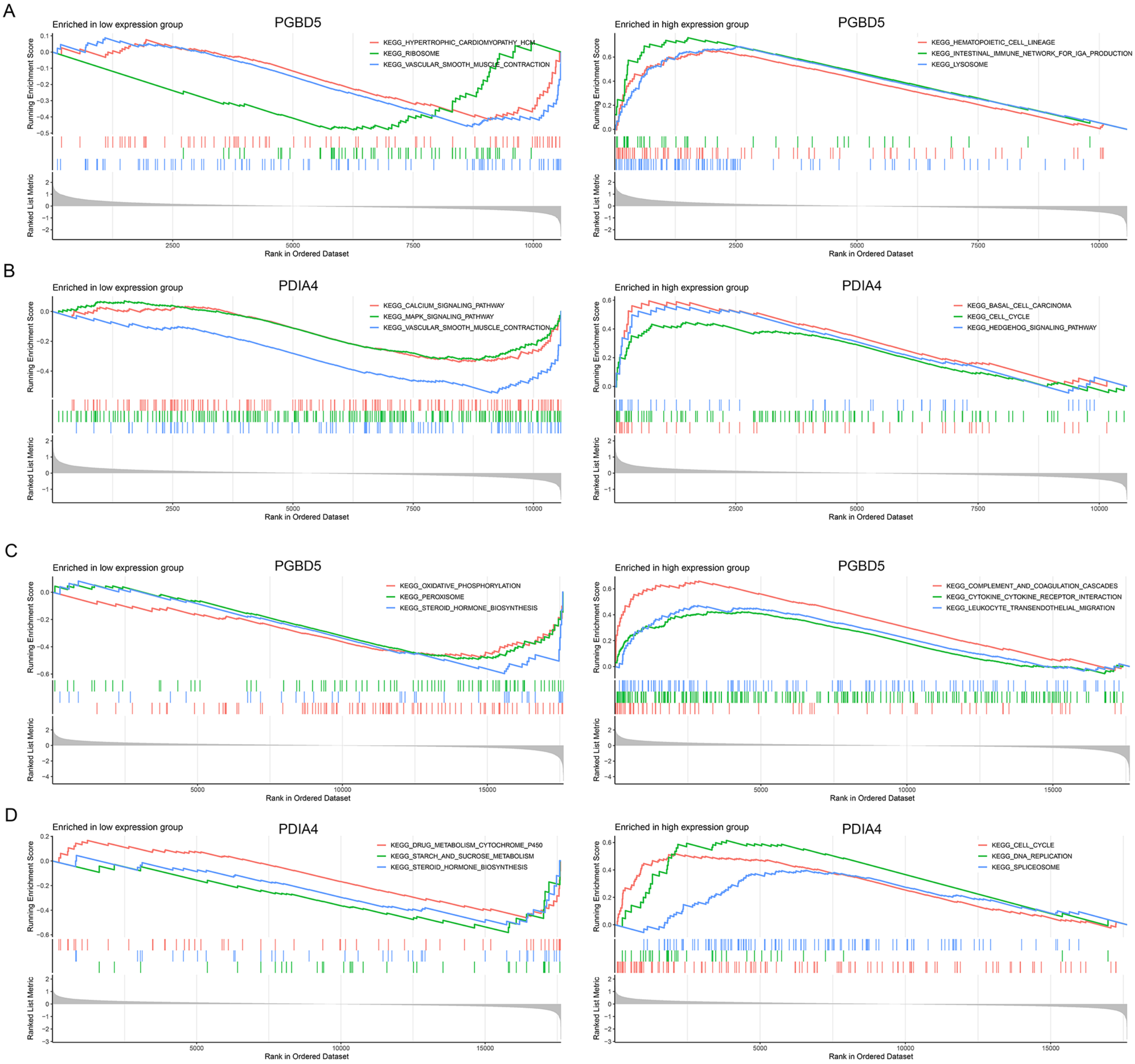


Fig. 10. GSEA for the Shared Diagnostic Genes. (**A**) GSEA for PGBD5 in EMs. (**B**) GSEA for PDIA4 in EMs. (**C**) GSEA for PGBD5 in RIF. (**D**) GSEA for PDIA4 in RIF. EMs, endometriosis; RIF, recurrent implantation failure; GSEA, Gene Set Enrichment Analysis.

Diverging from prior studies, our research integrates multi-omics data and employs a variety of machine learning models to identify the optimal diagnostic model, culminating in the construction of a disease prediction model. Compared to the commonly used single-omics data or traditional statistical methods in previous studies, our approach of multi-omics integration and machine learning demonstrates superior accuracy and robustness in model selection. Moreover, we have identified PDIA4 and PGBD5 as shared diagnostic genes EMs and RIF for the first time, an unprecedented finding. Previous research has predominantly focused on biomarkers associated with single diseases, whereas our study, through multi-omics integration and single-cell analysis, has revealed these shared genes between the two diseases for the first time, providing new targets for early diagnosis and targeted therapy. Single-cell analysis further confirmed that these genes are predominantly expressed in fibroblast subpopulations, elucidating the critical role of fibroblasts in the pathogenesis of EMs and RIF. ROC analysis indicated that these genes exhibit high accuracy in disease diagnosis (with AUC values all above 0.7), and the clinical diagnostic model constructed based on these genes also demonstrated good disease prediction capability.

PDIA4, a member of the protein disulfide isomerase (PDI) family, primarily functions to catalyze the proper folding of proteins and the correct formation of disulfide bonds in the endoplasmic reticulum[33]. Studies have indicated that PDIA4 plays a pivotal role in the development of insulin resistance (IR), particularly in skeletal muscle[34]. Furthermore, an increase in PDIA4 expression levels has been found to be correlated with elevated levels of IR and inflammatory cytokines[34]. PGBD5, a gene associated with the PiggyBac transposon, has been demonstrated to induce transposition of genomic DNA[35]. Currently, research on PGBD5 is predominantly focused on the field of oncology[36,37]. Although studies on PDIA4 and PGBD5 in EMs and RIF have not

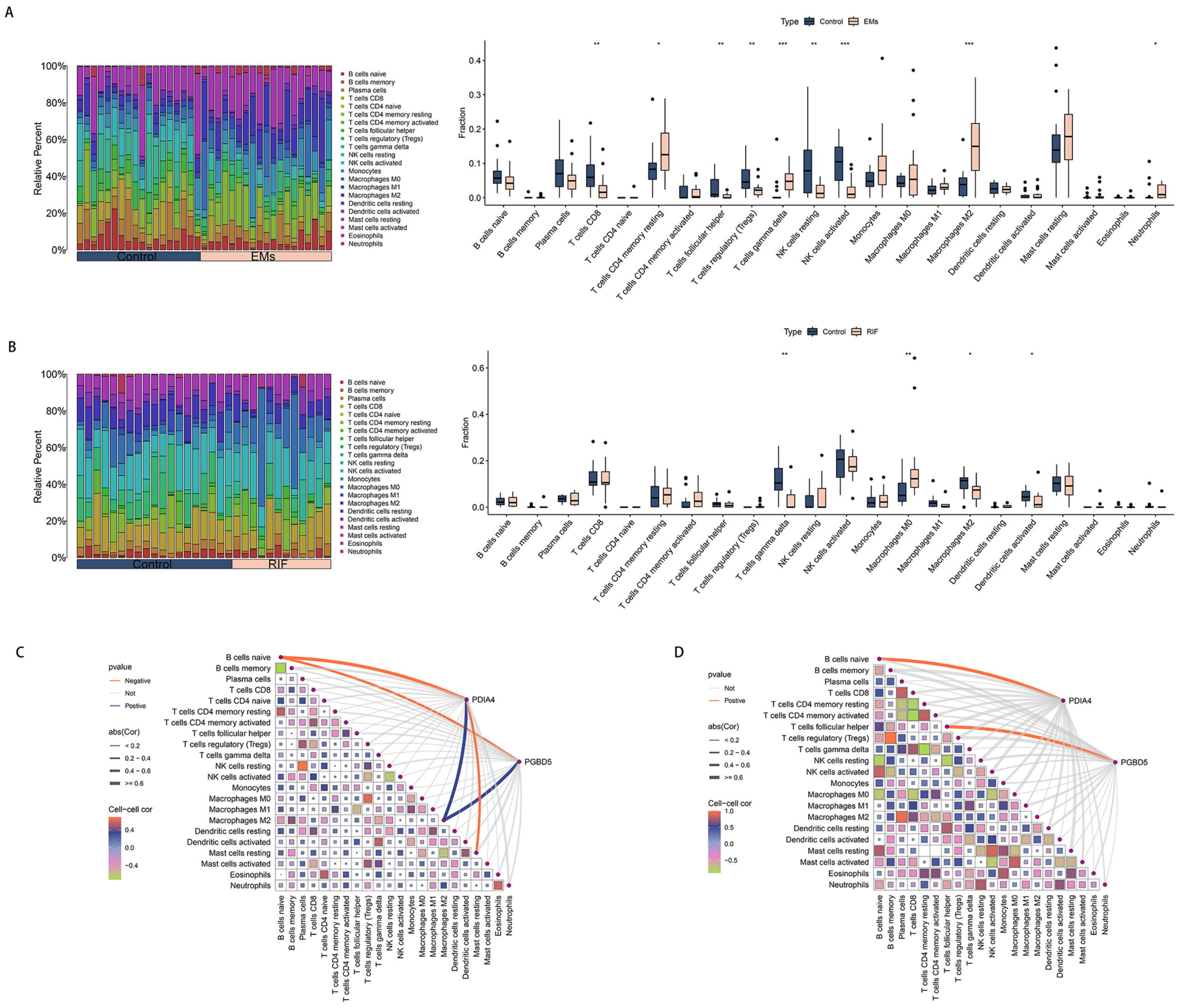


Fig. 11. Immunoinfiltration Analysis in EMs and RIF. (**A**) Immune Cell Distribution in the Training Set of EMs. (**B**) Immune Cell Distribution in the Training Set of RIF. (**C**) The correlation between PDIA4 and PGBD5 with immune cells in EMs. (**D**) The correlation between PDIA4 and PGBD5 with immune cells in RIF. EMs, endometriosis; RIF, recurrent implantation failure.

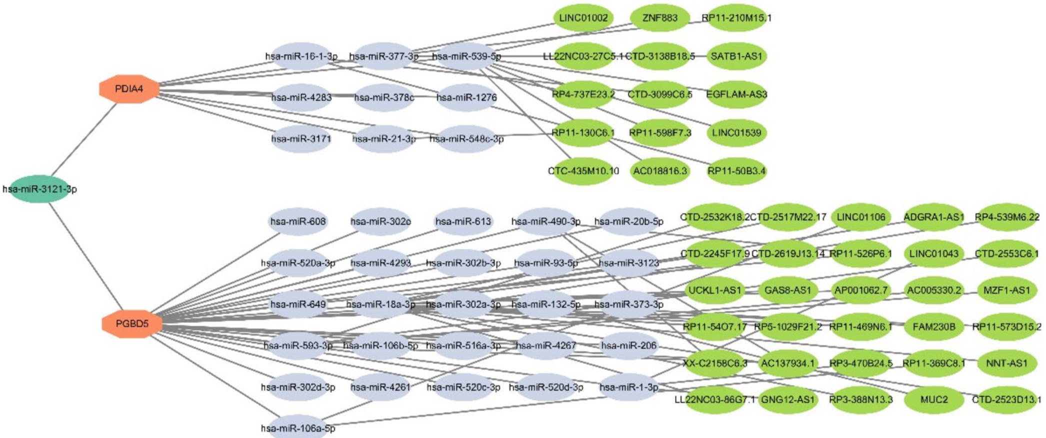


Fig. 12. Constructing a ceRNA Network Based on Shared Diagnostic Genes.

been reported, our analysis revealed the association of these two genes with multiple biological pathways in EMs and RIF. In EMs, the expression of PDIA4 is significantly correlated with signaling pathways related to vascular smooth muscle contraction and muscle contraction, while PGBD5 is associated with pathways related to vascular function and immunity. In RIF, PDIA4 is related to signaling pathways associated with drug metabolism and carbohydrate metabolism, and PGBD5 is related to pathways associated with steroid hormone biosynthesis and oxidative phosphorylation. The activation of these signaling pathways may be closely related to the pathophysiology of EMs and RIF. Specifically, the activation of vascular smooth muscle contraction and immunity-related signaling pathways may be associated with inflammatory responses and angiogenesis in EMs[38–40], while the activation of drug metabolism and cell cycle-related signaling pathways may be related to implantation failure and cell proliferation in RIF[41]. Therefore, the expression patterns of PDIA4 and PGBD5 in EMs and RIF, and their association with related signaling pathways, suggest that they may play a key role in the occurrence and development of these diseases and could potentially become new therapeutic targets for EMs-related RIF treatment. Based on diagnostic genes, we constructed a ceRNA network to explore the correlation between miRNAs and lncRNAs. We ultimately identified a miRNA that is associated with both diagnostic genes: miR-3121-3p. As a microRNA, miR-3121-3p plays a key role in various biological processes. Current research indicates that miR-3121-3p is significantly involved in cell proliferation, migration, and inflammatory responses[42,43]. These biological processes are associated with RIF related to EMs. Therefore, miR-3121-3p may play an important role in EMs-related RIF through the PDIA4 and PGBD5. However, research on the role of miR-3121-3p in EMs and RIF is currently insufficient. Our inference is based solely on the role of miR-3121-3p in existing studies. Further research requires more comprehensive experiments to validate the role of miR-3121-3p.

Currently, research on RIF and EMS primarily focuses on endometrial epithelial cells and immune cell subsets. Studies have shown that abnormal gene expression in endometrial epithelial cells may affect embryo implantation[44]. Additionally, RIF patients exhibit immune cell imbalances in the endometrium, including abnormalities in the number and function of immune cells such as natural killer (NK) cells, T cells, and macrophages. These immune cell abnormalities may interfere with the embryo implantation process, leading to the occurrence of RIF[45–47]. However, our study found significant differences in diagnostic genes in fibroblasts of EMS and RIF patients, suggesting that fibroblasts may be a key cell subset in EMS-related RIF. Fibroblasts in the endometrium have multiple functions, including the synthesis and remodeling of the extracellular matrix, intercellular signaling, and regulation of immune cells[44]. Differences in gene expression in fibroblasts may affect their function and subsequently interfere with the embryo implantation process. Despite these findings, many questions remain unresolved: the high heterogeneity of fibroblasts necessitates further research into the characteristics and functional differences of their subsets; direct validation of fibroblast function is currently lacking, and future studies need to explore their role in RIF through in vitro experiments and animal models[44]; moreover, interactions between fibroblasts and other cell subsets (such as epithelial and immune cells) need to be investigated to fully understand their mechanisms in RIF. Despite these uncertainties, our research findings provide new directions for the clinical treatment of EMS and RIF. Fibroblasts can serve as potential therapeutic targets, and future studies can explore targeted therapies for fibroblasts, such as improving endometrial receptivity by modulating fibroblast gene expression or function. Additionally, based on the gene expression differences in fibroblasts, personalized treatment plans can be developed. For example, by detecting specific gene expression patterns in fibroblasts, it is possible to predict patients' responses to treatment and select the most appropriate therapeutic approaches.

Despite successfully identifying PDIA4 and PGBD5 as shared diagnostic genes for EMs and RIF through multi-omics analysis and machine learning methods, and uncovering the shared characteristics of these two diseases in cellular microenvironments and molecular mechanisms, our study still has several limitations. First, the study results have not yet been experimentally validated, and the specific functions and regulatory mechanisms of PDIA4 and PGBD5 in diseases still need further exploration. Second, the study did not adequately consider the impact of other potential comorbidities (such as adenomyosis or polycystic ovary syndrome) on the expression of diagnostic genes, which may lead to biases in the results. Additionally, the sample size was relatively small, and there may be geographical or population biases, which could affect the universality and statistical power of the study results. Nevertheless, this study still has significant scientific and clinical importance. Through multi-omics analysis and machine learning methods, we successfully identified PDIA4 and PGBD5 as potential diagnostic genes, providing a new direction for early diagnosis and targeted therapy. Moreover, the study revealed the shared characteristics of EMs and RIF in cellular microenvironments and molecular mechanisms, filling the gap in related research. These findings provide a theoretical basis for early prediction of IVF outcomes and optimization of treatment plans, and are expected to significantly reduce the incidence of RIF.

To overcome the limitations of the current study and further promote related research, future research directions include: First, verifying the functions of PDIA4 and PGBD5 through in vitro cell experiments and animal models, using gene-editing technologies (such as CRISPR/Cas9) to knock out or overexpress these genes in endometrial cells and fibroblasts, and studying their effects on cell functions. Second, incorporating samples from more potential comorbidities, assessing their impact on diagnostic gene expression through multifactorial analysis, and developing a combined diagnostic model that takes into account multiple factors to improve diagnostic accuracy and reliability. In addition, collaborating with multiple medical institutions to expand the sample size, enhance the universality and statistical power of the study results, and further integrate global public datasets to increase sample diversity, thereby providing stronger support for clinical diagnosis and treatment.

Data availability

The original contributions presented in the study are publicly available. Publicly available datasets used in this study can be downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). Data numbers include： GSE111691, GSE23339, GSE7305, GSE111974, GSE26787, GSE106602, GSE214411, GSE183837.

Received: 18 December 2024; Accepted: 5 March 2025
Published online: 17 March 2025

References

1. Coughlan, C. et al. Recurrent implantation failure: definition and management. *Reprod. Biomed. Online*. **28** (1), 14–38 (2014).
2. Maesawa, Y., Yamada, H., Deguchi, M. & Ebina, Y. History of biochemical pregnancy was associated with the subsequent reproductive failure among women with recurrent spontaneous abortion. *Gynecol. Endocrinol.* **31** (4), 306–308 (2015).
3. Bashiri, A., Halper, K. I. & Orvieto, R. Recurrent implantation Failure-update overview on etiology, diagnosis, treatment and future directions. *Reprod. Biol. Endocrinol.* **16** (1), 121 (2018).
4. Boucher, A. et al. Implantation failure in endometriosis patients: etiopathogenesis. *J. Clin. Med.* **11**, 18 (2022).
5. Arici, A. et al. The effect of endometriosis on implantation: results from the Yale university in vitro fertilization and embryo transfer program. *Fertil. Steril.* **65** (3), 603–607 (1996).
6. Simon, C. et al. Outcome of patients with endometriosis in assisted reproduction: results from in-vitro fertilization and oocyte donation. *Hum. Reprod.* **9** (4), 725–729 (1994).
7. Andreoli, L. et al. EULAR recommendations for women's health and the management of family planning, assisted reproduction, pregnancy and menopause in patients with systemic lupus erythematosus and/or antiphospholipid syndrome. *Ann. Rheum. Dis.* **76** (3), 476–485 (2017).
8. Craciunas, L. et al. Conventional and modern markers of endometrial receptivity: a systematic review and meta-analysis. *Hum. Reprod. Update*. **25** (2), 202–223 (2019).
9. Albaghdadi, A. J. H. & Kan, F. W. K. Therapeutic potentials of Low-Dose tacrolimus for aberrant endometrial features in polycystic ovary syndrome. *Int. J. Mol. Sci.* ;**22**(6). (2021).
10. Simon, A. & Laufer, N. Assessment and treatment of repeated implantation failure (RIF). *J. Assist. Reprod. Genet.* **29** (11), 1227–1239 (2012).
11. Dubsky, P. et al. BRCA genetic testing and counseling in breast cancer: how do we Meet our patients' needs? *NPJ Breast Cancer*. **10** (1), 77 (2024).
12. Mrozikiewicz, A. E., Ozarowski, M. & Jedrzejczak, P. Biomolecular markers of recurrent implantation Failure-A review. *Int. J. Mol. Sci.* ;**22**(18), 10082 (2021).
13. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43** (7), e47 (2015).
14. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The Sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28** (6), 882–883 (2012).
15. Jin, M., Ni, D., Cai, J. & Yang, J. Identification and validation of immunity- and disulfidptosis-related genes signature for predicting prognosis in ovarian cancer. *Heliyon* **10** (12), e32273 (2024).
16. Wang, P. et al. The changes of gene expression profiling between segmental vitiligo, generalized vitiligo and healthy individual. *J. Dermatol. Sci.* **84** (1), 40–49 (2016).
17. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53** (D1), D672–D7 (2025).
18. Kanehisa, M. Toward Understanding the origin and evolution of cellular organisms. *Protein Sci.* **28** (11), 1947–1951 (2019).
19. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33** (5), 495–502 (2015).
20. Stuart, T. et al. Comprehensive integration of Single-Cell data. *Cell* **177** (7), 1888–1902 (2019). e21.
21. Booeshaghi, A. S. & Pachter, L. Normalization of single-cell RNA-seq counts by log(x + 1)dagger or log(1 + x)dagger. *Bioinformatics* **37** (15), 2223–2224 (2021).
22. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*. **16** (12), 1289–1296 (2019).
23. Khan, K. N., Guo, S. W., Ogawa, K., Fujishita, A. & Mori, T. The role of innate and adaptive immunity in endometriosis. *J. Reprod. Immunol.* **163**, 104242 (2024).
24. Li, F. et al. Potential biomarkers and endometrial immune microenvironment in recurrent implantation failure. *Biomolecules* ;**13**(3), 406 (2023).
25. Ding, H., Xu, H., Zhang, T. & Shi, C. Identification and validation of M2 macrophage-related genes in endometriosis. *Heliyon* **9** (11), e22258 (2023).
26. Park, W. et al. Female reproductive disease, endometriosis: from inflammation to infertility. *Mol. Cells*. **48** (1), 100164 (2025).

27. Senturk, L. M. & Arici, A. Immunology of endometriosis. *J. Reprod. Immunol.* **43** (1), 67–83 (1999).
28. Reis, J. L. et al. The role of NK and T cells in endometriosis. *Int. J. Mol. Sci.* **25**, 18 (2024).
29. Zhou, P. et al. Integrated transcriptomic analysis reveals dysregulated immune infiltration and pro-inflammatory cytokines in the secretory endometrium of recurrent implantation failure patients. *Life Med.* ;**3**(5), 036 (2024).
30. Hogg, C., Horne, A. W. & Greaves, E. Endometriosis-Associated macrophages: origin, phenotype, and function. *Front. Endocrinol. (Lausanne).* **11**, 7 (2020).
31. Song, L., Yang, C., Ji, G. & Hu, R. The role and potential treatment of macrophages in patients with infertility and endometriosis. *J. Reprod. Immunol.* **166**, 104384 (2024).
32. Hu, Y. et al. Gammadelta T cells: origin and fate, subsets, diseases and immunotherapy. *Signal. Transduct. Target. Ther.* **8** (1), 434 (2023).
33. Wang, Z., Zhang, H. & Cheng, Q. PDIA4: the basic characteristics, functions and its potential connection with cancer. *Biomed. Pharmacother.* **122**, 109688 (2020).
34. Lee, C. H. et al. PDIA4, a new Endoplasmic reticulum stress protein, modulates insulin resistance and inflammation in skeletal muscle. *Front. Endocrinol. (Lausanne).* **13**, 1053882 (2022).
35. Helou, L. et al. The piggyBac-derived protein 5 (PGBD5) transposes both the closely and the distantly related piggyBac-like elements Tcr-pble and Ifp2. *J. Mol. Biol.* **433** (7), 166839 (2021).
36. Henssen, A. G. et al. PGBD5 promotes site-specific oncogenic mutations in human tumors. *Nat. Genet.* **49** (7), 1005–1014 (2017).
37. Yamada, M. et al. Childhood cancer mutagenesis caused by transposase-derived PGBD5. *Sci. Adv.* **10** (12), eadn4649 (2024).
38. Hayrabedyan, S., Kyurkchiev, S. & Kehayov, I. FGF-1 and S100A13 possibly contribute to angiogenesis in endometriosis. *J. Reprod. Immunol.* **67** (1–2), 87–101 (2005).
39. Hung, S. W. et al. Pharmaceuticals targeting signaling pathways of endometriosis as potential new medical treatment: A review. *Med. Res. Rev.* **41** (4), 2489–2564 (2021).
40. Hattori, K. et al. Lymphangiogenesis induced by vascular endothelial growth factor receptor 1 signaling contributes to the progression of endometriosis in mice. *J. Pharmacol. Sci.* **143** (4), 255–263 (2020).
41. Liu, R. et al. MUC1 promotes RIF by regulating macrophage ROS-SHP2 signaling pathway to up-regulate inflammatory response and inhibit angiogenesis. *Aging (Albany NY).* **16** (4), 3790–3802 (2024).
42. Xu, M. et al. MiR-3121-3p promotes tumor invasion and metastasis by suppressing Rap1GAP in papillary thyroid cancer in vitro. *Ann. Transl Med.* **8** (19), 1229 (2020).
43. Sukegawa, M. et al. The BCR/ABL tyrosine kinase inhibitor, nilotinib, stimulates expression of IL-1beta in vascular endothelium in association with downregulation of miR-3p. *Leuk. Res.* **58**, 83–90 (2017).
44. Zhang, H., Zhang, C., Zhang, S. & Single-Cell, R. N. A. Transcriptome of the human endometrium reveals epithelial characterizations associated with recurrent implantation failure. *Adv. Biol. (Weinh).* **8** (1), e2300110 (2024).
45. Marder, W. Update on pregnancy complications in systemic lupus erythematosus. *Curr. Opin. Rheumatol.* **31** (6), 650–658 (2019).
46. Marder, W., Littlejohn, E. A. & Somers, E. C. Pregnancy and autoimmune connective tissue diseases. *Best Pract. Res. Clin. Rheumatol.* **30** (1), 63–80 (2016).
47. Radin, M. et al. Pregnancy outcomes in mixed connective tissue disease: a multicentre study. *Rheumatol. (Oxford).* **58** (11), 2000–2008 (2019).

Acknowledgements

Not applicable.

Author contributions

Dongxu Qin and Yongquan Zheng were primarily involved in the design of the experimental approach and data analysis, and in writing the manuscript. Libo Wang was mainly involved in the statistics of the experimental results. Zhenyi Lin was primarily responsible for the correction of the results figures. Yao Yao and Weidong Fei were mainly involved in the revision of the manuscript. Caihong Zheng provided financial support for the experiment and contributed to the revision of the manuscript. All authors reviewed the manuscript.

Funding

This study was supported by the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (Grant No. LHDMZ23H190002).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.Q. or Y.Z.

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

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