scientific reports

Check for updates

OPEN

Machine learning selection of basement membrane-associated genes and development of a predictive model for kidney fibrosis

Ziwei Yuan[1,7], Guangjia Lv[2,7], Xinyan Liu[3]✉, Yanyi Xiao[4,5]✉, Yuanfang Tan[1]✉ & Youyou Zhu[6]✉

This study investigates the role of basement membrane-related genes in kidney fibrosis, a significant factor in the progression of chronic kidney disease that can lead to end-stage renal failure. The authors aim to develop a predictive model using machine learning techniques due to the limitations of existing diagnostic methods, which often lack sensitivity and specificity. Utilizing gene expression data from the GEO database, the researchers applied LASSO, Random Forest, and SVM-RFE methods to identify five pivotal genes: ARID4B, EOMES, KCNJ3, LIF, and STAT1. These genes were analyzed across training and validation datasets, resulting in the development of a Nomogram prediction model. Performance metrics, including the area under the ROC curve (AUC), calibration curves, and decision curve analysis, indicated excellent predictive capabilities with an AUC of 0.923. Experimental validation through qRT-PCR in clinical samples and TGF-β-treated HK-2 cells corroborated the expression patterns identified in silico, showing upregulation of ARID4B, EOMES, LIF, and STAT1, and downregulation of KCNJ3. The findings emphasize the importance of basement membrane-related genes in kidney fibrosis and pave the way for enhanced early diagnosis and targeted therapeutic strategies.

Keywords Basement membrane, Renal fibrosis; machine learning, Predictive models, Gene enrichment analysis; Immune cells.

Abbreviations

(AUC)	Area Under the ROC Curve
(CKD)	Chronic Kidney Disease
(DCA)	Decision Curve Analysis
(DEGs)	Differentially Expressed Basement Membrane-Related Genes
(ESRD)	End-Stage Renal Disease
(ECM)	Excessive Extracellular Matrix
(GEO)	Gene Expression Omnibus
(GSEA)	Gene Set Enrichment Analysis
(PPI)	Protein, Protein Interaction
(RF)	Random Forest
(SVM-RFE)	Support Vector Machine Recursive, Feature Elimination
(SLE)	Systemic Lupus Erythematosus

Kidney fibrosis is a pivotal pathological feature in chronic kidney disease (CKD) progression and a major contributor to end-stage renal disease (ESRD)[1,2]. This process is characterized by the accumulation of extracellular matrix (ECM) components and the aberrant activation of fibrosis-related signaling pathways[3,4], leading to declining renal function. Despite advances in understanding the mechanisms of kidney fibrosis,

[1]Department of Laboratory Medicine, The Third People's Hospital of Ganzhou, 341000 Ganzhou, China. [2]College of Life Sciences, Northeast Forestry University, Harbin 150004, China. [3]Zhanggong District Maternal and Child Health Hospital, Ganzhou 341000, China. [4]Department of Thyroid and Breast Surgery, Wenzhou Central Hospital, Wenzhou 325000, China. [5]Zhejiang Key Laboratory of Intelligent Cancer Biomarker Discovery and Translation, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China. [6]Department of pathology, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai 317000, China. [7]Ziwei Yuan and Guangjia Lv contributed equally. ✉email: 41022187@qq.com; 15967411128@163.com; 2635043620@qq.com; zhyy@enzemed.com

effective early diagnostic tools remain scarce, hindering timely intervention. Identifying molecular markers that can predict disease progression is therefore crucial for preventing further renal damage.

The basement membrane plays a fundamental role in maintaining the structural integrity of renal tissues, supporting tubules and glomeruli, and regulating cellular signaling[5–9]. Alterations in its structure and function are frequently observed in kidney fibrosis, suggesting that basement membrane-related genes may be key drivers in the fibrotic process[10,11]. However, the precise mechanisms and potential clinical applications of these genes, particularly in early diagnosis and prognosis, remain underexplored.

Advances in bioinformatics and machine learning now enable the identification of potential molecular markers through the analysis of high-throughput genomic data[12,13]. Machine learning algorithms excel at handling complex, high-dimensional biological datasets, uncovering hidden disease-related features, and offering novel insights for accurate diagnosis and prognosis[14,15]. Feature selection methods, such as LASSO regression, Random Forest (RF), and Support Vector Machine Recursive Feature Elimination (SVM-RFE), can effectively isolate disease-associated genes, facilitating the creation of highly accurate diagnostic and predictive models.

In this study, we leveraged gene expression profile data and multiple machine learning algorithms to investigate the role of basement membrane-related genes in kidney fibrosis, constructing a Nomogram prediction model based on five key genes (ARID4B, EOMES, KCNJ3, LIF, and STAT1). Using data from kidney fibrosis patients and healthy controls in the GEO database, we identified these key genes via feature selection. To further confirm their function, we used qRT-PCR to assess their expression in TGF-β-treated HK-2 cells and clinical specimens. Experimental results revealed significant upregulation of ARID4B, EOMES, LIF, and STAT1, and downregulation of KCNJ3, consistent with our bioinformatics predictions.

By integrating bioinformatics with experimental validation, this study developed a highly accurate predictive model for kidney fibrosis. Our model provides a valuable molecular tool for early diagnosis and offers insights into the role of basement membrane-related genes in kidney fibrosis, with implications for future research. Additionally, the study identifies potential molecular targets for personalized treatment, underscoring its clinical relevance.

Materials and methods

Data collection and preprocessing

Kidney fibrosis-related gene expression data were retrieved from the Gene Expression Omnibus (GEO) database. The training set (GSE76882) comprised 99 control and 175 fibrosis samples, while the validation set (GSE22459) included 25 control and 40 fibrosis samples. Data quality was assessed using QC methods to identify outliers and batch effects. Take log2 of the expression data and use the limma package for normalization, and average the same genes. Low-expression genes were filtered out to reduce noise, ensuring more reliable results for downstream analysis. Basement membrane-related genes were obtained from the MSigDB database, and the screening criterion was p less than 0.05. Basement membrane-related terms such as "extracellular matrix", "basal lamina", and "collagen" were used to filter genes. To ensure the relevance of selected genes, we cross-referenced the list with known gene signatures and databases to confirm their involvement in basement membrane biology. We selected a final set of 374 genes (Supplementary Table S1), which were found to be significantly enriched in basement membrane-related functions. Data were transformed using log2 and further processed using quantile normalization to ensure uniformity across different samples[16].

Differential expression and PPI network analysis

The "limma" R package was employed for differential expression analysis, identifying basement membrane-related genes that showed significant differences between control and fibrosis samples ($p < 0.05$). The selected differentially expressed genes were used to construct a protein-protein interaction (PPI) network via the STRING database (https://string-db.org) to visualize potential interactions between these genes. Furthermore, we performed GO annotation and KEGG pathway enrichment analysis using the "clusterProfiler," "org.Hs.eg.db," and "DOSE" R packages to explore the biological functions of these genes in the context of kidney fibrosis[16–18].

Diagnostic marker selection using machine learning

Three machine learning approaches were utilized to identify potential diagnostic markers for kidney fibrosis: LASSO logistic regression[19], RF, and SVM-RFE[20]. LASSO was performed using the "glmnet" R package to identify key fibrosis-related genes[21,22]. The RF analysis identified the top 30 significant genes, from which the top 10 were selected for further investigation. Subsequently, SVM-RFE was utilized to refine this gene selection process. The consensus across the three algorithms was used to select candidate diagnostic markers. These markers were validated through qRT-PCR, confirming their expression levels in real-world samples.

Immune cell infiltration and correlation

We used the CIBERSORT algorithm to assess the proportions of 22 immune cell subtypes in each sample[23,24]. Spearman's rank correlation analysis was subsequently conducted to explore the relationship between gene expression and immune cell abundance, shedding light on the role of immune cells in kidney fibrosis. The "ggplot2" R package was used to visualize the correlations between genes and immune cells.

GSEA for functional insights

Gene Set Enrichment Analysis (GSEA) was conducted to investigate the biological functions of the candidate genes. The gene set "c2.cp.kegg.v7.0.symbols.gmt" from the MSigDB database was used as a reference to identify pathway enrichment. GSEA provided deeper insights into the functional roles of the candidate genes in relevant biological processes[25,26].

Cell culture and treatment

HK-2 cells were cultured in DMEM/F12 (Gibco, USA) supplemented with 10% FBS and 1% penicillin/streptomycin at 37 °C in a humidified incubator with 5% CO_2. Upon reaching 80–90% confluence, cells were prepared for treatment. To model renal fibrosis, cells were divided into control and TGF-β treatment groups. The treatment group received 20 ng/mL TGF-β (Reprotech, USA) for 48 h, while the control group received serum-free medium. After treatment, cells were harvested for subsequent analysis.

Clinical specimens

Kidney biopsy specimens were obtained from patients at the First Affiliated Hospital of Wenzhou Medical University. The study included 9 patients diagnosed with renal fibrosis and 5 normal kidney tissue samples from patients. All tissue samples were immediately preserved in RNAlater solution and stored at −80 °C until RNA extraction. This study was approved by the Ethics Committee of the First Affiliated Hospital of Wenzhou Medical University [Approval Number: (2023) No. (106)]. We confirm that this retrospective clinical study was performed in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants.

qRT-PCR

Total RNA was isolated from HK-2 cells (a gift from the First Affiliated Hospital of Wenzhou Medical University) of both the control and TGF-β-treated groups. The RNA purity and concentration were assessed, ensuring a 260/280 ratio between 1.8 and 2.0. After extraction, 1 μg of RNA was immediately reverse transcribed into cDNA using a reverse transcription kit (Vazyme, Nanjing, China). qRT-PCR was carried out using the SYBR Green detection system. The reaction mix consisted of 10 μL SYBR Green Master Mix, 0.5 μL each of forward and reverse primers (10 μM), and 2 μL cDNA, with a total volume of 20 μL. The PCR cycle conditions were: initial denaturation at 95 °C for 5 min, followed by 40 cycles of 95 °C for 15 s, 60 °C for 30 s, and 72 °C for 30 s[27]. Data analysis was conducted using the $2^{-\Delta\Delta Ct}$method, with β-actin serving as the reference gene[28]. Gene expression levels in the TGF-β group were compared to those in the control group to identify genes that showed significant differential expression. Primer sequences used in the RT-PCR assay in Supplementary Table S2.

Statistical analyses

Statistical analysis was conducted using R (version 4.0.3) and GraphPad Prism (version 9.0). Quantitative data were presented as mean ± SEM, with comparisons between groups evaluated using Student's t-test ($p < 0.05$). Differential gene expression was analyzed using "limma," and the Benjamini-Hochberg method was applied to control for false discovery rate (FDR), with an adjusted p-value < 0.05[26]. Correlations between gene expression and immune cell proportions were assessed using Spearman's correlation, with significance defined as $r > 0.3$ and $p < 0.05$[29]. GSEA was considered significant with $p < 0.05$ and $q < 0.25$. Cross-validation was used to evaluate machine learning results to prevent overfitting.

Results

Identification and enrichment analysis of basement membrane-related differentially expressed genes

We first applied the "limma" R package to identify 209 basement membrane-related genes that were differentially expressed between kidney fibrosis and control samples in the GSE76882 dataset (Fig. 1A), and 35 such genes in the GSE22459 dataset (Fig. 1B). By taking the intersection of these two sets, we obtained 29 differentially expressed basement membrane-related genes (DEGs) (Fig. 1C). To visualize potential interactions between these DEGs, a PPI (protein-protein interaction) network was constructed using the STRING online platform (https://string-db.org) (Fig. 1D). Enrichment analysis was then performed to investigate their biological functions. GO enrichment analysis showed that the DEGs were significantly involved in molecular functions such as signaling receptor activator activity, receptor ligand activity, and cytokine receptor binding. They were also enriched in specific cellular components like the sarcolemma, transporter complex, and transmembrane transporter complex. In terms of biological processes, these genes were related to mesenchyme development, mesenchymal cell differentiation, and epithelial to mesenchymal transition (Fig. 1E). KEGG pathway enrichment analysis further revealed that the DEGs were primarily enriched in pathways such as Th17 cell differentiation, cytokine-cytokine receptor interaction, and the AGE-RAGE signaling pathway in diabetic complications (Fig. 1F).

Identification of diagnostic markers for kidney fibrosis

After identifying the basement membrane genes that are differentially expressed in relation to renal fibrosis and understanding their biological roles, we looked into how to use these findings for clinical diagnosis. To achieve this, we applied several machine learning algorithms to pinpoint key genes that could act as diagnostic biomarkers for renal fibrosis. We employed three machine learning algorithms to identify diagnostic markers for kidney fibrosis. The LASSO regression algorithm first identified 17 potential biomarkers (Fig. 2A and B). The RF analysis identified the top 30 significant genes, from which the top 10 were selected for further investigation (Fig. 2C and D). Additionally, SVM-RFE analysis, conducted on the differentially expressed basement membrane-related genes, indicated that 10 genes in total could serve as diagnostic markers (Fig. 2E and F). By taking the intersection of the results from all three algorithms, we identified five common biomarkers: ARID4B, EOMES, KCNJ3, LIF, and STAT1 (Fig. 2G).

Evaluation of the diagnostic performance of five candidate biomarkers for kidney fibrosis

We performed a detailed evaluation of the five candidate genes across both the training and validation cohorts. ROC curves were generated for each of the biomarkers, demonstrating that ARID4B, EOMES, KCNJ3, LIF, and

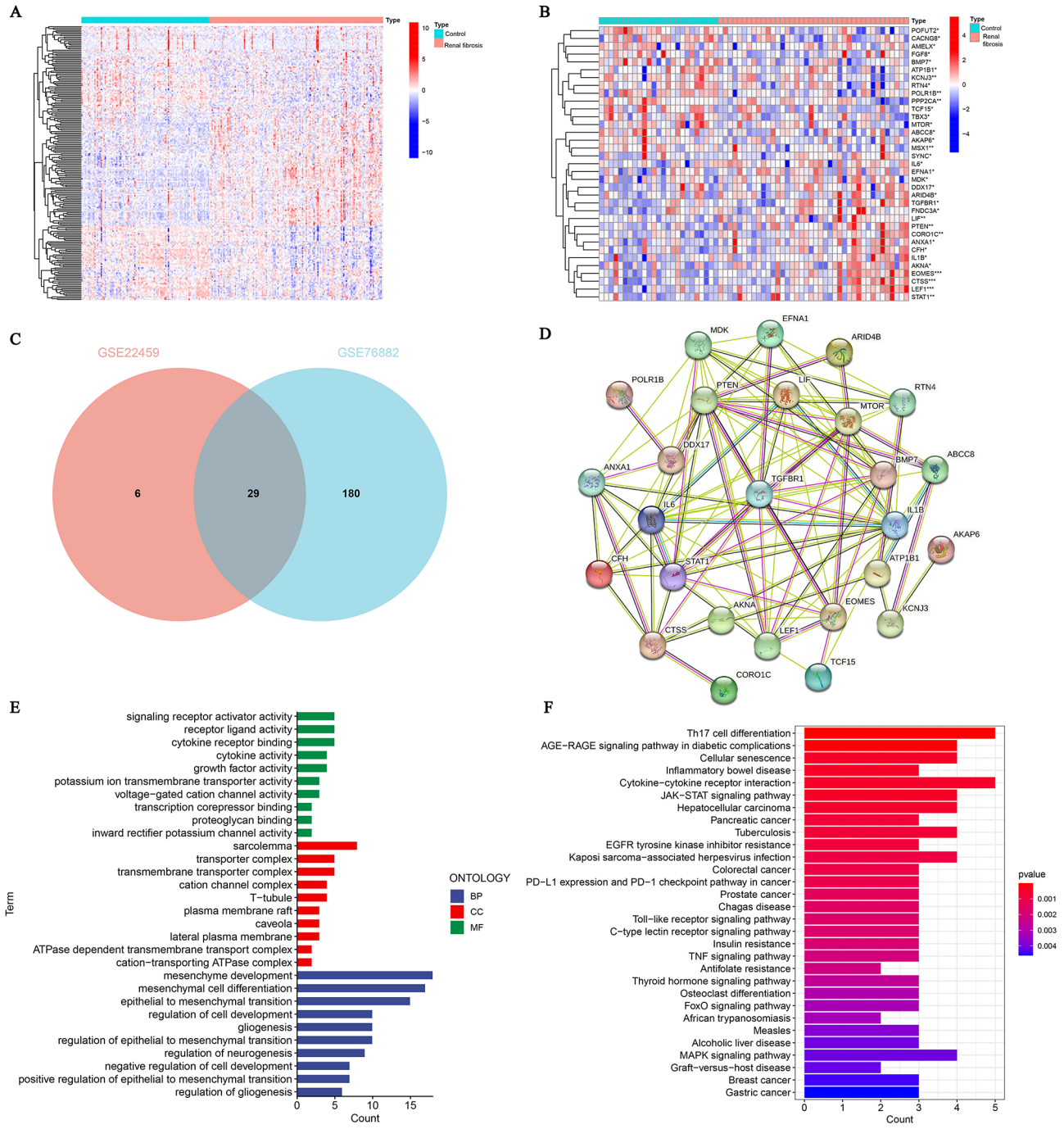

Fig. 1. Identification and enrichment analysis of basement membrane-related differentially expressed genes. **A** Heatmap illustrating the expression patterns of DEGs in the GSE76882 dataset, highlighting differential gene expression across samples. **B** Heatmap showcasing the expression of DEGs identified in the GSE22459 dataset. **C** Venn diagram showing the common DEGs between the GSE76882 and GSE22459 datasets, with a focus on the 29 genes related to the basement membrane. **D** PPI network showing interactions among DEGs and highlighting key hubs. **E** GO enrichment analysis of DEGs, covering biological processes, molecular functions, and cellular components. **F** KEGG pathway enrichment analysis highlighting pathways significantly linked to the DEGs.

STAT1 exhibited robust diagnostic performance. In the training set, the AUC values were 0.668 for ARID4B, 0.851 for EOMES, 0.792 for KCNJ3, 0.776 for LIF, and 0.827 for STAT1 (Fig. 3A). Similarly, in the validation set, the AUC values for ARID4B, EOMES, KCNJ3, LIF, and STAT1 were 0.652, 0.741, 0.696, 0.700, and 0.692, respectively (Fig. 3B). These findings indicate that these five genes possess strong diagnostic potential for detecting kidney fibrosis.

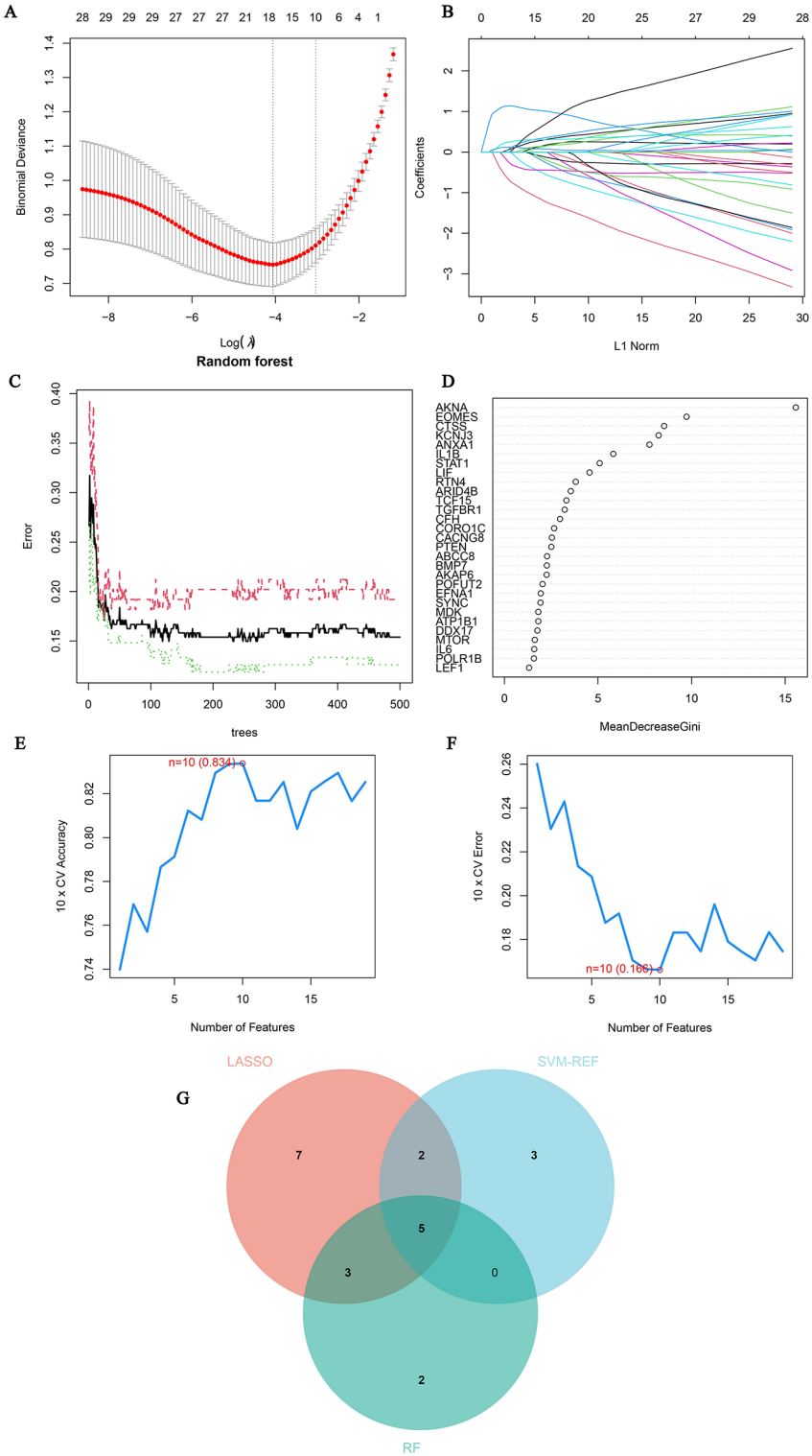

Fig. 2. Identification of diagnostic markers for kidney fibrosis. **A-B** LASSO regression results, indicating the gene coefficients that contributed most to kidney fibrosis classification. **C-D** SVM-RFE results, depicting the most important gene features. **E-F** Variable importance plot from Random Forest, showing the top-ranking genes. **G** A Venn diagram illustrates the overlap among three machine learning models.

Expression levels of five candidate biomarkers in kidney fibrosis

We further analyzed the expression patterns of ARID4B, EOMES, KCNJ3, LIF, and STAT1 in both the training and validation cohorts. As shown in Fig. 4A, the expression of ARID4B, EOMES, LIF, and STAT1 was significantly upregulated in kidney fibrosis compared to the control group, while KCNJ3 was downregulated in the training

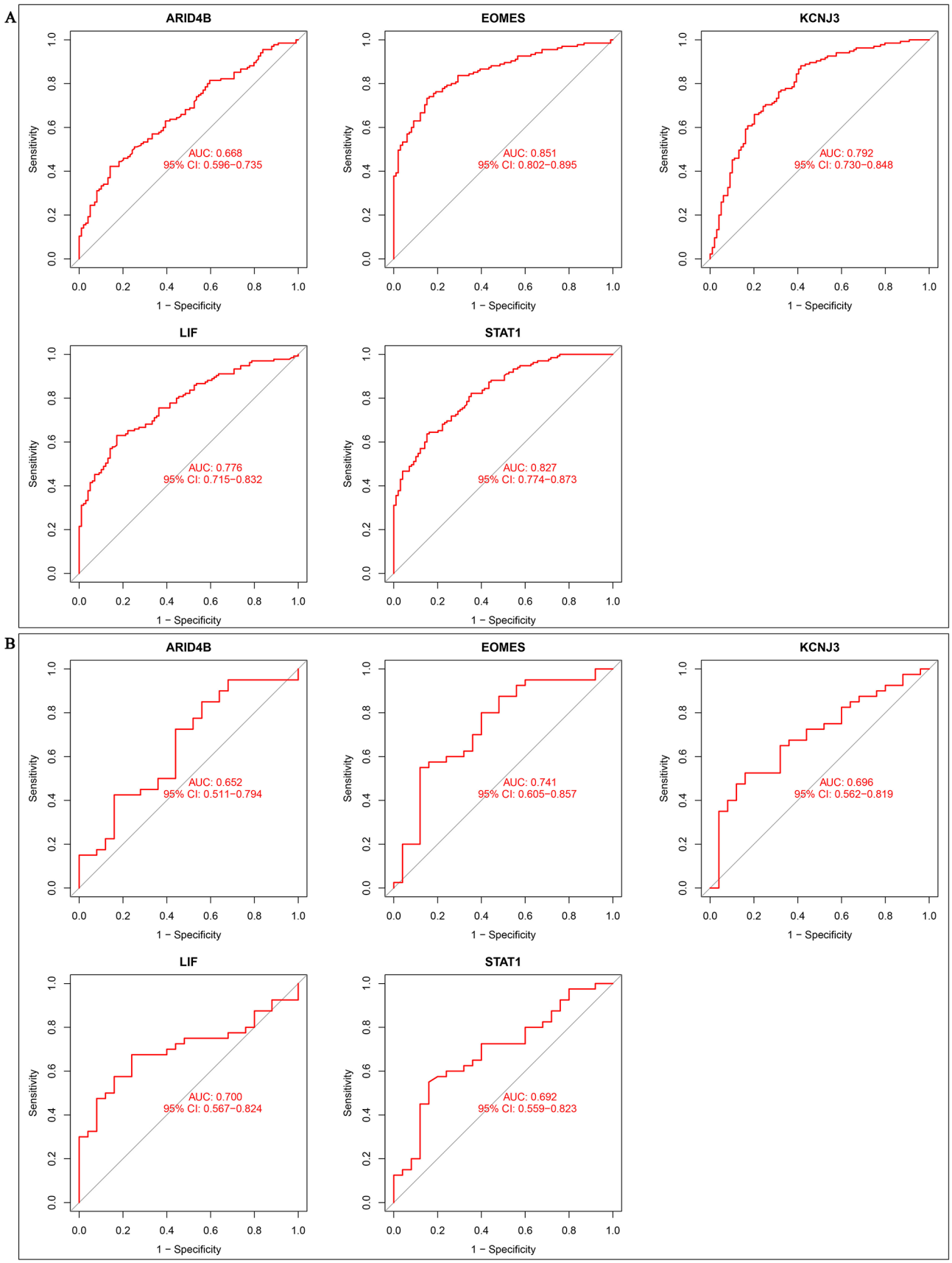

Fig. 3. Evaluation of the diagnostic performance of five candidate biomarkers for kidney fibrosis. **A** ROC curves for the five candidate genes (ARID4B, EOMES, KCNJ3, LIF, and STAT1) in the training set, showing their AUC values and contribution to the model. **B** ROC curves for the validation set, confirming the diagnostic value of the genes with high AUC values, proving their effectiveness for kidney fibrosis diagnosis.

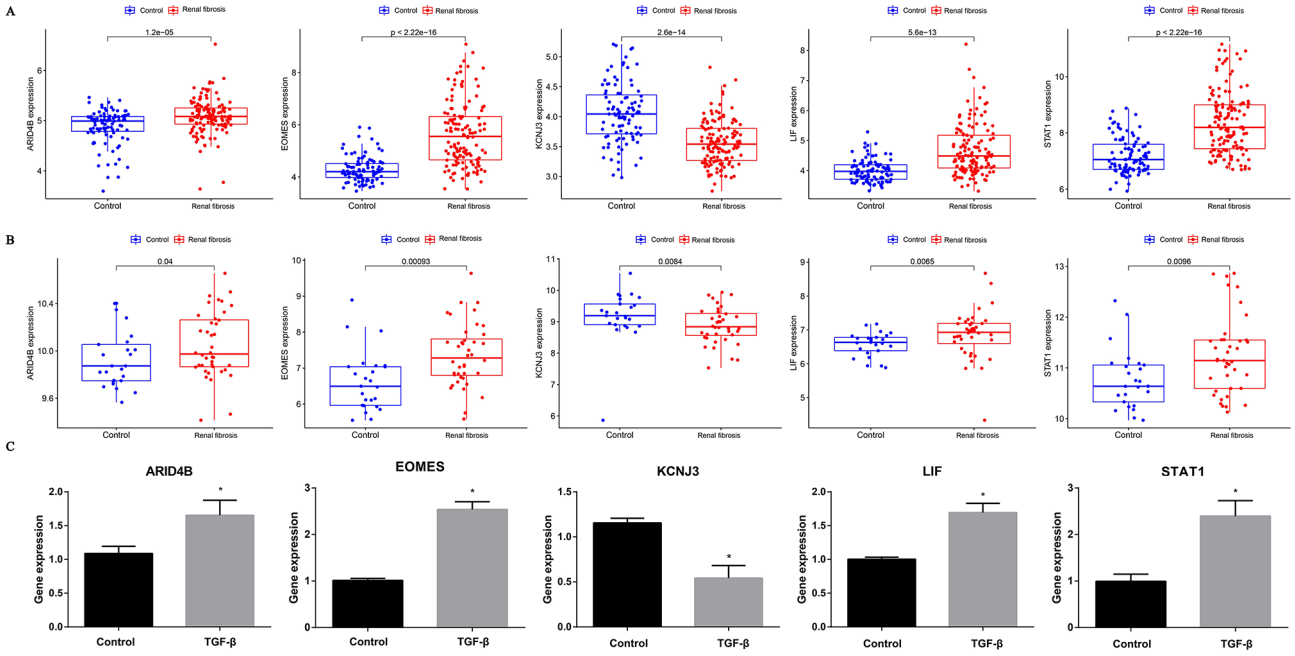

Fig. 4. Expression levels of five candidate biomarkers in kidney fibrosis. **A** Expression levels of ARID4B, EOMES, KCNJ3, LIF, and STAT1 in the training set, showing significant differences between fibrosis and non-fibrosis samples. **B** Expression levels of these genes in the validation set, confirming consistent fibrosis-related patterns. **C** qRT-PCR validation in HK-2 cells with/without TGF-β treatment, showing increased ARID4B, EOMES, LIF, and STAT1, and decreased KCNJ3, consistent with predictions.

set. This trend was also consistent in the validation set (Fig. 4B). Moreover, qRT-PCR validation confirmed that in TGF-β-treated HK-2 cells, the expression of ARID4B, EOMES, LIF, and STAT1 increased, while KCNJ3 expression decreased, aligning with our bioinformatics findings (Fig. 4C). To validate these findings in clinical samples, we performed qRT-PCR analysis using kidney biopsy specimens from patients with renal fibrosis ($n = 9$) and normal kidney tissues ($n = 5$) collected from the First Affiliated Hospital of Wenzhou Medical University. The results showed similar expression patterns, further supporting the potential diagnostic value of these basement membrane-related genes in renal fibrosis (Figure S1). These results from clinical specimens further support the potential diagnostic value of our identified biomarker panel in renal fibrosis.

Correlation between five candidate biomarkers and immune cells

Renal fibrosis triggers various changes in the proportions and functions of immune cells. To understand the involvement of immune cells in this condition, we examined the interactions between five candidate diagnostic genes and 22 types of immune cells. ARID4B showed a positive correlation with activated CD4 memory T cells, eosinophils, and activated mast cells, while it was negatively correlated with M0 macrophages, resting NK cells, and activated NK cells (Fig. 5A). EOMES was positively associated with gamma delta T cells, CD8 T cells, and activated CD4 memory T cells, but negatively associated with resting mast cells, activated NK cells, and resting CD4 memory T cells (Fig. 5B). KCNJ3 exhibited a positive correlation with activated NK cells, resting mast cells, and resting CD4 memory T cells, and a negative correlation with activated CD4 memory T cells, CD8 T cells, and follicular helper T cells (Fig. 5C). LIF was positively correlated with activated CD4 memory T cells, follicular helper T cells, and eosinophils, and negatively correlated with resting mast cells, activated NK cells, and resting CD4 memory T cells (Fig. 5D). STAT1 showed a positive correlation with activated CD4 memory T cells, M1 macrophages, and gamma delta T cells, and a negative correlation with resting mast cells, resting CD4 memory T cells, and M0 macrophages (Fig. 5E).

GSEA analysis of five candidate diagnostic genes

To explore the potential biological functions of the five candidate diagnostic genes, we performed GSEA. The analysis revealed that ARID4B is predominantly involved in ECM receptor interactions, JAK-STAT signaling pathways, and natural killer cell-mediated cytotoxicity (Fig. 6A). EOMES is mainly associated with systemic lupus erythematosus, hematopoietic cell lineage, and Leishmania infection (Fig. 6B). KCNJ3 is primarily involved in peroxisome functions (Fig. 6C). LIF is significantly related to Leishmania infection and NOD-like receptor signaling pathways (Fig. 6D). STAT1 is chiefly associated with systemic lupus erythematosus and Leishmania infection (Fig. 6E).

Development and validation of the nomogram model

To enhance the accuracy of diagnosing kidney fibrosis, we developed a Nomogram prediction model based on five key genes identified in our study (Fig. 7A). This model provides a visual scoring system by integrating

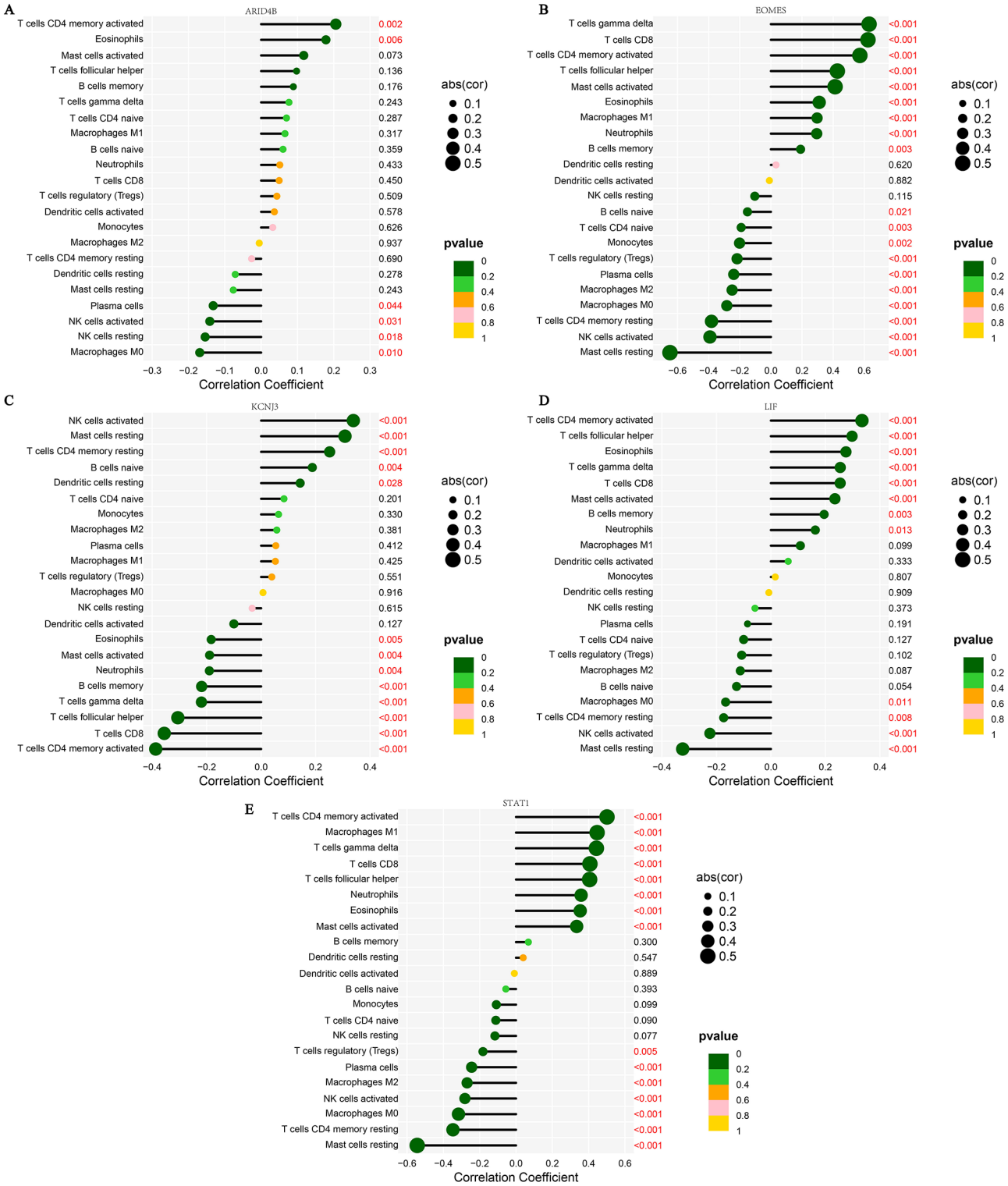

Fig. 5. Correlation between five candidate biomarkers and immune cells. **A-E** Correlation analysis between ARID4B (A), EOMES (B), KCNJ3 (C), LIF (D), STAT1 (E), and 22 immune cell types, showing distinct associations and suggesting roles in immune regulation during kidney fibrosis.

the expression levels of these genes, allowing clinicians to estimate an individual's risk of developing kidney fibrosis more efficiently. It offers personalized diagnostic insights, facilitating quicker decision-making in clinical settings. Performance evaluation demonstrated that the model achieved an impressive AUC of 0.923 (Fig. 7B), signifying its strong diagnostic capability when using the five-gene combination. In addition, we verified the model with an additional external dataset GSE65326 (Figure S2) with an AUC area of 1.00. Furthermore,

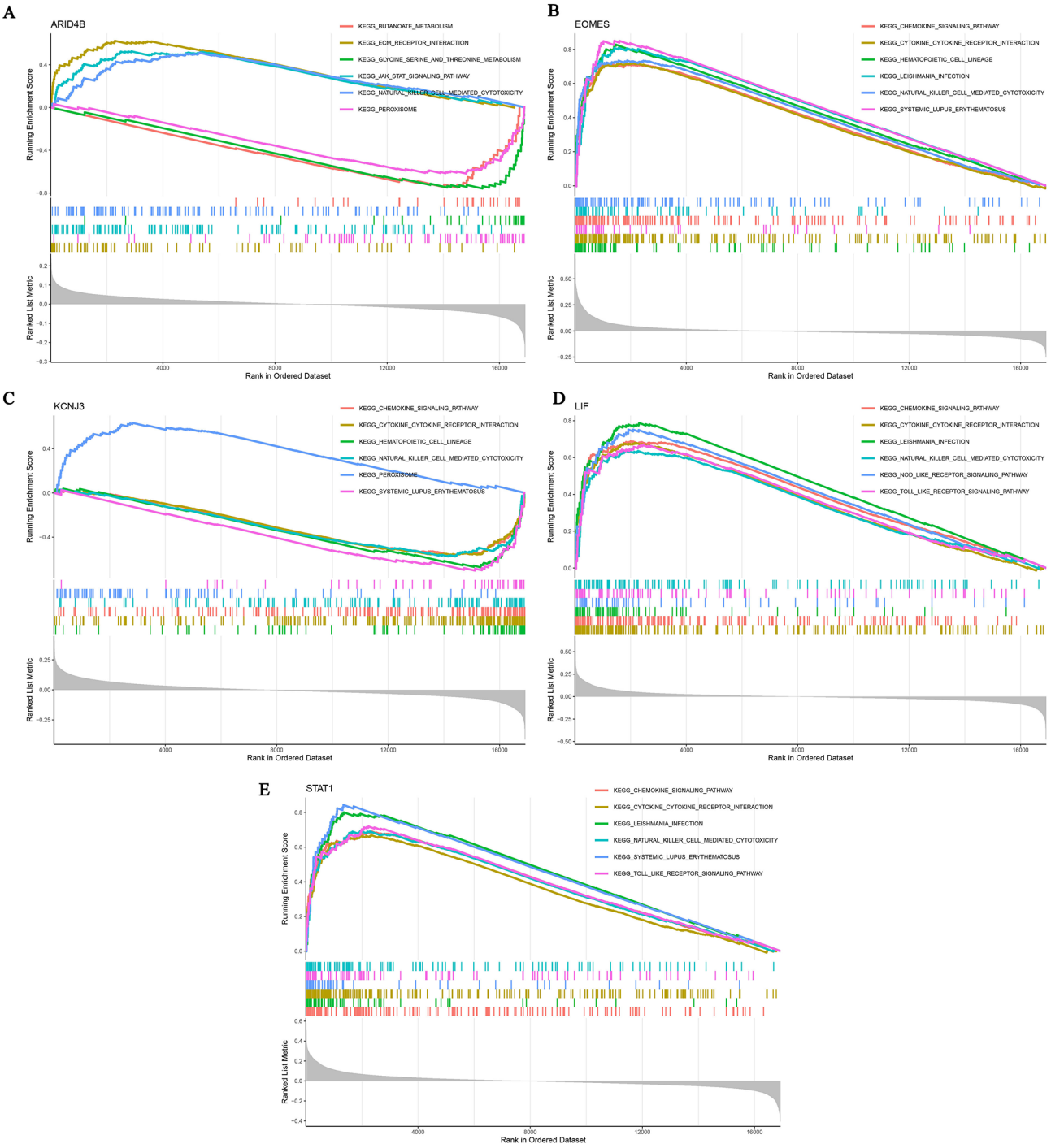

Fig. 6. GSEA analysis of five candidate diagnostic genes. **A-E** GSEA identified the top six significantly enriched signaling pathways for each of the five candidate diagnostic genes: ARID4B (A), EOMES (B), KCNJ3 (C), LIF (D), and STAT1 (E).

calibration curves (Fig. 7C) confirmed the model's excellent fit. To assess the practical utility of the model in a clinical context, we conducted decision curve analysis (DCA, Fig. 7D). The DCA results indicated that across a broad range of threshold probabilities, the net benefit of the Nomogram significantly exceeded that of conventional diagnostic methods, underscoring its clinical advantages. Thus, the Nomogram we developed not only offers robust predictive accuracy but also holds significant promise for future clinical application.

Discussion

This study combined machine learning, bioinformatics, and experimental validation to develop a predictive model for kidney fibrosis. It identified five key genes (ARID4B, EOMES, KCNJ3, LIF, STAT1) and created a

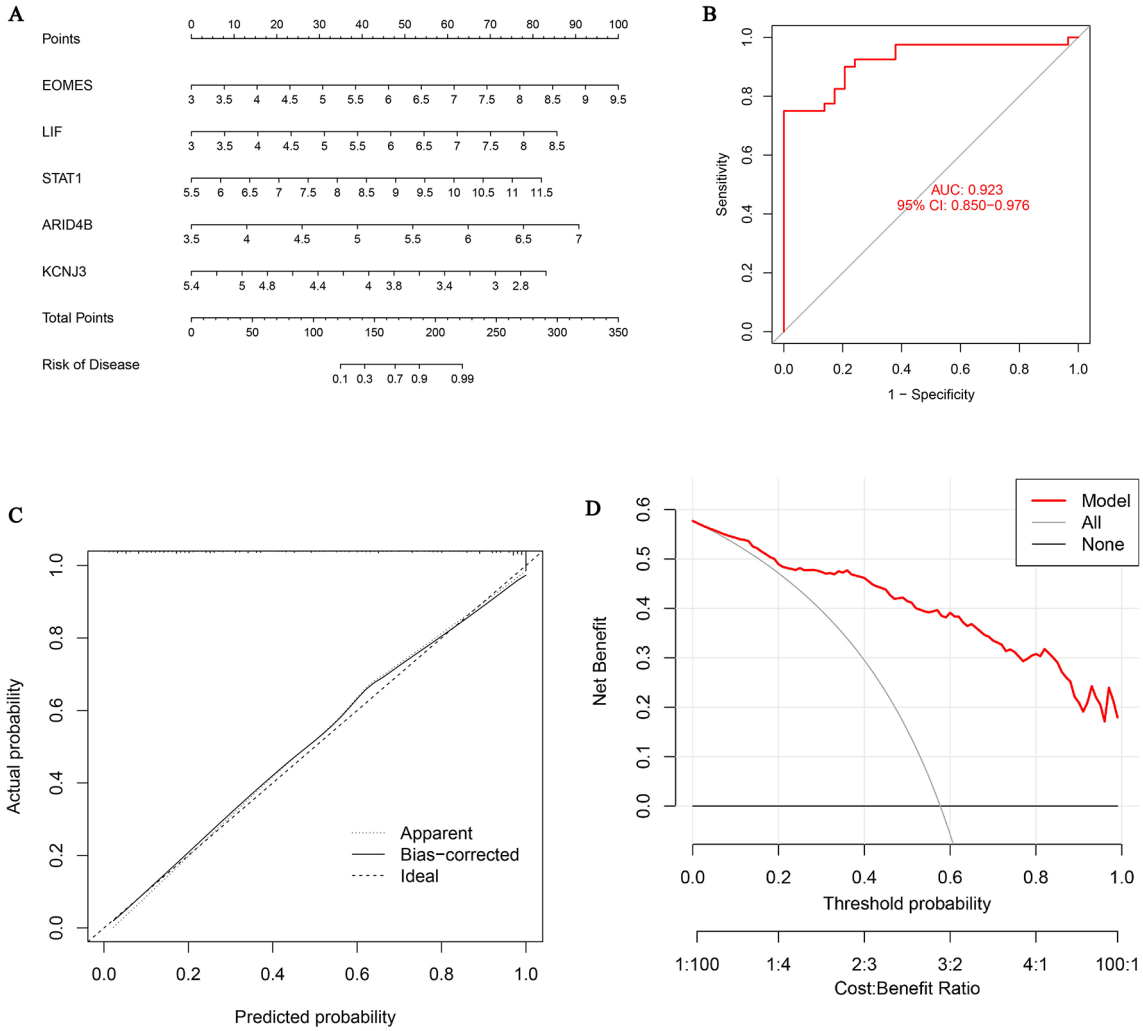

Fig. 7. Development and validation of the nomogram model. **A** Nomogram model for predicting kidney fibrosis risk using ARID4B, EOMES, KCNJ3, LIF, and STAT1. **B** ROC curve of the five-gene signature, showing high diagnostic performance. **C** Calibration curve indicating the accuracy of predicted vs. observed kidney fibrosis probabilities. **D** DCA assessing the clinical utility and net benefit of the nomogram.

Nomogram model for fibrosis risk prediction. qRT-PCR on TGF-β-treated HK-2 cells confirmed increased expression of ARID4B, EOMES, LIF, and STAT1, while KCNJ3 was downregulated, aligning with bioinformatics predictions. These findings emphasize the basement membrane's role in kidney stability and provide new insights into fibrosis progression.

Through gene enrichment and PPI network analyses, our study unveiled the potential roles of the five selected genes in renal fibrosis. ARID4B, a gene typically known for its role in chromatin remodeling and transcriptional regulation, is well-studied in cancer, where it controls cell proliferation and the cell cycle[30–32]. However, its role in fibrosis remains unclear. In renal fibrosis, changes in chromatin and transcription may be pivotal, and our finding of ARID4B's enrichment in the ECM receptor interaction pathway suggests that it could be involved in ECM accumulation, a key factor in fibrosis progression. This points to a novel function of ARID4B in the fibrotic process. EOMES, known for its role in T cell differentiation and regulation of CD8 + T cells and NK cells[33–36], was linked to systemic lupus erythematosus (SLE) and immune regulation in our study. This aligns with the known role of EOMES in autoimmune and inflammatory diseases, such as lupus nephritis. In the context of renal fibrosis, EOMES might regulate immune cell recruitment to the kidney, particularly T cells and NK cells, thereby exacerbating chronic inflammation and promoting fibrosis. KCNJ3, a potassium channel gene involved in ion transport and peroxisomal function[37,38], has not been well-explored in fibrosis. However, altered ion transport can impact cellular metabolism and stress responses, which may indirectly promote fibrosis. By modulating cellular homeostasis and oxidative stress responses, KCNJ3 could indirectly impact the fibrotic process, particularly in the context of renal tubule damage and repair. This offers a new perspective on the connection between ion transport dysfunction and fibrosis, especially in the kidney. LIF is well-established for its role in inflammation and tissue repair[39–41]. Our analysis found LIF associated with the NOD-like receptor signaling pathway, reaffirming its involvement in immune regulation and inflammation-central processes in fibrosis development. LIF's pro-

inflammatory properties may also contribute to ECM deposition and the recruitment of inflammatory cells to the site of injury, supporting its role in the progression of kidney fibrosis. These findings further support the view that LIF is linked to fibrosis in various tissues, including the kidney. STAT1, a major transcription factor in immune and inflammatory responses, has been shown to play a role in kidney inflammation[42,43]. The increased expression of STAT1 in our study is consistent with previous findings that indicate its role in promoting fibrosis through chronic inflammation. STAT1's involvement in the JAK-STAT signaling pathway also underscores its key role in fibrosis. STAT1 activates pro-inflammatory cytokines, driving immune cell infiltration and persistent inflammation in the kidney. This inflammation creates a microenvironment that favors fibrosis, particularly by enhancing the production of ECM proteins and activating fibroblasts. Thus, STAT1's role in fibrosis is likely mediated through its effects on immune cell activation and chronic inflammation. In conclusion, our study provides new mechanistic insights into renal fibrosis by integrating these genes into a predictive machine learning model. While genes such as EOMES, STAT1, and LIF have been previously associated with fibrosis and inflammation, our inclusion of ARID4B and KCNJ3 brings attention to underexplored pathways in fibrosis, particularly the regulation of immune responses and cellular stress.

This study has several important limitations. First, while we used comprehensive public databases, the results need further clinical validation in diverse patient groups to confirm their broader applicability. While qRT-PCR showed consistent gene expression changes, these findings suggest correlations, not direct causal relationships, with renal fibrosis. The molecular mechanisms behind these associations still need further research. To improve diagnostic accuracy, future research should combine multiple methods, like advanced imaging and biochemical profiles. This approach could provide a clearer understanding of how gene expression affects disease progression. Finally, factors such as patient age, comorbidities, medications, and environment could influence gene expression and disease outcomes. These variables should be considered in future clinical studies to strengthen the findings across different patient groups.

In conclusion, this study combines multiple analytical techniques to explore the role of basement membrane genes in kidney fibrosis and develops a highly accurate predictive model. This model provides support for early diagnosis and paves the way for precision treatment in the future.

Data availability

The data sets used and analyzed in this study have been annotated in the manuscript.

Received: 17 October 2024; Accepted: 7 February 2025
Published online: 24 February 2025

References

1. Capasso, A. et al. Summary of the International Conference on Onco-Nephrology: an emerging field in medicine. *Kidney Int.* **96**, 555–567. https://doi.org/10.1016/j.kint.2019.04.043 (2019).
2. Fogo, A. B. Mechanisms of progression of chronic kidney disease. *Pediatr. Nephrol.* **22**, 2011–2022. https://doi.org/10.1007/s00467-007-0524-0 (2007).
3. Sun, Q. et al. Elastin imaging enables noninvasive staging and treatment monitoring of kidney fibrosis. *Sci. Transl Med.* **11** https://doi.org/10.1126/scitranslmed.aat4865 (2019).
4. Zhou, X. et al. Enhancer of Zeste Homolog 2 inhibition attenuates renal fibrosis by maintaining Smad7 and phosphatase and Tensin Homolog expression. *J. Am. Soc. Nephrol.* **27**, 2092–2108. https://doi.org/10.1681/ASN.2015040457 (2016).
5. Inoue, K. et al. Podocyte histone deacetylase activity regulates murine and human glomerular diseases. *J. Clin. Invest.* **129**, 1295–1313. https://doi.org/10.1172/JCI124030 (2019).
6. Sirokmany, G. et al. Peroxidasin-mediated crosslinking of collagen IV is independent of NADPH oxidases. *Redox Biol.* **16**, 314–321. https://doi.org/10.1016/j.redox.2018.03.009 (2018).
7. Mota, C. et al. From tissue and Organ Development to in Vitro models. *Chem. Rev.* **120**, 10547–10607. https://doi.org/10.1021/acs.chemrev.9b00789 (2020).
8. Ishihara, J. et al. Laminin heparin-binding peptides bind to several growth factors and enhance diabetic wound healing. *Nat. Commun.* **9**, 2163. https://doi.org/10.1038/s41467-018-04525-w (2018).
9. Sachs, N. et al. Blood pressure influences end-stage renal disease of Cd151 knockout mice. *J. Clin. Invest.* **122**, 348–358. https://doi.org/10.1172/JCI58878 (2012).
10. Zhao, X., Chen, J., Sun, H., Zhang, Y. & Zou, D. New insights into fibrosis from the ECM degradation perspective: the macrophage-MMP-ECM interaction. *Cell. Biosci.* **12**, 117. https://doi.org/10.1186/s13578-022-00856-w (2022).
11. Wang, Y. et al. COL4A3 gene variants and Diabetic kidney disease in MODY. *Clin. J. Am. Soc. Nephrol.* **13**, 1162–1171. https://doi.org/10.2215/CJN.09100817 (2018).
12. Le, N. Q. K., Li, W. & Cao, Y. Sequence-based prediction model of protein crystallization propensity using machine learning and two-level feature selection. *Brief. Bioinform.* **24** https://doi.org/10.1093/bib/bbad319 (2023).
13. Kha, Q. H., Le, V. H., Hung, T. N. K., Nguyen, N. T. K. & Le, N. Q. K. Development and validation of an explainable machine learning-based prediction model for drug-food interactions from Chemical structures. *Sens. (Basel).* 23. https://doi.org/10.3390/s23083962 (2023).
14. Alsini, R. et al. Deep-VEGF: deep stacked ensemble model for prediction of vascular endothelial growth factor by concatenating gated recurrent unit with two-dimensional convolutional neural network. *J. Biomol. Struct. Dyn.* 1–11. https://doi.org/10.1080/07391102.2024.2323144 (2024).
15. Almusallam, N. et al. Multi-headed ensemble residual CNN: a powerful tool for fibroblast growth factor prediction. *Results Eng.* **24**, 103348. https://doi.org/10.1016/j.rineng.2024.103348 (2024).
16. Yuan, Z. et al. Investigating the impact of inflammatory response-related genes on renal fibrosis diagnosis: a machine learning-based study with experimental validation. *J. Biomol. Struct. Dyn.* 1–13. https://doi.org/10.1080/07391102.2024.2317992 (2024).
17. Ba, R. et al. FOXG1 drives transcriptomic networks to specify principal neuron subtypes during the development of the medial pallium. *Sci. Adv.* **9**, eade2441. https://doi.org/10.1126/sciadv.ade2441 (2023).
18. Liang, L. et al. Mutation-associated transcripts reconstruct the prognostic features of oral tongue squamous cell carcinoma. *Int. J. Oral Sci.* **15**, 1. https://doi.org/10.1038/s41368-022-00210-3 (2023).
19. Ali, F. et al. DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. *J. Comput. Aided Mol. Des.* **33**, 645–658. https://doi.org/10.1007/s10822-019-00207-x (2019).

20. Ali, F. et al. DBPPred-PDSD: machine learning approach for prediction of DNA-binding proteins using Discrete Wavelet transform and optimized integrated features space. *Chemom Intell. Lab. Syst.* **182**, 21–30. https://doi.org/10.1016/j.chemolab.2018.08.013 (2018).
21. Dawkins, J. J. et al. Gut metabolites predict Clostridioides difficile recurrence. *Microbiome* **10**, 87. https://doi.org/10.1186/s40168-022-01284-1 (2022).
22. Zhang, Q. et al. Gammaproteobacteria, a core taxon in the guts of soil fauna, are potential responders to environmental concentrations of soil pollutants. *Microbiome* **9**, 196. https://doi.org/10.1186/s40168-021-01150-6 (2021).
23. Zhang, S. L. et al. Pectin supplement significantly enhanced the anti-PD-1 efficacy in tumor-bearing mice humanized with gut microbiota from patients with colorectal cancer. *Theranostics* **11**, 4155–4170. https://doi.org/10.7150/thno.54476 (2021).
24. Tekpli, X. et al. An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat. Commun.* **10**, 5499. https://doi.org/10.1038/s41467-019-13329-5 (2019).
25. Ye, H. et al. The hepatic Microenvironment uniquely protects leukemia cells through induction of growth and survival pathways mediated by LIPG. *Cancer Discov.* **11**, 500–519. https://doi.org/10.1158/2159-8290.CD-20-0318 (2021).
26. Chandler, B. C. et al. TTK inhibition radiosensitizes basal-like breast cancer through impaired homologous recombination. *J. Clin. Invest.* **130**, 958–973. https://doi.org/10.1172/JCI130435 (2020).
27. Bustin, S. A. & Nolan, T. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J. Biomol. Tech.* **15**, 155–166 (2004).
28. Radonic, A. et al. Guideline to reference gene selection for quantitative real-time PCR. *Biochem. Biophys. Res. Commun.* **313**, 856–862. https://doi.org/10.1016/j.bbrc.2003.11.177 (2004).
29. Wang, Y. F. et al. The gut microbiota-inflammation-brain axis in end-stage renal disease: perspectives from default mode network. *Theranostics* **9**, 8171–8181. https://doi.org/10.7150/thno.35387 (2019).
30. Young, I. C. et al. Differentiation of fetal hematopoietic stem cells requires ARID4B to restrict autocrine KITLG/KIT-Src signaling. *Cell. Rep.* **37**, 110036. https://doi.org/10.1016/j.celrep.2021.110036 (2021).
31. Wu, R. C. et al. Identification of the PTEN-ARID4B-PI3K pathway reveals the dependency on ARID4B by PTEN-deficient prostate cancer. *Nat. Commun.* **10**, 4332. https://doi.org/10.1038/s41467-019-12184-8 (2019).
32. Wu, M. Y., Eldin, K. W. & Beaudet, A. L. Identification of chromatin remodeling genes Arid4a and Arid4b as leukemia suppressor genes. *J. Natl. Cancer Inst.* **100**, 1247–1259. https://doi.org/10.1093/jnci/djn253 (2008).
33. Lazarevic, V., Glimcher, L. H. & Lord, G. M. T-bet: a bridge between innate and adaptive immunity. *Nat. Rev. Immunol.* **13**, 777–789. https://doi.org/10.1038/nri3536 (2013).
34. Curran, M. A. et al. Systemic 4-1BB activation induces a novel T cell phenotype driven by high expression of Eomesodermin. *J. Exp. Med.* **210**, 743–755. https://doi.org/10.1084/jem.20121190 (2013).
35. Jiang, X., Chen, Y., Peng, H. & Tian, Z. Single line or parallel lines: NK cell differentiation driven by T-bet and Eomes. *Cell. Mol. Immunol.* **9**, 193–194. https://doi.org/10.1038/cmi.2012.8 (2012).
36. Pearce, E. L. et al. Control of effector CD8+T cell function by the transcription factor eomesodermin. *Science* **302**, 1041–1043. https://doi.org/10.1126/science.1090148 (2003).
37. Pijnappels, D. A. et al. Resynchronization of separated rat cardiomyocyte fields with genetically modified human ventricular scar fibroblasts. *Circulation* **116**, 2018–2028. https://doi.org/10.1161/CIRCULATIONAHA.107.712935 (2007).
38. Yamada, M., Inanobe, A. & Kurachi, Y. G protein regulation of potassium ion channels. *Pharmacol. Rev.* **50**, 723–760 (1998).
39. Yu, H. et al. LIF negatively regulates tumour-suppressor p53 through Stat3/ID1/MDM2 in colorectal cancers. *Nat. Commun.* **5**, 5218. https://doi.org/10.1038/ncomms6218 (2014).
40. Dallagi, A. et al. The activating effect of IFN-gamma on monocytes/macrophages is regulated by the LIF-trophoblast-IL-10 axis via Stat1 inhibition and Stat3 activation. *Cell. Mol. Immunol.* **12**, 326–341. https://doi.org/10.1038/cmi.2014.50 (2015).
41. Hu, W., Feng, Z., Teresky, A. K. & Levine, A. J. p53 regulates maternal reproduction through LIF. *Nature* **450**, 721–724. https://doi.org/10.1038/nature05993 (2007).
42. Fu, Y. et al. The STAT1/HMGB1/NF-kappaB pathway in chronic inflammation and kidney injury after cisplatin exposure. *Theranostics* **13**, 2757–2773. https://doi.org/10.7150/thno.81406 (2023).
43. Ouyang, W., Rutz, S., Crellin, N. K., Valdez, P. A. & Hymowitz, S. G. Regulation and functions of the IL-10 family of cytokines in inflammation and disease. *Annu. Rev. Immunol.* **29**, 71–109. https://doi.org/10.1146/annurev-immunol-031210-101312 (2011).

Acknowledgements

The authors would like to thank all individuals and organizations for their valuable support and contributions to this study.

Author contributions

ZW-Y: Conceptualization, Data curation, Writing-Original Draft.GJ-L: Methodology, Formal Analysis, Writing-Review & Editing.YY-X: Funding and Formal AnalysisXY-L and YF-T: Visualization, Supervision, Writing-Review & Editing.YY-Z: Project Administration, Supervision.

No funding was provided for this study.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-89733-3.

Correspondence and requests for materials should be addressed to X.L., Y.X., Y.T. or Y.Z.