Iranian Journal of Basic Medical Sciences
ijbms.mums.ac.ir

A lipid metabolism–based gene signature defines diagnosis and molecular heterogeneity in diabetic nephropathy

Qingbu Mei [1#], Xinyu Liu [2#], Jing Xu [2], Shuang Li [2], Tao Wang [3], Hui Yuan [2, 4*]

[1] School of Basic Medical Sciences, Qiqihar Medical University, Qiqihar, China
[2] School of Basic Medical Sciences, Mudanjiang Medical University, Mudanjiang, China
[3] Department of Anesthesiology, Mudanjiang Traditional Chinese Medicine Hospital, Mudanjiang 157011, China
[4] School of Stomatology, Mudanjiang Medical University, Mudanjiang, China

ARTICLE INFO

Article type:
Original

Article history:
Received: Feb 2, 2026
Accepted: Jun 15, 2026

Keywords:
Diabetic nephropathy
Immune infiltration
Lipid metabolism
Machine learning
Molecular heterogeneity

ABSTRACT

Objective(s): To identify lipid metabolism-related biomarkers of diabetic nephropathy (DN), develop and validate a diagnostic model, and further explore immune infiltration patterns and molecular subtypes of DN.

Materials and Methods: Multiple public transcriptomic datasets were integrated to identify lipid metabolism-related biomarkers for DN using differential expression analysis, weighted gene co-expression network analysis, and machine learning. A diagnostic model was then constructed from these biomarkers, and its performance was evaluated using ROC curve and nomogram analyses, followed by validation in an independent cohort. In addition, immune cell infiltration was assessed with the CIBERSORT algorithm, and consensus clustering identified molecular subtypes of DN. Finally, the findings were validated in a model of type 2 diabetes.

Results: A total of 28 lipid metabolism–related differentially expressed genes were identified, and five hub biomarkers (G0S2, PTGDS, CA2, CYP27B1, and HSD17B14) were screened and demonstrated favorable diagnostic performance in both training and validation cohorts. Functional enrichment analysis revealed that these genes were mainly involved in fatty acid metabolism and PPAR signaling pathways. Immune infiltration analysis revealed significant correlations between dysregulated lipid metabolism and immune cell infiltration in DN. Consensus clustering identified distinct molecular subtypes of DN with different lipid metabolic characteristics. Critically, bioinformatic predictions were validated in db/db mice, showing significantly elevated mRNA and protein levels of all five biomarkers in DN kidneys.

Conclusion: Our findings reveal a validated lipid-metabolism-based diagnostic signature with potential for early detection, molecular stratification, and mechanistic investigation.

▶ Please cite this article as:
Mei Q, Liu X, Xu J, Li S, Wang T, Yuan H. A lipid metabolism–based gene signature defines diagnosis and molecular heterogeneity in diabetic nephropathy . Iran J Basic Med Sci 2026; 29:

Introduction

Diabetic nephropathy (DN) is a major microvascular complication of diabetes, and DN is also the leading cause of chronic kidney disease (CKD) and end-stage renal disease (ESRD) worldwide (1). In recent years, numerous studies have shown that multiple mechanisms contribute to the pathogenesis of DN, including metabolic disorders (2, 3), hemodynamic changes, inflammatory responses, and fibrosis (2). A hallmark of DN is the accumulation of collagen, fibronectin, and other extracellular matrix (ECM) proteins. This accumulation leads to tubulointerstitial fibrosis, hypertrophy and dilation of the glomerular mesangium, thickening of the glomerular basement membrane, loss of podocyte processes, and inflammation caused by infiltration by monocytes and macrophages (3-5).

Lipid metabolism imbalance is regarded as one of the important pathogenic mechanisms of DN. Hyperglycemia and insulin resistance in the diabetic state promote abnormal lipid deposition in renal cells, including podocytes and renal tubular epithelial cells (6, 7). The abnormal accumulation of these lipids and their toxic intermediate products can trigger oxidative stress, endoplasmic reticulum (ER) stress, mitochondrial dysfunction, and a vicious cycle of pro-inflammatory and pro-fibrotic signaling cascades (8-10). Multiple studies have confirmed that the imbalance in the activities of AMP-activated protein kinase (AMPK), mammalian target of rapamycin (mTOR), and peroxisome proliferator-activated receptor (PPAR) signaling pathways plays a significant role in DN (11-13). Although there have been advancements in blood glucose control and existing kidney protection treatments, a significant number of patients still progress to irreversible renal failure. However, commonly used clinical diagnostic indicators, such as estimated glomerular filtration rate (eGFR) and urinary albumin excretion rate, typically change only after significant damage to kidney structure and function has occurred (14,

*_**Corresponding author:**_ Hui Yuan. School of Basic Medical Sciences and School of Stomatology, Mudanjiang Medical University, No. 3 Tongxiang Street, Aimin District, Mudanjiang, Heilongjiang Province, China. Email: huiyuan1982@126.com
# These authors contributed equally to this work

15). Currently reported lipid metabolism-related molecular markers still exhibit significant deficiencies in stability, diagnostic efficacy, and clinical translation potential. There is an urgent need in clinical practice to identify new molecular markers with clear pathophysiological significance and high diagnostic efficacy to enable early diagnosis, accurate risk stratification, and targeted therapeutic strategies. This issue represents a critical and pressing challenge in current research on DN.

High-throughput transcriptomics provides an effective means to discover potential molecular mechanisms and candidate markers in DN systematically. However, many existing studies remain limited to a single dataset or to traditional differential gene expression analysis, and they lack systematic integration of gene co-expression networks and their functional relationships. This lack of integration may lead to the omission of key regulatory modules and poor reproducibility of findings (16). Machine learning and systems biology methods have provided novel approaches for high-dimensional data analysis. LASSO regression can reduce feature redundancy and select key variables, while the RF algorithm performs exceptionally well at capturing nonlinear relationships among variables. By integrating WGCNA with these methods, core network modules can be identified, thereby enhancing the stability and biological interpretability of candidate markers (17). Furthermore, some studies have attempted to use these algorithms to screen and validate candidate biomarkers for DN and to investigate their correlations with immune cell infiltration, thereby providing insights into the interaction between immune metabolism and disease progression (18, 19).

In this study, we proposed the hypothesis that lipid metabolic reprogramming is the key pathogenic mechanism of DN and that the expression profiles of lipid metabolism-related genes (LMGs) can serve as efficient molecular diagnostic markers that reflect immune microenvironment remodeling. We integrated multiple GEO datasets and conducted differential expression analysis and WGCNA network modeling, followed by cross-screening with lipid metabolism-related gene sets. We then combined LASSO and RF algorithms to identify candidate LMGs, thereby constructing a lipid metabolism feature model consisting of five genes with high diagnostic efficacy. Furthermore, we evaluated the correlation between this feature and renal immune cell infiltration. We also conducted molecular classification of DN based on this feature to reveal its potential immune-metabolic heterogeneity; these findings were subsequently validated using a DN animal model. This provides a theoretical basis and potential molecular targets for the early diagnosis, risk stratification, and individualized therapies of DN (Figure 1).

Materials and Methods

Differentially expressed gene identification

Raw expression data from four GEO datasets (GSE47185, GSE96804, GSE104948, and GSE99340) were retrieved using the GEOquery package. To ensure cross-platform comparability, we first employed the sva package to correct for batch effects and normalize expression values prior to integrative analysis. Differential expression analysis was then performed using the limma package in R (20), with genes satisfying |log2 fold change| >0.585 designated as significantly differentially expressed genes (DEGs). Volcano plots and heatmaps were generated using ggplot2 and ComplexHeatmap, respectively, to visualize the distribution of DEGs.

Weighted gene co-expression network construction

A weighted gene co-expression network was constructed from raw expression profiles using the WGCNA package (21, 22). The optimal soft-thresholding power was determined via the "Pick Soft Threshold" function to achieve scale-free topology. Pairwise topological overlap was calculated using the topological overlap matrix (TOM) similarity function, and the resulting TOM was then used as input for module detection. Co-expression modules were identified through dynamic tree cutting (cutree Dynamic) with a specified minimum module size. Module-trait relationships were

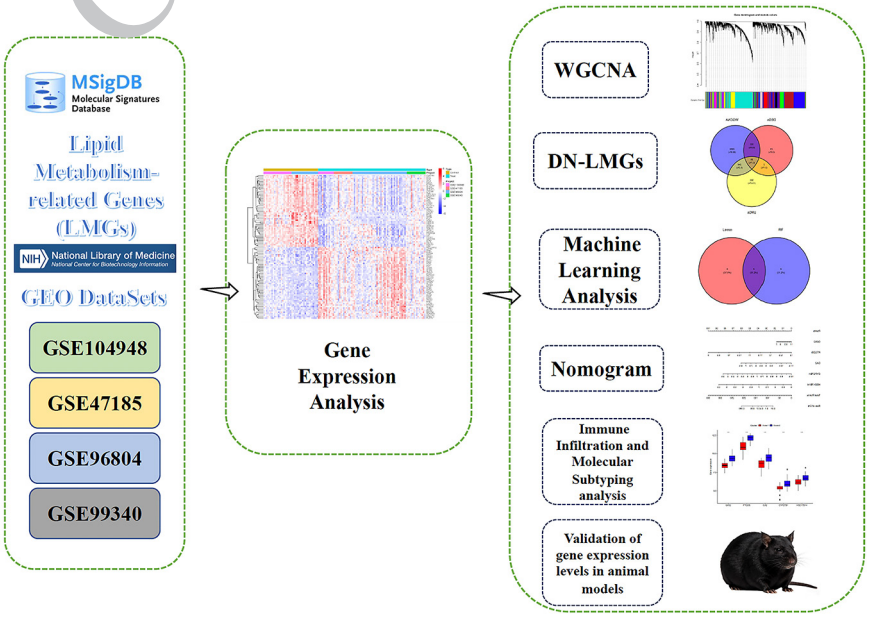

Figure 1. Schematic illustration of the overall general workflow of this study

quantified by calculating the Pearson correlation between module eigengenes and clinical phenotypes. Significance was assessed using the "cor *P*-value Student" function. Modules that demonstrated a significant association with DN were prioritized for subsequent analyses.

Functional enrichment analysis

Overlapping DEGs were subjected to Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis using the clusterProfiler package (23). Statistical significance was defined as *P*-value<0.05 and false discovery rate (FDR)<0.05, following the Benjamini-Hochberg multiple testing correction. Only terms satisfying both thresholds were considered biologically significant.

Machine learning-based biomarker selection

Candidate biomarkers were identified from 28 hub genes using an integrated LASSO-RF framework implemented in R v4.3.2. LASSO regression was performed with the glmnet package (v4.1-8) using 10-fold cross-validation to determine the optimal λ (λ min) that minimized the mean cross-validated error (24). Genes retaining non-zero coefficients at λ min were selected. Subsequently, RF analysis was conducted using the random Forest package (v4.7-1.1) with 5,000 decision trees. Feature importance was ranked by mean decrease in Gini index (MDG) following grid search optimization of the mtry parameter, which specifies the number of variables randomly sampled as candidates at each split. The final five-gene signature (G0S2, PTGDS, CA2, CYP27B1, and HSD17B14) comprised the intersection of genes selected by both algorithms.

Diagnostic performance evaluation

The prognostic accuracy of the five-gene signature was evaluated by receiver operating characteristic (ROC) curve analysis using the pROC package (v1.18.5)(25). Area under the curve (AUC) values and 95 % confidence intervals (CIs) were derived from 2,000 bootstrap replicates, with optimal cutoffs determined by maximizing Youden's index. The model's accuracy was validated using GEO datasets (GSE30528 and GSE104954). A nomogram integrating the five-gene signature was developed via multivariable logistic regression (rms v6.7-0) to predict disease status. The accuracy and clinical net benefit of the Nomogram were assessed through calibration curves and decision curve analysis (DCA).

Immune infiltration and molecular subtyping

Immune cell composition was estimated using the CIBERSORT algorithm (v1.06) with the LM22 reference matrix (1,000 permutations). Differences in immune cell infiltration between DN and control samples were assessed by the Wilcoxon rank-sum test. Correlations among 22 immune cell subsets were evaluated using Spearman's rank correlation. Associations between the five-gene signature and immune cell abundances were examined similarly. Molecular subtypes were identified using Consensus Cluster Plus (v1.60.0) based on expression of the five core genes. Cluster stability was evaluated across K=2-6 using 80% resampling and 1,000 iterations. The optimal number of clusters was determined using consensus matrices, cumulative distribution function (CDF) curves, and delta area plots. Subtype separation was confirmed by principal component analysis (PCA), with gene expression patterns visualized by two-dimensional hierarchical clustering.

Experimental animal model

Male C57BL/6 and db/db mice (20±2 g) were obtained from GemPharmatech Co., Ltd. The animals were assigned to two groups: the wild-type (WT) group and the type 2 diabetes mellitus (T2DM) group. After one week of acclimatization under SPF conditions, 7-week-old db/db mice with random blood glucose levels ≥16.7 mmol/l were selected as the T2DM model. All animal experimental protocols were approved by the Institutional Animal Care and Use Committee of Mudanjiang Medical University (Approval No. IACUC-20250824-172) and were conducted in accordance with the Guide for the Care and Use of Laboratory Animals published by the National Institutes of Health (NIH).

Serum measurements

Blood glucose levels and glucose tolerance were evaluated using tail vein blood collected from mice with a glucometer (ACCU-CHEK, Roche, Germany). At week 12, blood samples were collected from the medial canthus vein and centrifuged to obtain serum for biochemical analyses. Serum insulin concentrations were determined using a commercial ELISA kit (Beyotime, Shanghai, China). Serum triacylglycerol (TG) and total cholesterol (TC) levels were measured using standard biochemical assay kits (Beyotime, Shanghai, China). Serum creatinine (SCR) and blood urea nitrogen (BUN) concentrations were quantified using assay kits from Bioengineering Institute (Jiancheng, Jiangsu, China), according to the manufacturers' instructions.

Doppler ultrasonography analysis

A Doppler ultrasonography machine was used to noninvasively assess renal morphology and hemodynamics (Mylab Delta-vet, Esaote, Italy). All the mice were anesthetized with 2% isoflurane for ultrasonography. Renal size, cortical thickness, and renal arterial blood flow parameters were recorded to evaluate renal perfusion and functional alterations.

Transmission electron microscopy (TEM)

Kidney tissues were fixed in 2.5 % glutaraldehyde at 4 ˚C overnight, followed by post-fixation in 1 % osmium tetroxide for 2 h. The samples were then dehydrated through a graded ethanol series, embedded in epoxy resin, and examined using a transmission electron microscope (H-7650; Hitachi, Japan).

Histology analysis

The kidneys were rapidly excised and rinsed with pre-cooled phosphate-buffered saline (PBS). The tissues were fixed in 10% buffered paraformaldehyde, embedded in paraffin, and sectioned at a thickness of 4 μm for morphological evaluation. Hematoxylin and eosin (H&E) staining and Masson's trichrome staining were performed according to standard protocols. Stained sections were examined using a computer-assisted color image analysis system (Leica DM2700M microscope, Germany).

Quantitative real-time polymerase chain reaction (qRT-PCR) analysis

Total RNA was extracted from kidney tissue using the

Qiagen RNeasy Mini Kit. One microgram of RNA was reverse-transcribed to cDNA, and 2 µl (0.4 µg) of the resulting cDNA was used for PCR with gene-specific primers (Table 1). mRNA expression of G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 was measured using the miScript SYBR Green PCR Kit (Qiagen) on a CFX96 Touch Real-Time PCR System (Bio-Rad, MA, USA). Relative transcript levels were quantified by the ΔΔCT method, with GAPDH serving as the internal reference.

Western blot analysis

Mouse kidney tissues were homogenized in RIPA buffer (1:1000 PMSF) at 4 ˚C, centrifuged at 13,500 rpm for 25 min, and protein concentrations were measured by BCA assay. Lysates were separated by SDS-PAGE, transferred to PVDF membranes, blocked with 5 % milk, and incubated overnight with primary antibodies (G0S2, PTGDS, CA2, CYP27B1, HSD17B14, β-actin, 1:1000). After 1 h with HRP-conjugated secondary antibodies (1:10,000), proteins were detected using ECL and a fluorescent imaging system.

Statistical analyses

All experiments were performed at least three times independently. Data are presented as mean±standard error of the mean (SEM). Statistical significance was determined using a two-tailed Student's t-test or one-way analysis of variance (ANOVA) with Bonferroni's *post hoc* test (GraphPad Prism 9.0). $P<0.05$ was considered significant.

Results

Identification of differentially expressed genes in diabetic nephropathy

We conducted principal component analysis (PCA) by integrating four publicly available transcriptome datasets (GSE104948, GSE47185, GSE96804, GSE99340). Firstly, PCA was performed on the raw unstandardized data, and the results indicated significant batch effects among the different datasets (Figure 2A). Subsequently, the data from the four datasets were normalized. The results showed that batch effects were effectively mitigated, and the samples were clearly classified by disease status rather than dataset source (Figure 2B). To identify differentially expressed genes (DEGs) between the DN and control groups, a differential analysis was conducted, resulting in the identification of inter-group DEGs (adjusted P-value<0.05, |log2 fold change (FC)| > 0.585)(Figure 2C). The top 100 most significantly differentially expressed genes are shown in the heatmap, indicating a clear distinction between the disease and control groups (Figure 2D).

Table 1. Nucleotide sequences of the primers used for real-time polymerase chain reaction (PCR) to investigate the mRNA expression levels of hub genes

Genes	Forward (5′–3′)	Reverse (5′–3′)
GAPDH	GAAGGTGAAGGTCGGAGTC	GAAGATGGTGATGGGATTTC
G0S2	CGTTCTTCGGCGTGGTCAT	CGACTTCTTGCTCTGCTCCA
PTGDS	GCTTCACAGAGGATACCATT	TTGCTTCCGGAGTTTATTGT
CA2	GCTCCATGTCCGTCATC	TATGCCCAGGTCTGTAGG
CYP27B1	CCATGTGGCAGA AGGGATAA	AAACCGTAA ACCAGGCTAGG
HSD17B14	TGTGACAAGGATGAGGCTGG	CCACACAATCCAGGTGACCA

Identify the key module driving genes

To obtain the gene modules related to the DN score, we constructed a WGCNA model to identify the gene modules

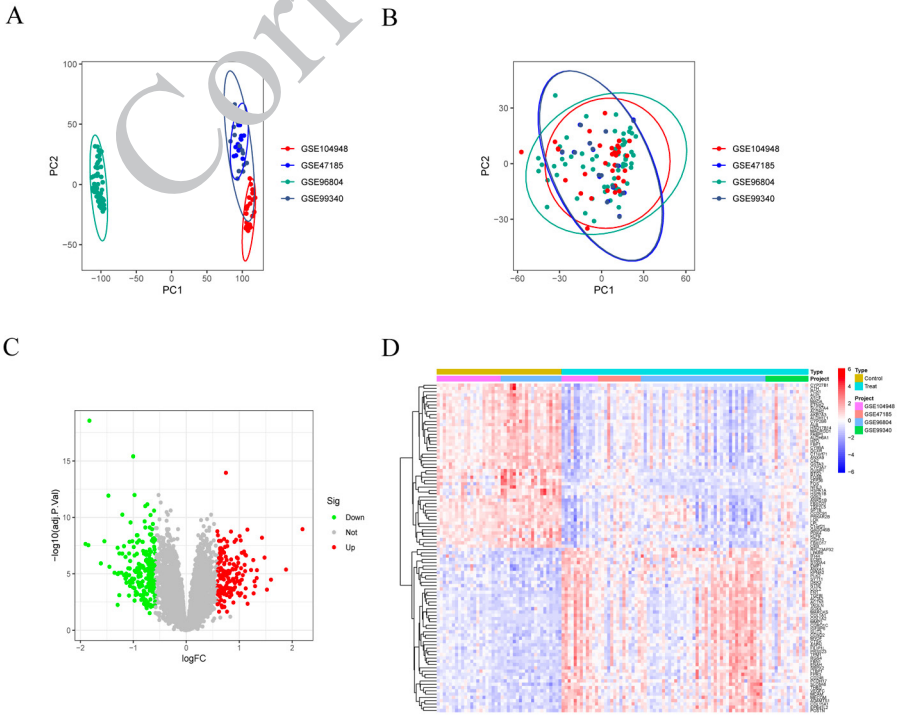

Figure 2. Data quality control and differentially expressed gene screening
(A) Principal component analysis (PCA) of four datasets (GSE104948, GSE47185, GSE96804, and GSE99340), PCA of the four datasets showing PC1 vs PC2, performed without prior standardization. (B) PCA was performed on PC1 and PC2 after normalizing the data obtained from the four datasets. (C) Volcano plot of DEGs between control and diabetic nephropathy (DN). (D) Heat map of the top 100 DEGs between control and DN

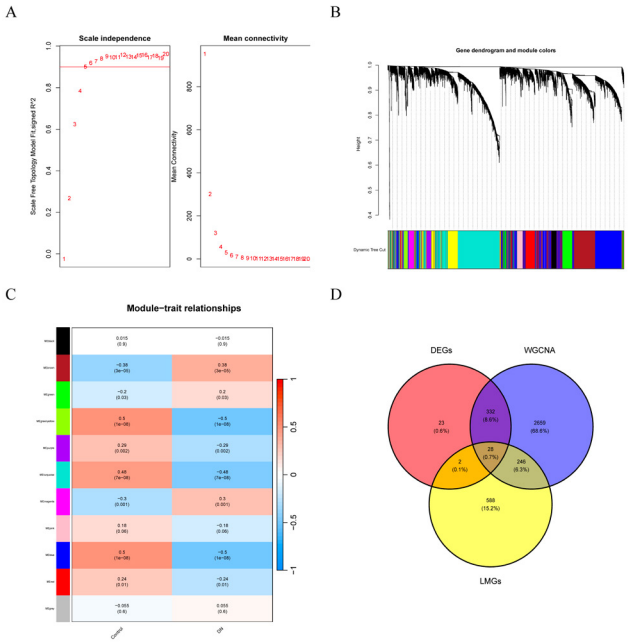

Figure 3. Weighted gene co-expression network analysis (WGCNA)
(A) Scale-free fit index analysis with different soft thresholds and average connectivity analysis with different soft threshold powers. (B) Heatmap of gene clustering trees with hierarchical clustering results. (C) Heatmap of gene and feature correlations. (D) Venn diagram showing the intersection of DEGs, WGCNA module genes, and LMGs, identifying 28 consensus hub genes as candidate biomarkers for diabetic nephropathy

associated with DN (Figure 3A). The "PickSoft Threshold" function was used to determine the soft threshold to construct a scale-free network. Then, module partitioning of the genes was performed, and a hierarchical clustering tree diagram was built using "Dynamic tree cutting" to display the cluster tree diagrams of different gene modules (Figure 3B). The correlation between the gene modules and the grouping was analyzed using the Pearson correlation coefficient and the *P*-value. Genes from modules exhibiting significant trait correlations (|r| ≥0.3) were merged for further analysis (Figure 3C). Subsequently, the candidate core genes selected from this key module were cross-matched with the list of DEGs and then integrated with the LMGs from the MSigDB database. By integrating these three gene sets, 28 common hub genes were identified as DN-LMGs for subsequent studies (Figure 3D).

Identification of candidate genes and GO/KEGG enrichment analysis

Firstly, we conducted a GO analysis on these 28 DN-LMGs. The results showed significant enrichment in biological processes related to lipid metabolism, fatty acid β-oxidation, and steroid biosynthesis (Figure 4A). The cellular components are significantly enriched in regions such as the mitochondrial matrix and peroxisomes, while the molecular functions include redox enzymes and hydrolase activities. To investigate the biological processes associated with the candidate genes, KEGG pathway analysis further confirmed their core roles in key metabolic pathways, including the PPAR signaling pathway, fatty acid degradation, and biosynthesis of unsaturated fatty acids (Figures 4B, C). Furthermore, PPI analysis was performed to explore the functional relationships among the candidate genes. Based on the STRING database and visualized using Cytoscape, the PPI network revealed a highly interconnected module centered on metabolic regulation (Figure 4D).

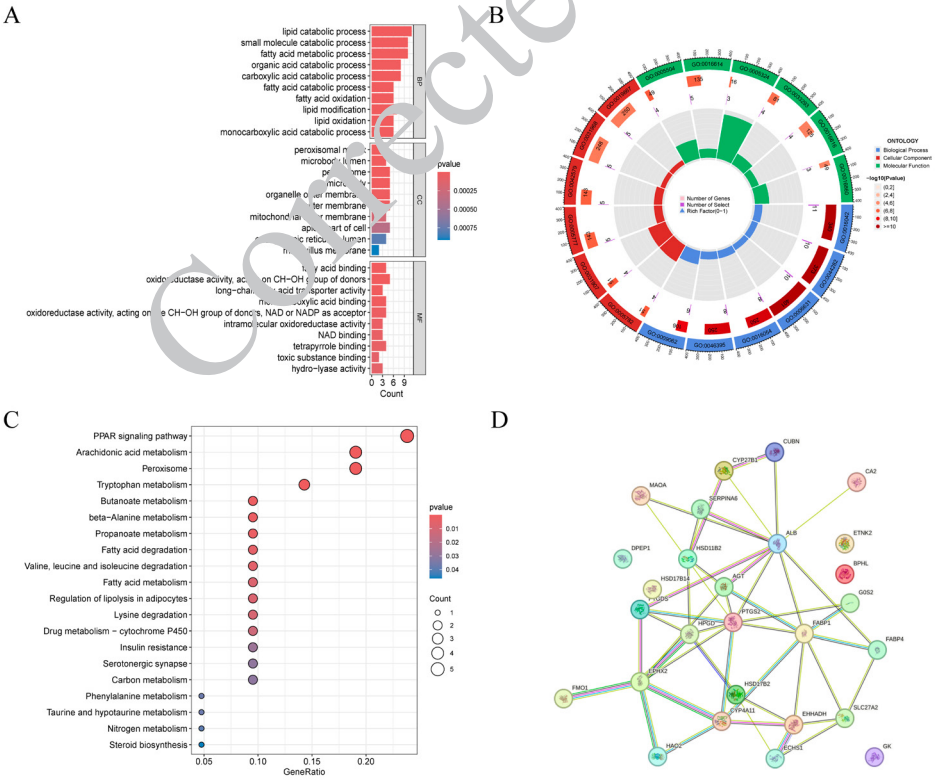

Figure 4. Functional enrichment and gene interaction analysis
(A) Gene Ontology (GO) enrichment analysis of consensus hub differentially expressed genes for biological process (BP), cellular component (CC), and molecular function (MF). (B–C) KEGG pathway enrichment analysis showing significantly enriched metabolic and signaling pathways; dot size represents gene count and color indicates adjusted P-value. (D) PPI network according to the STRING database

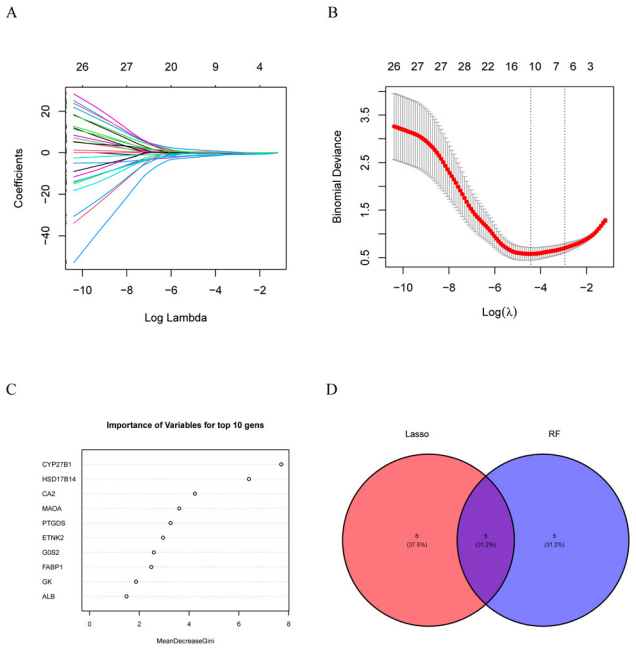

Figure 5. Machine learning-based screening of pivotal genes
(A) LASSO regression for 28 hub genes. (B) Cross-validation of parameter selection in LASSO regression. (C) Top 10 genes ranked by mean decrease in Gini index from random forest analysis, revealing variable importance for classification. (D) Venn diagram intersection of genes selected by LASSO and Random Forest to identify five core biomarkers (G0S2, PTGDS, CA2, CYP27B1, HSD17B14)

Machine learning for screening key genes

To further refine the biomarker panel and identify key genes, we employed two complementary machine learning algorithms. First, LASSO regression analysis with 10-fold cross-validation was conducted in the training set to screen candidate genes (Figures 5A-B). Additionally, the RF model ranked the genes by the average decrease in the Gini index, reflecting their importance (Figure 5C). The intersection of genes selected by both LASSO and RF algorithms yielded five core biomarkers: G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 (Figure 5D).

Nomogram risk assessment

To translate biomarkers into clinical application tools, we developed a diagnostic model based on the characteristics of a five-gene signature and performed multidimensional validation. To facilitate interpretation, a nomogram was established based on this model, which visually demonstrated the contribution of each gene to the risk of DN (Figure 6A). In the training set, the logistic regression model constructed by integrating the expression values of GOS2, PTGDS, CA2, CYP27B1, and HSD17B14 demonstrated remarkable discrimination ability. The area under the receiver operating characteristic curve was 0.957, indicating excellent diagnostic performance (Figure 6B). Additionally, the clinical nomogram used to predict an individual's disease risk demonstrated high predictive accuracy, as assessed by the calibration plot and DCA curve (Figures 6C-D). To ensure the robustness and generalizability of the model, we used an external validation set from independent datasets (GSE30528 and GSE104954), and the results showed that its diagnostic accuracy remained high in this cohort (Figure 6E). It is worth noting that each gene with this five-gene signature also demonstrated significant independent diagnostic value in the validation set. The AUCs for G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 were 0.836, 0.800, 0.885, 0.906, and 0.916, respectively (Figures 6F). These results collectively indicate that the lipid metabolism-related characteristics of this five-gene signature are not the result of overfitting, but rather represent a stable and applicable diagnostic marker for DN.

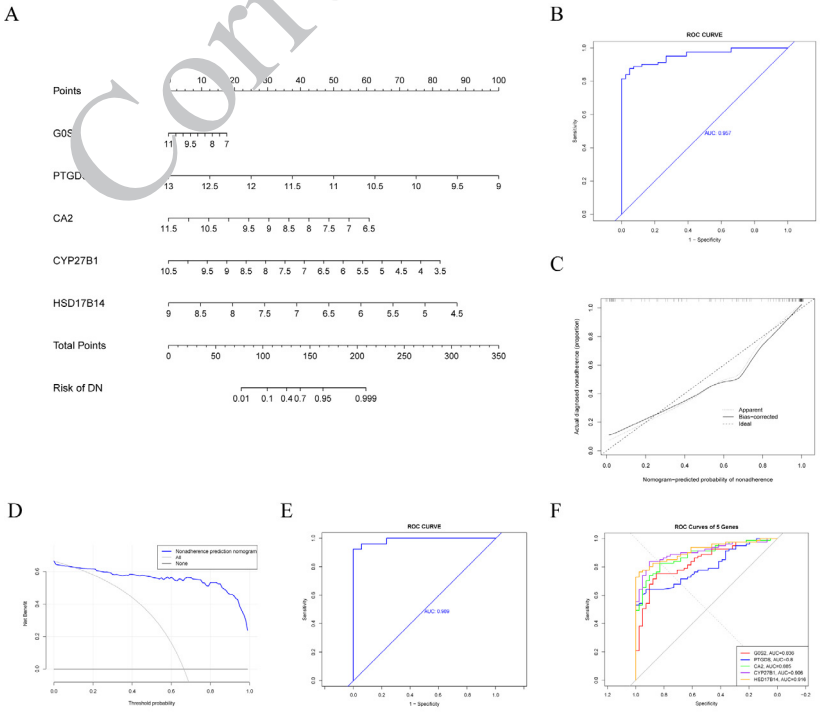

Figure 6. Diagnostic performance validation and clinical model construction
(A) The nomogram for five biomarkers (G0S2, PTGDS, CA2, CYP27B1, HSD17B14) in the training set. (B) Receiver operating characteristic curve (ROC) curve of the combined five-gene signature in the training cohort. (C) Calibration curve of the five-gene nomogram model in the training cohort. (D) DCA curve for the nomogram. (E) ROC curve of the combined five-gene signature in the validation cohort. (F) ROC curves for G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 in the validation cohort

The connection between genetic characteristics and the immune microenvironment

To investigate differences in the immune microenvironment among samples from different groups (normal and DN samples), CIBERSORT was used on the training set to assess the abundance of immune cell infiltration across all samples. The analysis revealed significant changes in the renal immune environment in diabetic kidney disease. Compared to the control group, the infiltrations of M2-type macrophages and specific T cell subsets were increased in DN tissues. Other immune cell changes were also observed (Figure 7A). Correlation analysis revealed complex interactions among different immune cell types in the renal microenvironment (Figure 7B). More importantly, the expression levels of five identified genetic features were significantly correlated with the abundance of specific immune cell types, especially showing a strong positive correlation with the infiltration of M2-type macrophages (Figure 7C). This indicates an association between lipid metabolism reprogramming and pro-fibrotic immune remodeling.

Molecular subtypes for identifying lipid metabolism characteristics

To investigate the heterogeneity of DN, a co-clustering analysis was applied to these five key LMGs, successfully dividing the DN samples into two distinct molecular subtypes. When the clustering number (K) was set to 2, the consensus matrix heatmap displayed a clear block-like structure and high stability in sample co-clustering, indicating the presence of two internally consistent molecular subtypes (Figure 8A). The consensus cumulative distribution function (CDF) curve and ΔK analysis further confirmed that K=2 was the optimal number of clusters as clustering stability peaked at this value (Figures 8B-C). The sample assignment-tracking diagram visually depicts the evolution of sample clustering relationships

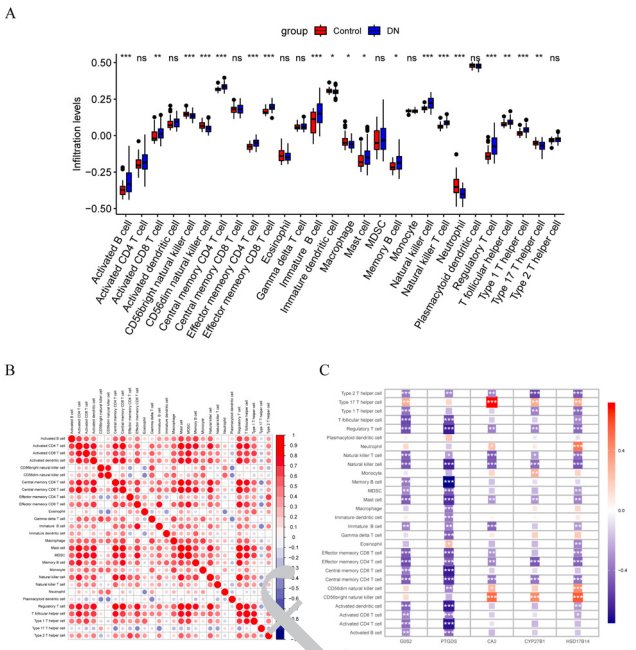

Figure 7. Immune cell infiltration analysis
(A) Violin plot showing differential immune cell infiltration between diabetic nephropathy (DN) and control groups calculated by the CIBERSORT algorithm, displaying relative abundance of 22 immune cell subsets (*P<0.05; **P<0.01; ***P<0.001). (B) Correlation network heatmap depicting interrelationships among different immune cells, with colors representing correlation coefficients. (C) Correlation heatmap between five-gene biomarker signature and infiltrating immune cells, with color intensity indicating correlation strength and asterisks marking significant correlations (P<0.05)

from K=2 to K=6 (Figure 8D). Based on these five genetic characteristics, principal component analysis was conducted for dimensionality reduction, and the results showed a clear separation between the two subtypes, indicating significant transcriptional differences reflected in the data distribution (Figure 8E). To further verify the expression differences of these five LMGs in DN disease subtypes, Boxplots

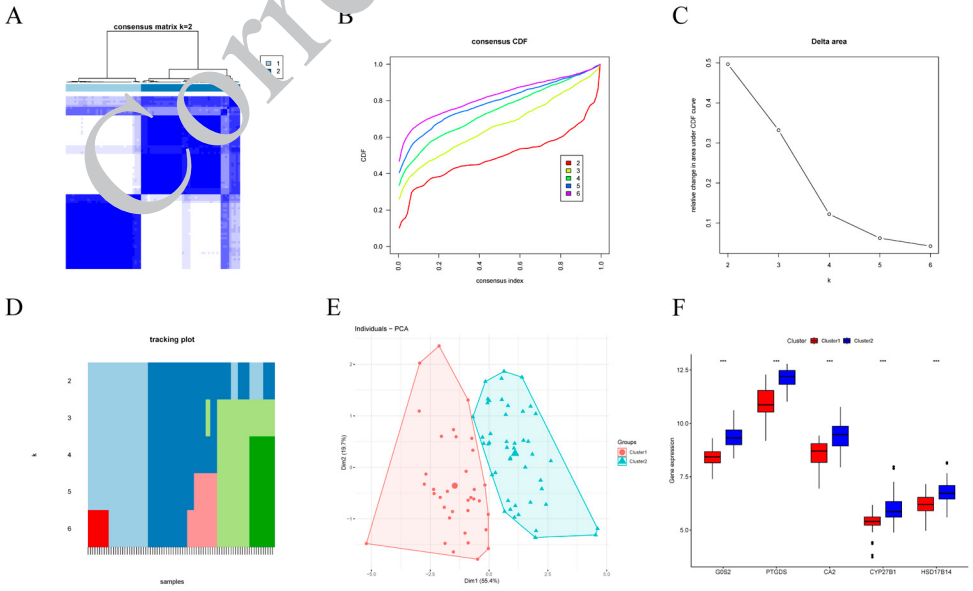

Figure 8. Consensus clustering identification of diabetic nephropathy (DN) molecular subtypes
(A)Consensus matrix heatmap (K=2) showing sample clustering stability, with color intensity indicating co-clustering probability. (B) Cumulative distribution function (CDF) curve of consensus index for evaluating stability across different cluster numbers (K=2–6). (C) Delta analysis of cluster number stability to determine optimal cluster number via relative change in CDF curves. (D) Tracking plot showing sample membership changes from K=2 to K=6. (E) Principal component analysis of molecular subtypes based on a five-gene signature to validate inter-subtype separation. (F) The expression levels of the five biomarkers between cluster 1 and cluster 2

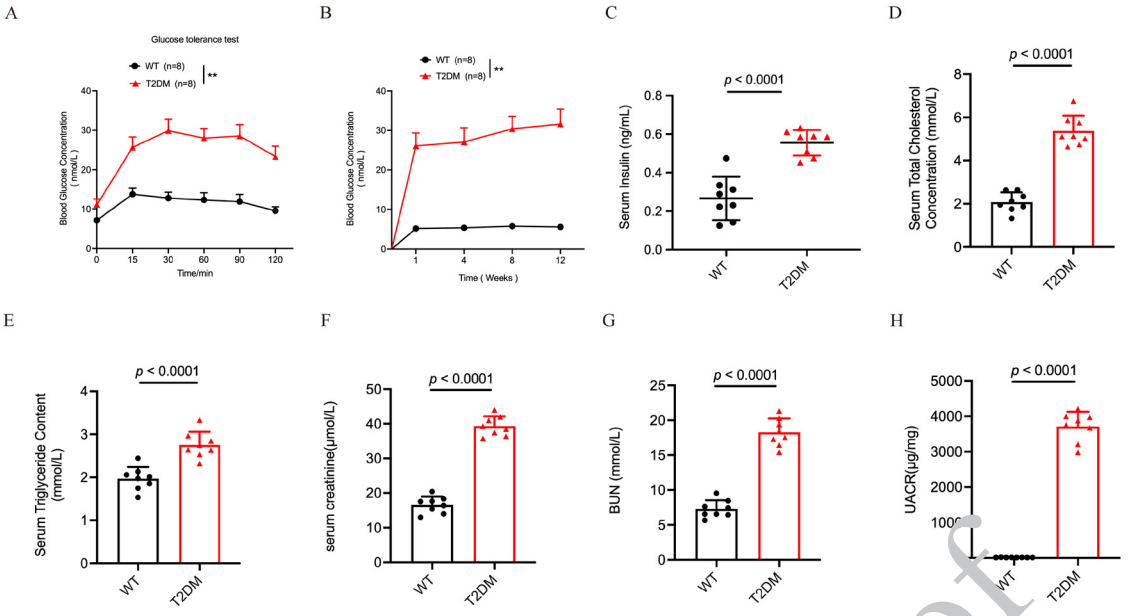

Figure 9. Establishment of a type 2 diabetes mellitus (T2DM) mouse model
Db/db mice were selected as the research object of the T2DM model, and the related indices were determined at week 12. (A) Insulin resistance test. (B) Blood glucose concentration. (C) Serum insulin concentration. (D) Total cholesterol. (E) Triacylglycerol. (F) Serum creatinine. (G) Blood urea nitrogen concentration. (H) The urinary albumin/creatinine ratio (UACR). $^{**}P<0.01$ vs wild-type (WT) group

(Figure 8F) display the expression levels of LMGs and the variations between the two DN subtypes. This study is the first to classify DN into two stable molecular subtypes based on hub LMGs, providing a theoretical basis for the subsequent classification and treatment strategies targeting the pathological mechanism of DN.

Establishment of animal models of diabetic nephropathy

We used 8-week-old wild-type and db/db mice to model type 2 diabetes mellitus. At week 12, we measured fasting glucose, glucose tolerance, insulin, TG, TC, SCR, and BUN reflecting the hallmark features of T2DM. Compared with WT mice, T2DM mice exhibited significantly impaired glucose tolerance and blood glucose levels at all time points (Figure 9A-B). Insulin levels were markedly elevated (Figure 9C), and TG, TC, SCR, BUN, and UACR were also significantly increased, indicating metabolic dysregulation and renal impairment in the T2DM group (Figure 9D-H).

Validation of hub gene expression in a DN animal model

To assess renal dysfunction in db/db mice, Doppler ultrasonography was performed. Compared with the WT group, the db/db mice exhibited pronounced hydronephrosis in the T2DM group (Figure 10A). TEM images revealed thickening of the glomerular basement membrane and endothelial injury in db/db mouse kidneys, along with fusion or detachment of podocyte foot processes and pronounced mitochondrial swelling with loss of cristae in renal tubular epithelial cells (Figure 10B). H&E staining revealed pronounced thickening of the basement membrane, enlarged and swollen glomeruli, and extensive deposition of membranous material near the mesangium. The renal tubular lumens were markedly narrowed, with hypertrophy of tubular epithelial cells. Collagen accumulation in certain regions indicated fibrosis, further confirmed by Masson's trichrome staining, which demonstrated extensive fibrotic deposition in the kidneys of DN mice (Figure 10C-D). We further demonstrated that, compared with WT mice, both the gene and protein levels of G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 were markedly elevated in the kidneys of DN mice. These results suggest that the upregulation of these genes may contribute to the metabolic and structural alterations characteristic of DN, underscoring their potential role in the progression of renal injury in DN (Figure 10E-G).

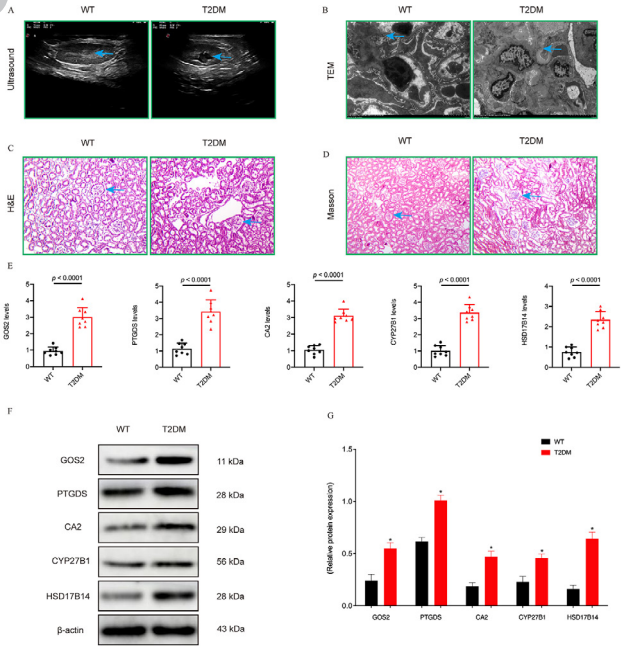

Figure 10. Verify the expression of hub genes in the diabetic nephropathy (DN) animal model
(A) Doppler ultrasonography. (B) Transmission electron microscopy. (C) H&E staining. (D) Masson Staining. (E) The gene expression levels of G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 were detected by qRT-PCR. (F) The protein expression levels of G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 were detected by Western blot. $^{*}P<0.05$ vs WT group

Discussion

DN represents the leading cause of CKD and end-stage renal failure worldwide. However, early diagnostic biomarkers with robust pathobiological relevance remain limited[26]. To address this critical gap, we developed a multi-layered integrative framework that leverages transcriptomic meta-analysis, network topology, and machine learning to derive a five-gene lipid metabolism signature with exceptional discriminatory power (AUC=0.957) for DN detection. Our findings not only identify a clinically translatable biomarker panel but also clarify the central role of lipid metabolic reprogramming in shaping the renal immune microenvironment and molecular disease heterogeneity.

Integrative strategy and novel biomarker discovery

The strength of our approach lies in the systematic convergence of three independent gene prioritization strategies—differential expression, WGCNA, and a priori lipid metabolism gene curation from MSigDB. This tripartite filtering mitigates the false discovery rates inherent in single-algorithm analyses, yielding 28 high-confidence hub genes that represent the nexus of transcriptional dysregulation, topological importance, and metabolic pathobiology. The subsequent application of two complementary machine learning algorithms, LASSO regression and RF, further distilled this set to five core biomarkers: G0S2, PTGDS, CA2, CYP27B1, and HSD17B14. This dual-machine-learning intersection ensures both parsimony and robustness: LASSO selects for synergistic predictive features while penalizing multicollinearity, whereas RF captures non-linear interactions and ranks variables by their contribution to classification accuracy. The resultant signature's superior diagnostic performance, validated through bootstrap-resampled ROC analysis and an independent external cohort, underscores its generalizability beyond the training datasets.

Pathobiological implications of the five-gene signature

Each biomarker encapsulates a distinct facet of lipid-mediated renal injury, offering mechanistic insights beyond simple classification. G0S2, a potent inhibitor of adipose triglyceride lipase (ATGL), emerges as a critical gatekeeper of lipotoxicity; its up-regulation likely promotes lipid droplet accumulation and ceramide-driven podocyte apoptosis, a hallmark of DN (27). PTGDS (prostaglandin D2 synthase) synthesizes PGD2, which exerts context-dependent effects. While PGD2 is traditionally viewed as anti-inflammatory, its metabolite 15d-PGJ2 can activate PPARγ pathways that paradoxically promote fibrogenesis in diabetic milieus (28). The inclusion of CA2 (carbonic anhydrase II) links acid-base homeostasis to lipid metabolism; CA2 catalyzes the bicarbonate equilibrium essential for β-oxidation and has been implicated in regulating fatty acid synthase, suggesting a coupling between metabolic acidosis and lipogenesis in diabetic kidneys (29). CYP27B1 (25-hydroxyvitamin D3-1α-hydroxylase) integrates vitamin D signaling with lipid homeostasis. Its down-regulation in DN may impair the conversion of vitamin D to its active metabolite. This impairment thereby disrupts VDR-mediated suppression of NF-κB-driven inflammation and epithelial-mesenchymal transition. This observation aligns with emerging evidence that vitamin D deficiency exacerbates renal lipotoxicity through impaired LXRs and PPARα signaling (30, 31). In addition, HSD17B14, a hydroxysteroid dehydrogenase involved in steroid and fatty acid desaturation, may modulate the production of pro-inflammatory eicosanoids, thereby bridging sex steroid metabolism and renal inflammation (32). Collectively, these genes form a coherent network wherein lipid accumulation, oxidative stress, and immunometabolic dysregulation converge to drive glomerular and tubular injury.

Lipid metabolism as a central hub in DN pathogenesis

Our GO and KEGG enrichment analyses revealed that hub genes are significantly enriched in PPAR signaling, fatty acid degradation, and steroid biosynthesis pathways, reinforcing lipotoxicity as a central pathogenic mechanism. The "lipid metabolic reprogramming" hypothesis posits that hyperglycemia and insulin resistance trigger ectopic lipid deposition in renal cells, which in turn fosters a self-perpetuating cycle of inflammation, endoplasmic reticulum stress, and fibrosis (33). The identification of these five LMGs as predictive features validates this hypothesis at the transcriptional level. Moreover, it suggests that therapeutic modulation of specific lipid handling pathways, rather than broad-spectrum lipid lowering, may be required for renal protection. For instance, targeting the GOS2-ATGL protein-protein interaction could enhance lipophagy, while CYP27B1 agonism might restore anti-inflammatory vitamin D signaling (34, 35).

Immune microenvironment remodeling and molecular subtypes

CIBERSORT analysis revealed profound shifts in renal immune composition, and the altered abundances of M2 macrophages, CD4+ memory T cells, and regulatory T cells correlated with the five-gene signature, which represents a specific molecular profile. Lipid metabolic perturbations in renal parenchymal cells actively shape the immune landscape. This likely occurs via metabolic byproducts such as succinate, ceramides, and oxidized phospholipids, which act as damage-associated molecular patterns (DAMPs). Notably, the signature showed the strongest correlation with M2 macrophage polarization-a key driver of fibrosis-implicating lipid metabolism in the metabolic programming of pro-fibrotic immune cells. Consensus clustering based on the five-gene signature identified two stable molecular subtypes, providing a proof of concept for precision nephrology. This stratification may reflect underlying differences in lipotoxic susceptibility, genetic background, or duration of diabetes, offering a framework for patient enrichment in future clinical trials. Furthermore, the subtype-specific expression patterns visualized by two-way hierarchical clustering support the existence of distinct DN endophenotypes that may respond differentially to interventions targeting lipid metabolism.

Clinical translation and precision medicine implications

We developed a nomogram integrating the five-gene signature to create a clinically actionable tool for predicting disease risk. With an AUC of 0.957, the model achieves near-optimal discrimination-close to the highest achievable performance-outperforming traditional clinical markers such as albuminuria and eGFR, which detect DN only after substantial damage has occurred. The high calibration

accuracy suggests that the model's predicted probabilities reliably reflect true disease risk, supporting its potential use for early screening in diabetic patients. Moreover, the individual gene-specific ROC curves in the validation cohort confirm each component's independent contribution, thereby addressing concerns about overfitting.

Validation of five key biomarkers in a DN animal model

Using a db/db mouse model, we validated five key biomarkers associated with DN. The db/db mice exhibited hallmark metabolic abnormalities of type 2 diabetes, including persistent hyperglycemia, impaired glucose tolerance, hyperinsulinemia, and dyslipidemia, accompanied by elevated SCR, BUN, and UACR, confirming the successful establishment of DN. Pronounced structural and ultrastructural abnormalities paralleled renal functional impairment. Doppler ultrasonography revealed hydronephrosis, while TEM demonstrated glomerular basement membrane thickening, endothelial injury, podocyte foot process effacement, and mitochondrial damage in tubular epithelial cells. Histological analyses further showed glomerular hypertrophy, mesangial expansion, tubular injury, and extensive fibrotic deposition, reflecting progressive renal remodeling in DN. Within this pathological context, G0S2, PTGDS, CA2, CYP27B1, and HSD17B14 were significantly up-regulated at both the transcript and protein levels in DN kidneys. The strong concordance between their elevated expression and renal dysfunction suggests that these genes are closely linked to the metabolic, mitochondrial, and fibrotic alterations driving DN progression. Together, these findings establish these molecules as robust biomarkers of DN and highlight their potential involvement in disease pathogenesis.

Limitations and future directions

Despite these strengths, several caveats warrant consideration. First, our findings require further experimental validation using qRT-PCR, immunoblotting, and single-cell RNA sequencing (scRNA-seq) to confirm cell-type-specific expression patterns and protein-level alterations. Furthermore, future studies will primarily focus on applying scRNA-seq to achieve a more comprehensive understanding of cellular heterogeneity and molecular mechanisms in DN. Second, although expression validation was conducted in external datasets and animal models, the present study did not include loss-of-function or gain-of-function experiments to establish direct causal relationships between the identified biomarkers and DN progression. Therefore, further functional studies, including gene knockdown and overexpression experiments, are warranted to elucidate the underlying biological mechanisms and validate the functional roles of these biomarkers in DN. Third, the reproducibility and robustness of these subtypes still require validation in additional independent cohorts. Future studies with larger sample sizes and multicenter datasets are needed to confirm further the stability and clinical relevance of the identified molecular subtypes. Fourth, clinical metadata were limited; incorporating longitudinal eGFR decline, proteinuria levels, and HbA1c trajectories would strengthen predictive models and enable survival analysis.

Building on these findings, future work should prioritize functional studies using podocyte and tubular epithelial cell lines with G0S2 knockdown or CYP27B1 overexpression under high-glucose and lipid-loaded conditions to establish causality. *In vivo* validation using diabetic mouse models (e.g., db/db or Akita) treated with targeted interventions (e.g., G0S2 inhibitors or vitamin D analogs) will clarify the therapeutic potential. Additionally, prospective clinical trials integrating the five-gene signature into point-of-care devices for renal biopsy samples or urinary extracellular vesicles could revolutionize DN diagnostics. Finally, investigating the interaction between this lipid-centric signature and other pathways (e.g., TGF-β_1/SMAD, mTORC1) will provide a more holistic view of DN pathogenesis (36, 37).

Conclusion

In summary, we have developed and validated a five-gene lipid metabolism signature that robustly distinguishes DN from controls and reveals novel pathogenic axes linking lipid dysregulation, immune infiltration, and molecular heterogeneity. By integrating systems biology with machine learning, this study provides a blueprint for biomarker discovery in complex metabolic diseases. Furthermore, it highlights lipid metabolic reprogramming as a central and targetable mechanism in DN pathophysiology. This signature holds immediate translational potential for early diagnosis, risk stratification, and therapeutic guidance, marking a significant step toward precision medicine in diabetic kidney disease.

Acknowledgment

This research is supported by the Heilongjiang Natural Science Foundation Project (No. PL2025H223); Heilongjiang Province Traditional Chinese Medicine Research Project (No. ZHY2024-212); Basic Scientific Research Business Research Project of Heilongjiang (No.2024-KYYWF-0502); Program for Young Talents of Basic Research in Universities of Heilongjiang Province (No. YQJH2025260); Qiqihar Bureau of Science and Technology (No. LSFGG-2025135).

Authors' Contributions

H Y and Q M conceived and supervised the study; H Y, X L, J X, S L, and Q M analyzed the data; J X wrote the manuscript; H Y, X L, J X, S L, and Q M reviewed the results; H Y, X L, J X, S L, T W, and Q M approved the final version for publication.

Conflicts of Interest

The authors declare no competing interests

Declaration

We have not used any AI tools or technologies to prepare this manuscript.

References

1. Rayego-Mateos S, Rodrigues-Diez RR, Fernandez-Fernandez B, Mora-Fernández C, Marchant V, Donate-Correa J, *et al.* Targeting inflammation to treat diabetic kidney disease: The road to 2030. Kidney Int 2023; 103: 282-296.
2. Cooper ME. Interaction of metabolic and haemodynamic factors in mediating experimental diabetic nephropathy. Diabetologia 2001; 44: 1957-1972.
3. Pagtalunan ME, Miller PL, Jumping-Eagle S, Nelson RG, Myers BD, Rennke HG, *et al.* Podocyte loss and progressive glomerular injury in type II diabetes. J Clin Invest 1997; 99: 342-348.
4. Qian Y, Feldman E, Pennathur S, Kretzler M, Brosius FC, 3rd.

From fibrosis to sclerosis: Mechanisms of glomerulosclerosis in diabetic nephropathy. Diabetes 2008; 57: 1439-1445.
5. Kato M, Natarajan R. Epigenetics and epigenomics in diabetic kidney disease and metabolic memory. Nat Rev Nephrol 2019; 15: 327-345.
6. Bobulescu IA. Renal lipid metabolism and lipotoxicity. Curr Opin Nephrol Hypertens 2010; 19: 393-402.
7. Shen S, Ji C, Wei K. Cellular senescence and regulated cell death of tubular epithelial cells in diabetic kidney disease. Front Endocrinol (Lausanne) 2022; 13: 924299.
8. Opazo-Ríos L, Mas S, Marín-Royo G, Mezzano S, Gómez-Guerrero C, Moreno JA, *et al.* Lipotoxicity and diabetic nephropathy: Novel mechanistic insights and therapeutic opportunities. Int J Mol Sci 2020; 21: 2632.
9. Sharma K, Karl B, Mathew AV, Gangoiti JA, Wassel CL, Saito R, *et al.* Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease. J Am Soc Nephrol 2013; 24: 1901-1912.
10. Jiang Y, Peng Y, Yang X, Yu J, Yu F, Yuan J, *et al.* PM(2.5) exposure aggravates kidney damage by facilitating the lipid metabolism disorder in diabetic mice. PeerJ 2023; 11: e15856.
11. Huang S, Jin Y, Zhang L, Zhou Y, Chen N, Wang W. PPAR gamma and PGC-1alpha activators protect against diabetic nephropathy by suppressing the inflammation and NF-kappaB activation. Nephrology (Carlton) 2024; 29: 858-872.
12. Wu L, Liu C, Chang DY, Zhan R, Zhao M, Man Lam S, *et al.* The attenuation of diabetic nephropathy by annexin A1 via regulation of lipid metabolism through the AMPK/PPARα/CPT1b pathway. Diabetes 2021; 70: 2192-2203.
13. Shi J, Liu X, Jiao Y, Tian J, An J, Zou G, *et al.* mTOR pathway: A key player in diabetic nephropathy progression and therapeutic targets. Genes Dis 2025; 12: 101260.
14. Wang Y, Liu T, Wu Y, Wang L, Ding S, Hou B, *et al.* Lipid homeostasis in diabetic kidney disease. Int J Biol Sci 2024; 20: 3710-3724.
15. Mora-Fernández C, Domínguez-Pimentel V, de Fuentes MM, Górriz JL, Martínez-Castelao A, Navarro-González JF. Diabetic kidney disease: From physiology to therapeutics. J Physiol 2014; 592: 3997-4012.
16. Dong B, Liu X, Yu S. Utilizing machine learning algorithms to identify biomarkers associated with diabetic nephropathy: A review. Medicine (Baltimore) 2024; 103: e37232.
17. Sun D, Wei S, Wang D, Zeng M, Mo Y, Li H, *et al.* Integrative analysis of potential diagnostic markers and therapeutic targets for glomerulus-associated diabetic nephropathy based on cellular senescence. Front Immunol 2023; 14: 1327757.
18. Sun Y, Dai W, He W. Identification of key immune-related genes and immune infiltration in diabetic nephropathy based on machine learning algorithms. IET Syst Biol 2023; 17: 95-106.
19. Guo C, Wang W, Dong Y, Han Y. Identification of key immune-related genes and potential therapeutic drugs in diabetic nephropathy based on machine learning algorithms. BMC Med Genomics 2024; 17: 220.
20. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47.
21. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.
22. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. J Stat Softw 2012; 46: i11.
23. Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L, *et al.* Using clusterProfiler to characterize multiomics data. Nat Protoc 2024; 19: 3292-3320.
24. Tay JK, Narasimhan B, Hastie T. Elastic net regularization paths for all generalized linear models. J Stat Softw 2023; 106: 1.
25. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12: 77.
26. Brennan E, McEvoy C, Sadlier D, Godson C, Martin F. The genetics of diabetic nephropathy. Genes 2013; 4: 596-619.
27. Li X, Wu Y, Zhao J, Wang H, Tan J, Yang M, *et al.* Distinct cardiac energy metabolism and oxidative stress adaptations between obese and non-obese type 2 diabetes mellitus. Theranostics 2020; 10: 2675-2695.
28. Łukowicz K, Grygier B, Basta-Kaim A. Emerging role of neural stem/progenitor cell secretome in brain inflammatory response modulation. Pharmacol Rep 2025; 77: 907-920.
29. Mardones P, Chang C, Oude Elferink RP. Cyclic AMP and alkaline pH downregulate carbonic anhydrase 2 in mouse fibroblasts. Biochim Biophys Acta 2014; 1840: 1765-1770.
30. Lidberg KA, Annalora AJ, Jozic M, Elson DJ, Wang L, Bammler TK, *et al.* An isense oligonucleotide development for the selective modulation of CYP3A5 in renal disease. Sci Rep 2021; 11: 4722.
31. de Bragança AC, Volpini RA, Mehrotra P, Andrade L, Basile DP. Vitamin D deficiency contributes to vascular damage in sustained ischemic acute kidney injury. Physiol Rep 2016; 4: e12829.
32. Maciejewska-Skrendo A, Cięszczyk P, Chycki J, Sawczuk M, Smółka W. Genetic markers associated with power athlete status. J Hum Kinet 2019; 68: 17-36.
33. Lei L, Yu J, Zhang B, Liu P, Meng Z, Song Y, *et al.* Systematic profiling reveals hepatic immune and metabolic dysregulation in DNASE1L3-deficient mice. iScience 2025; 28: 114198.
34. Poursistany H, Azar ST, Azar MT, Raeisi S. The current and emerging Klotho-enhancement strategies. Biochem Biophys Res Commun 2024; 693: 149357.
35. Hummel DM, Thiem U, Höbaus J, Mesteri I, Gober L, Stremnitzer C, *et al.* Prevention of preneoplastic lesions by dietary vitamin D in a mouse model of colorectal carcinogenesis. J Steroid Biochem Mol Biol 2013; 136: 284-288.
36. Deshpande A, Loth M, Sidiropoulos DN, Zhang S, Yuan L, Bell ATF, *et al.* Uncovering the spatial landscape of molecular interactions within the tumor microenvironment through latent spaces. Cell Syst 2023; 14: 285-301.
37. Heydarpour F, Sajadimajd S, Mirzarazi E, Haratipour P, Joshi T, Farzaei MH, *et al.* Involvement of TGF-β and autophagy pathways in pathogenesis of diabetes: A comprehensive review on biological and pharmacological insights. Front Pharmacol 2020; 11: 498758.