Original Article

Identification of key biomarkers and immune infiltration in renal interstitial fibrosis

Zhanhong Hu[1#], Yumei Liu[2#], Ye Zhu[3#], Hongxia Cui[4], Jie Pan[1]

[1]Department of Pharmacy, The Second Affiliated Hospital of Soochow University, Suzhou, China; [2]College of Pharmaceutical Science, Soochow University, Suzhou, China; [3]Department of Nephrology, The Second Affiliated Hospital of Soochow University, Suzhou, China; [4]Department of Pathology, The Second Affiliated Hospital of Soochow University, Suzhou, China

Contributions: (I) Conception and design: Z Hu, J Pan; (II) Administrative support: J Pan; (III) Provision of study materials or patients: Y Liu, Y Zhu, H Cui; (IV) Collection and assembly of data: Z Hu, Y Liu; (V) Data analysis and interpretation: Z Hu, J Pan; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work and should be considered as co-first authors.

Correspondence to: Jie Pan. Department of Pharmacy, The Second Affiliated Hospital of Soochow University, Suzhou 215004, China. Email: pankypan@163.com; Hongxia Cui. Department of Pathology, The Second Affiliated Hospital of Soochow University, Suzhou 215004, China. Email: hongxia217@163.com.

Background: Renal interstitial fibrosis (RIF) is the common final pathway that mediates almost all progressive renal diseases. However, the underlying mechanisms of RIF have not been fully elucidated. Therefore, the current study aimed to explore the etiology of RIF and identify the key targets and immune infiltration patterns of RIF.

Methods: Ribonucleic acid (RNA)-seq data of RIF and normal samples were downloaded from the Gene Expression Omnibus (GEO) database. Weighted gene co-expression network analysis (WGCNA) was performed to screen relevant modules associated with RIF. Differentially expressed genes (DEGs) between the RIF and normal samples were identified using the limma package. Machine learning methods were used to identify hub gene signatures related to RIF. Further biochemical approaches including quantitative polymerase chain reaction (qPCR), immunoblotting and immunohistochemistry experiments were performed to verify the hub signatures in the RIF samples. Single sample gene set enrichment analysis (ssGSEA) was used to analyze the proportions of 28 immune cells in RIF and normal samples.

Results: WGCNA showed 121 RIF-related genes. A total of 523 DEGs were found between the RIF and normal samples. By overlapping these genes, we obtained 78 RIF-related genes, which were mainly enriched in Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with immunity and inflammation. Integrative analysis of machine learning methods showed prominin 1 (*PROM1*), tryptophan aspartate-containing coat protein (*CORO1A*), interferon-stimulated exonuclease gene 20 (*ISG20*), and tissue inhibitor matrix metalloproteinase 1 (*TIMP1*) as hub gene signatures in RIF. Further, receiver operating curve (ROC) curves implied the diagnostic role of *ISG20* and *CORO1A* in RIF. The expression levels of ISG20 and CORO1A were significantly higher in fibrotic tubular cells and renal tissues based on biochemical approaches. The immune microenvironment was found to be markedly altered in the RIF samples, as 21 differentially infiltrated immune cells (DIICs) were found between RIF and normal samples.

Conclusions: This study is the first to find that *ISG20* and *CORO1A* are key biomarkers and to examine the landscape of immune infiltration in RIF. Our findings provide novel insights into the mechanisms and treatment of patients with RIF.

Keywords: Renal interstitial fibrosis (RIF); immune infiltration; machine learning; *CORO1A*; *ISG20*

Introduction

Chronic kidney disease (CKD), which is characterized by renal dysfunction, has become an important public health problem because of its high rate of morbidity and mortality. CKD affects approximately 10–13% of the population worldwide (1). Renal fibrosis, particularly renal interstitial fibrosis (RIF), is the final common outcome of various types of CKD. Research has found that RIF is the most reliable predictor of the progression of CKD to end-stage renal failure (ESRD). However, the underlying mechanisms of RIF are yet to be fully elucidated, and current therapies only delay disease progression (2). Therefore, exploring the underlying mechanism and discovering new potential drug targets for RIF are important for future treatments of CKD.

RIF, a progressive and irreversible pathological feature, is characterized by inflammation, myofibroblast activation and migration, and matrix deposition and remodeling. It is currently believed that the RIF develops in response to the accumulation of extracellular matrix (ECM) due to epithelial-mesenchymal transition (EMT), transforming growth factor (TGF)-β signaling, oxidative stress and proteinuria (3). The immune system plays an important role in these changes (4). The immune system can be divided into the innate immune system, which mainly consists of monocytes, natural killer (NK), and dendritic cells (DC), and the adaptive immune system, which is represented by B and T lymphocytes. However, in patients with RIF, a comprehensive evaluation of the distribution of immune cells—the so-called immune infiltration, which might provide key insights into the development of novel therapeutic strategies to prevent the progression of RIF and CKD has not yet been reported.

With the development of microarray technologies, bioinformatics analyses have been widely used to identify disease-specific biomarkers. Recently, some critical genes with tubulointerstitial lesion in CKD were identified using weighted gene co-expression network analysis (WGCNA) by the analysis of GSE104954 and GSE47185 downloaded from the Gene Expression Omnibus (GEO) database (5). Samples of these two datasets were from CKD patients with various pathological type. These findings provided novel biomarkers associated with renal tubulointerstitial injury in CKD. However, this report has not solely focused on RIF. Most importantly, this report did not give any information related to immune system in RIF, which plays important roles in the progression of CKD.

In the current study, a bioinformatics analysis of expression data from RIF patients downloaded from GEO database was performed to explore the diagnostic biomarkers for RIF via WGCNA and machine learning. *ISG20* and *CORO1A* were identified as key hub genes closely correlated with RIF, which was further verified by biochemical studies with fibrotic tubular cells and renal tissue. Furthermore, potential drugs targeting *ISG20* and *CORO1A* were predicted. Most importantly, evaluation of immune cell infiltration showed that the abundance of most immune cells was found to significantly differ between RIF and normal samples. Collectively, these findings provide two novel molecular markers to further explore the pathological mechanisms and potential drug targets for RIF.

We present the following article in accordance with the MDAR reporting checklist (available at https://atm.amegroups.com/article/view/10.21037/atm-22-366/rc).

Methods

Data source

The expression data of 24 RIF and 25 normal samples in GSE22459, and 42 RIF and 99 normal samples in GSE76882, were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). To identify genes that are key to the progression of RIF, the genes in GSE22459 and GSE76882 datasets were first overlapped to obtain 15,779 genes from 66 RIF and 124 samples for further analysis (Figure S1A). Thereafter, the data of GSE22459 and GSE76882 were combined. The Sangerbox online tool (http://sangerbox.com) was applied to remove the batch effect of the two datasets (Figure S1B). Thereafter, a sample clustering tree map was constructed to detect and eliminate outliers, which showed that outlier samples were not detected (Figure S1C).

The Sangerbox online tool (http://sangerbox.com) was used to remove the batch effects of the two datasets.

WGCNA

WGCNA was performed based on gene expression profiles and sample traits (normal and RIF). To select the best soft threshold, the pick soft threshold function of WGCNA was used to calculate the value of β from 1 to 30. Based on the selected soft threshold, the adjacency matrix was converted to a topological overlap matrix to construct the network, and the gene dendrogram and module color were

established using the degree of dissimilarity. Further, the initial module was divided using dynamic tree cutting, and similar modules were merged. The Pearson correlation coefficients between the module eigengenes and sample traits were calculated to determine the relevant modules associated with RIF, where modules with |Correlation coefficient (cor)| >0.3 and a P value <0.05 were considered as key modules related to RIF. Using |module membership (MM)| >0.8 and |gene significance (GS)| >0.2 as criteria, genes in the relevant modules associated with RIF were further screened for downstream analysis.

Identification and functional analysis of candidate key genes in RIF

The Limma R package was used to identify differentially expressed genes (DEGs) between RIF and normal samples, with a threshold of |log$_2$FC| >0.5 and an adjusted P value <0.05. The ggplot and pheatmap R package were used to plot the volcano diagram and heatmap of DEGs, respectively. To obtain candidate key genes in RIF, we intersected DEGs with genes obtained from WGCNA. The ClusterProfiler R package was used to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of candidate key genes. Statistical significance was set at P<0.05. Furthermore, the protein-protein interaction network of candidate key genes was constructed and visualized using the STRING database (https://www.string-db.org/) and Cytoscape software (version 3.7.2).

Selection of hub gene signatures by machine learning

We applied the least absolute shrinkage and selection operator (LASSO) algorithm using the glmnet package in R software and screened the gene signatures under the optimal lambda with the smallest classification error. The random forest (RF) R package was used to construct the RF model to identify the important variables. The performance of the RF model was evaluated using the receiver operating curve (ROC). Support vector machine-recursive feature elimination (SVM-RFE) was also performed using the e1071 package in R software to identify the best variables (gene signatures) via the deletion of SVM-generated eigenvectors in conjunction with 5× fold cross validation. Finally, we identified hub gene signatures by overlapping gene signatures from the LASSO, RF, and SVM-RFE algorithms.

Identification of hub biomarkers in RIF

The diagnostic value of the hub gene signature was determined using ROC curves. Hub genes with areas under the curve (AUC) greater than 0.8 were identified as potential diagnostic biomarkers in RIF. The expression of diagnostic biomarkers in RIF and control samples were displayed in the box plot. The Wilcoxon test was used for the comparison with the ggpubr R package. Moreover, the chemicals interacting with the diagnostic biomarkers were predicted using the Comparative Toxicogenomics Database (CTD) (https://ctdbase.org). Accordingly, a chemical-diagnostic biomarker interaction network was constructed using the Cytoscape software.

Verification of hub biomarkers through biochemical experiments

Reagents and antibodies

TGF-β (240-B) was purchased from Bio-Techne (China). Adding TGF-β to epithelial cells in culture is a convenient way to construct fibrosis cell models, due to the important role of TGF-β in the induction of MCT (6). Therefore, TGF-β was used as stimulator to construct the renal fibrosis cell models in the present study. The following antibodies were used: anti-α-SMA antibody (19245T) from CST (USA); anti-CORO1A antibody (ab203698) from Abcam (Cambridge, UK); anti-ISG20 antibody (22097-1-AP) from Proteintech (USA); anti-Vimentin antibody (ab92547) from Abcam (Cambridge, UK); and anti-GAPDH antibody (60004-1-Ig) from Proteintech (USA).

Cell line and cell culture

HK-2 cells (Human renal tubular epithelial cells, 339833) were obtained from the Beijing Beina Chuanglian Biotechnology Research Institute (China). Cells were cultured in RPMI-1640 supplemented with 10% fetal bovine serum (FBS, EallBio, China), 100 U/mL penicillin, and 100 μg/mL streptomycin (NCM Biotech, China), and incubated at 37 ℃ in a 10% CO$_2$ atmosphere.

Western blotting analysis

Cell lysates from HK-2 cells stimulated with TGF-β were mixed with an appropriate amount of 5× sample loading buffer. Proteins were separated using sodium dodecylsulphate polyacrylamide gel electrophoresis (SDS-PAGE) and transferred onto a polyvinylidene fluoride membrane. The membrane was blocked with 5% non-

Page 4 of 16

Hu et al. Key biomarkers identification in renal interstitial fibrosis

fat milk and incubated with primary and secondary antibodies. The membrane was washed three times with Tris-Buffered Saline and Tween 20 (TBST) and incubated with each antibody. Protein bands were visualized with a hypersensitive ECL chemiluminescence kit (NCM Biotech, China), and signals were recorded using a Tanon 5200 chemiluminescence imaging system (China).

Quantitative polymerase chain reaction (qPCR)

Total RNA was isolated from HK-2 cells stimulated with TGF-β using TRIzol (Vazyme, China), and reverse transcribed into cDNA using the HiScript II Q RT SuperMix kit (Vazyme). qPCR was performed on an Applied Biosystems 7500 Real-Time PCR system using the 2× SYBR Green qPCR Master Kit (Biotool). The primer sequences were as follows: α-SMA, 5'- GACAATGGCTCTGGGCTCTGTAA -3' (forward) and 5'-CTGTGCTTCGTCACCCACGTA-3' (reverse); CORO1A, 5'-GCACCCAGACACGATCTACAG-3' (forward) and 5'-GGACGGTCCTTCTCAGCTAC-3' (reverse); ISG20, 5'-CTTCCAGGCACTGAAAGAGG-3' (forward) and 5'-ATCTTCCACCGAGCTGTGTC-3' (reverse); Vimentin 5'-GCGTGACGTACGTCAGCAATATGA-3' (forward) and 5'-GTTCCAGGGACTCATTGGTTCCTT-3' (reverse); and GAPDH 5'-TGCACCACCAACTGCTTAGC-3' (forward) and 5'-ACAGTCTTCTGGGTGGCAGTG (reverse).

Immunohistochemistry

Twenty formalin-fixed and paraffin-embedded fibrotic renal tissue samples were obtained from CKD patients who underwent a renal biopsy at the Second Affiliated Hospital of Soochow University from August 2019 to October 2021. All procedures performed in this study involving human participants were in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the ethics committee of the Second Affiliated Hospital of Soochow University (No. JD-HG-2021-52) and informed consent was taken from all the patients.

Immunohistochemical staining was performed to detect expression of ISG20 and CORO1A according to a previously described method (7). The staining intensity was scored as 0 (negative), 1 (weak), 2 (medium), or 3 (strong), whereas the staining area was scored as 0 (0% of the staining area), 1 (1–25%), 2 (26–50%), 3 (51–75%), and 4 (76–100%). The final scores for ISG20 and CORO1A expression were obtained by calculating the products of the scores of staining intensity and staining area.

Masson staining

To evaluate the degree of interstitial fibrosis in fibrotic renal tissues, 3-μm-thick sections were generated for Masson trichrome staining. Aipathwell (Wuhan servicebio technology Co., Ltd., China) was used to quantify the positive staining area. Thereafter, the proportion of positive staining was calculated.

Distribution of immune cells in RIF

The single sample gene set enrichment analysis (ssGSEA) algorithm in "GSVA" R package was applied to calculate the infiltration levels of 28 immune cell types in RIF and normal samples, including immature B cells, memory B cells, activated B cells, central memory CD8+ T cells, central memory CD4+ T cells, activated CD4+ T cells, activated CD8+ T cells, NK cells, effector memory CD8+ T cells, MDSCs, type 2 T helper cells, regulatory T cells, effector memory CD4+ T cells, gamma delta T cells, CD56dim NK cells, CD56bright NK cells, type 1 T helper cells, type 17 T helper cells, monocytes, macrophages, NK T cells, immature DCs, activated DCs, plasmacytoid DCs, mast cells, T follicular helper cells, eosinophils, and neutrophils. The correlations of the infiltration levels among the 28 immune cells were calculated. Comparisons of immune cell infiltration between RIF and normal samples were conducted using the Wilcoxon test. Immune cells with a P value <0.05 were considered to be differentially infiltrated immune cells (DIICs). The t-SNE method was then applied to cluster and visualize the distribution of DIICs in the GEO cohort. Spearman correlations between DIICs and diagnostic biomarkers were calculated.

Statistical analysis

Both the correlations between the module eigengenes and sample traits and the correlations among immune cells were calculated by Pearson's correlation coefficient. The comparisons of immune cell infiltration between RIF and normal samples were calculated by Wilcoxon test. Student's *t*-test (two-sided) was performed to compare the data from different groups in the biochemical experiments. The values are expressed as mean ± standard deviation (SD). All values of P<0.05 were considered to be statistically significant.
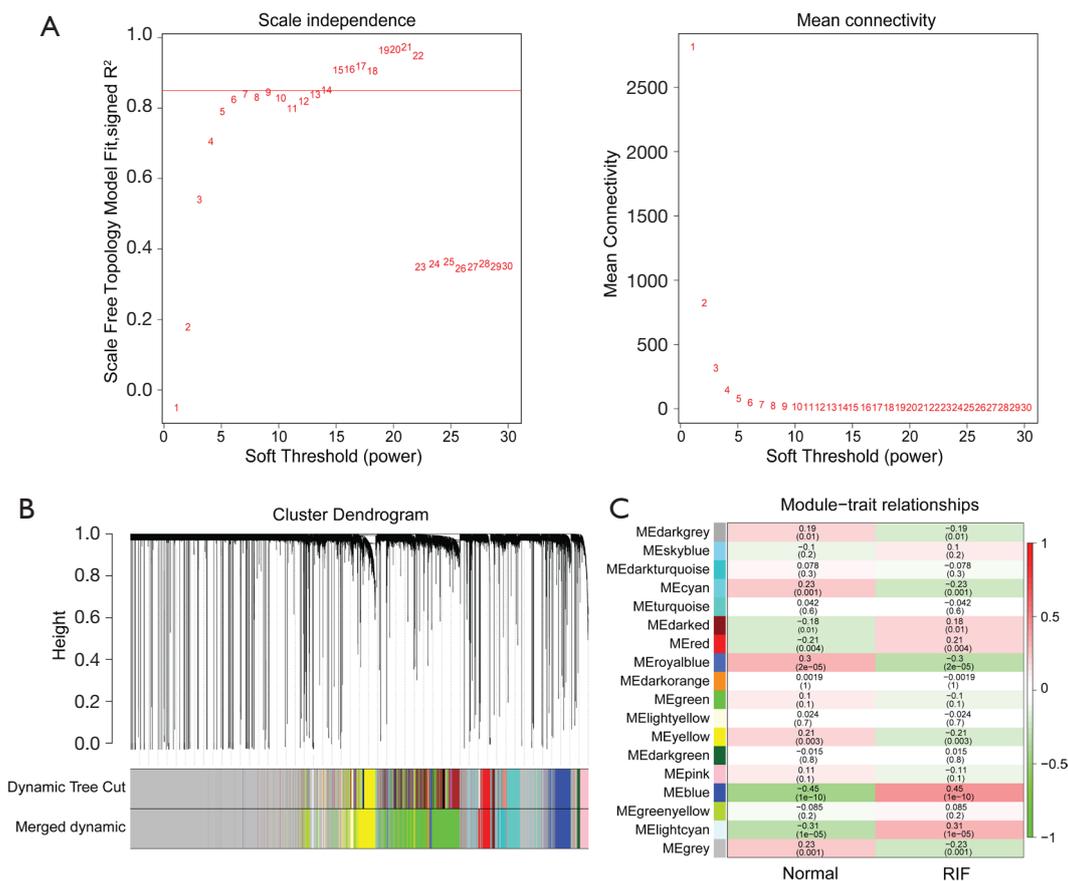
Figure 1 Identification of the key modules associated with RIF. (A) Determination of soft-threshold power via WGCNA. Analysis of the scale-free index and mean connectivity for various soft threshold powers (β) was performed. (B) Identification of the gene co-expression modules via hierarchical cluster analysis. Each branch of the tree diagram represents genes, and genes clustered into the same module are assigned the same module color. (C) Heatmap of the correlation between the module eigengenes and clinical traits of RIF. The Pearson correlation coefficient between the module eigengenes and sample traits were calculated, and modules with |cor| >0.3 and P value <0.05 were considered as key modules related to RIF. RIF, renal interstitial fibrosis; WGCNA, weighted gene co-expression network analysis.

Results

Identification of the key modules associated with RIF

By using the pick soft threshold function of WGCNA, we found that the optimal soft threshold power was 14, with an R^2 of approximately 0.85 (*Figure 1A*). After merging similar modules, 18 modules were identified from the co-expression network (Figure S2 and *Figure 1B*). According to the module-trait relationships in *Figure 1C*, the blue module cor =0.45; P value <0.05) and light cyan module (cor =0.31; P value <0.05) were found to be relevant to RIF. Therefore, 746 genes in the blue module and 390 genes in the light cyan module were extracted. Then, using |GS| >0.2 and |MM| >0.8, a total of 121 genes were selected from the

blue and light cyan modules for downstream analysis.

Identification of candidate key genes in RIF

We explored the DEGs involved in the RIF. A total of 523 DEGs were identified, including 386 upregulated and 137 downregulated genes in RIF samples compared to normal samples (https://cdn.amegroups.cn/static/public/atm-22-366-1.xlsx) (*Figure 2A*). The expression levels of the top 100 DEGs in each sample are displayed in a heatmap (*Figure 2B*). To screen for RIF-related DEGs, we overlapped 121 genes identified via WGCNA and 523 DEGs and obtained a total of 78 genes that may have been key to the progression of RIF (candidate key genes;
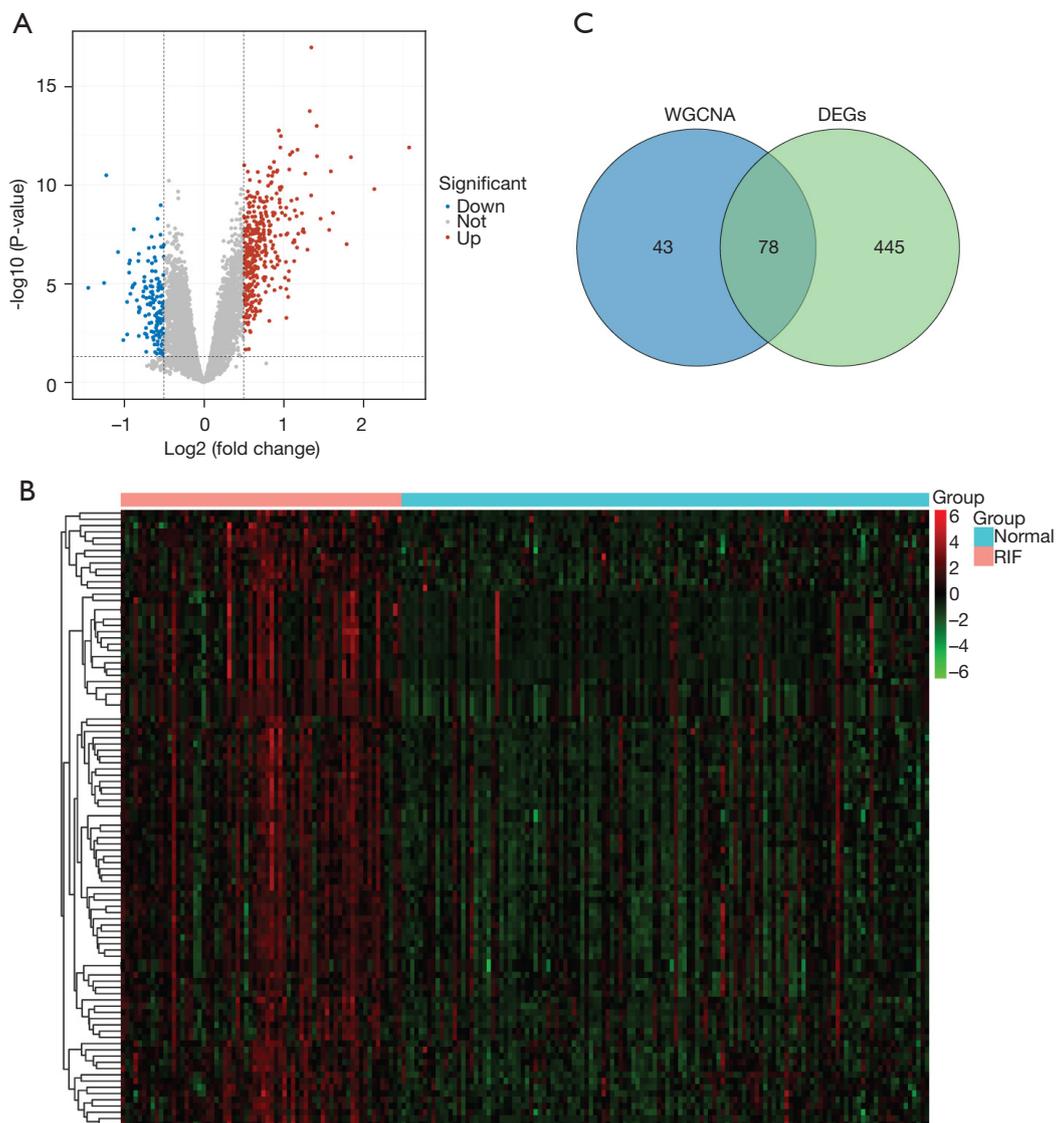
Page 6 of 16

Hu et al. Key biomarkers identification in renal interstitial fibrosis

Figure 2 Identification of candidate key genes in RIF. (A) Volcano plot of DEGs. Volcano plot illustrates the DEGs between normal and RIF samples. Red and blue dots above the dashed curves represent proteins significantly upregulated or downregulated in RIF, with fold change >2. (B) Heatmap of the top 100 DEGs. Each row represents a gene, and each column represents a sample. The red and green colors of the tile indicate high or low expression, respectively. (C) Venn diagrams of RIF-related DEGs. A Venn diagram was used to identify genes shared between WGCNA and DEGs. The blue circle indicates genes from WGCNA, and the green circle indicates DEGs. RIF, renal interstitial fibrosis; WGCNA, weighted gene co-expression network analysis; DEG, differentially expressed gene.

Figure 2C). The potential biological functions of candidate key genes in RIF were investigated using clusterProfiler. A total of 378 biological processes, 29 cellular components, and 11 molecular functions were significantly enriched (https://cdn.amegroups.cn/static/public/atm-22-366-2.xlsx). The top 10 biological processes were closely associated with immunity and inflammation, such as leukocyte proliferation,

regulation of immune effector processes, regulation of lymphocyte proliferation, regulation of mononuclear cell proliferation, regulation of innate immune response, and neutrophil degranulation (*Figure 3A*). Further, candidate key genes were found to be significantly enriched in 32 KEGG pathways (https://cdn.amegroups.cn/static/public/atm-22-366-3.xlsx). Consistent with the GO results, the top
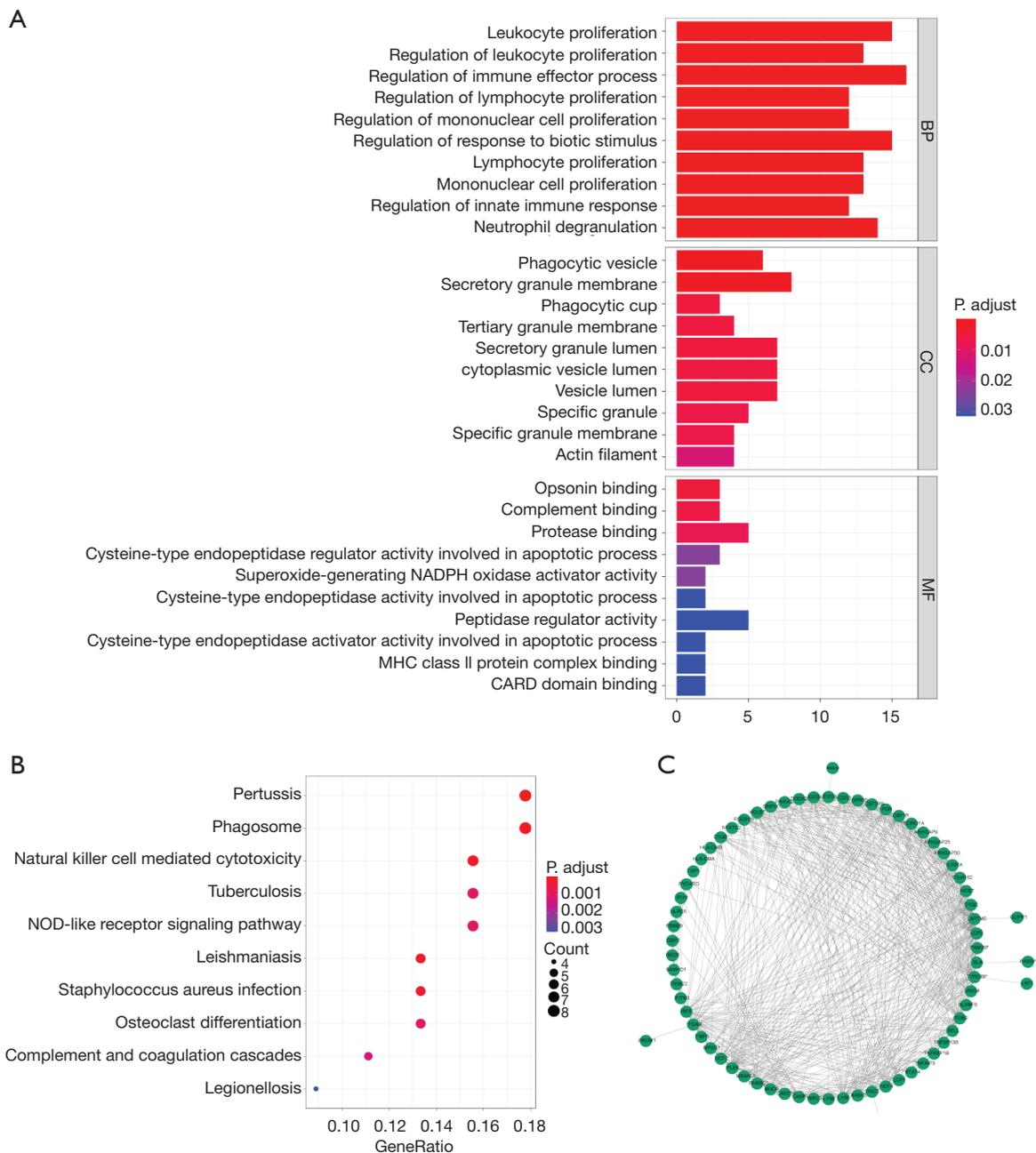
Figure 3 Functional analysis of candidate key genes in RIF. (A,B) GO annotation and KEGG pathway enrichment analysis of key genes. The top 10 enriched GO BP, CC, and MF terms (A) as well as KEGG (B) pathways were analyzed. (C) PPI network analysis of the key genes in RIF. A PPI network between the 78 RIF-related DEGs was constructed using the STRING database (http://www.string-db.org/) and visualized using Cytoscape software. BP, biological process; CC, cell component; MF, molecular function; NADPH, nicotinamide adenine dinucleotide phosphate; MHC, major histocompatibility complex; CARD, caspase recruitment domain; NOD, nucleotide oligomerization domain; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; PPI, protein-protein interaction; RIF, renal interstitial fibrosis; DEG, differentially expressed gene.

Page 8 of 16

Hu et al. Key biomarkers identification in renal interstitial fibrosis

10 pathways related to immune diseases and inflammation, such as pertussis, leishmaniasis, NK cell-mediated cytotoxicity and NOD-like receptor signaling pathways (*Figure 3B*). These results indicate that immunity and inflammation play an important role in RIF. In the protein-protein interaction (PPI) network, 78 candidate key genes in RIF were found to interact with each other (*Figure 3C*).

Identification of ISG20 and CORO1A as key biomarkers of RIF

The 78 candidate key genes were employed in the LASSO, RF, and SVM-RFE analyses to screen hub signatures in RIF. First, we performed the LASSO algorithm, which showed nine signatures, including *CORO1A*, *ISG20*, *LCP1*, *PLEK*, *PROM1*, *RASSF2*, *RASSF2*, *TIMP1*, and *TNFAIP3* (*Figure 4A*,*4B*). Using RF, we selected the top five most important variables, which were *PROM1*, *CSF1R*, *CORO1A*, *ISG20*, and *TIMP1* (*Figure 4C*). The ROC of the RF model showed good performance, with an AUC of 0.817 (*Figure 4D*). We also performed the SVM-RFE algorithm and identified 41 gene signatures at 5-fold cross-validation (*Figure 4E*,*4F*). After overlapping the gene signatures selected by LASSO, RF, and SVM-RFE, *PROM1*, *CORO1A*, *ISG20*, and *TIMP1* were identified as hub signatures in RIF (*Figure 5A*). Subsequently, ROC curves were then used to evaluate the diagnostic value of the hub signature. The AUC values of *CORO1A*, *ISG20*, *PROM1*, and *TIMP1* were 0.816, 0.832, 0.726, and 0.763, respectively (*Figure 5B-5E*), suggesting that *ISG20* and *CORO1A* had better performance at distinguishing RIF and normal samples. Therefore, *ISG20* and *CORO1A* were identified as diagnostic biomarkers of RIF. The expression levels of *ISG20* and *CORO1A* were markedly elevated in RIF samples compared with normal samples (*Figure 5F*,*5G*).

Validation of the expression of ISG20 and CORO1A in fibrotic tubular cells and renal tissue

To further determine the contribution of *ISG20* and *CORO1A* to RIF, we validated the expression levels of ISG20 and CORO1A in fibrotic tubular cells and renal tissue. The mRNA expression levels of *ISG20* and *CORO1A* in TGF-β-stimulated HK-2 cells were significantly higher than levels in the control (*Figure 6A*). Immunoblotting analyses further showed that fibrotic tubular cells expressed higher protein levels of ISG20 and CORO1A (*Figure 6B*). Finally, 20 fibrotic renal tissues were selected from patients

with CKD receiving renal biopsy. The characteristics of included patients were showed in https://cdn.amegroups.cn/static/public/atm-22-366-4.xlsx. The samples were divided into two groups based on the median score of the percentages of positive staining for Masson on wax-embedded renal tissues. Samples with a higher amount of Masson staining had higher CORO1A and ISG20 protein levels (*Figure 6C*,*6D*), indicating that ISG20 and CORO1A were positively correlated with the degree of renal fibrosis. Collectively, these data further highlight the key role of *ISG20* and *CORO1A* in RIF.

Identification of potential drugs for the treatment of RIF

We predicted the potential 77 and 46 chemicals that could either increase or decrease the expression of *ISG20* and *CORO1A*, respectively, using the CTD database. Based on their interaction, we constructed a chemical-diagnostic biomarker network, which showed that *ISG20* and *CORO1A* had common and specific interacting chemicals (*Figure 7*). Based on the elevated expression patterns of *ISG20* and *CORO1A* in RIF, 29 and 25 chemicals may serve as potential drugs that can reduce the expression of *ISG20* and *CORO1A*, respectively (https://cdn.amegroups.cn/static/public/atm-22-366-5.xlsx), and ultimately treat RIF. Notably, antirheumatic agents, cadmium, chenodeoxycholic acid, dimethylselenide, methotrexate, methyl methanesulfonate, tetrachlorodibenzodioxin, and tretinoin were common chemicals targeting both *ISG20* and *CORO1A* (https://cdn.amegroups.cn/static/public/atm-22-366-6.xlsx).

Distribution of immune cells in RIF

Based on the GO and KEGG results, we investigated and compared the immune microenvironment between RIF and normal samples. Using ssGSEA, the infiltration levels of 28 immune cells in the RIF and normal samples were calculated (*Figure 8A* and Figure S3). Moderate to high correlations were found among these immune cells (*Figure 8B*), indicating their interactions in the immune microenvironment. Further, the abundance of 21 immune cells was found to significantly differ between RIF and normal samples, including activated B cells, central memory CD4+ T cells, activated CD4+ T cells, activated CD8+ T cells, NK cells, effector memory CD8+ T cells, MDSCs, type 2 T helper cells, regulatory T cells, gamma delta T cells, CD56dim NK cells, CD56bright NK cells, type
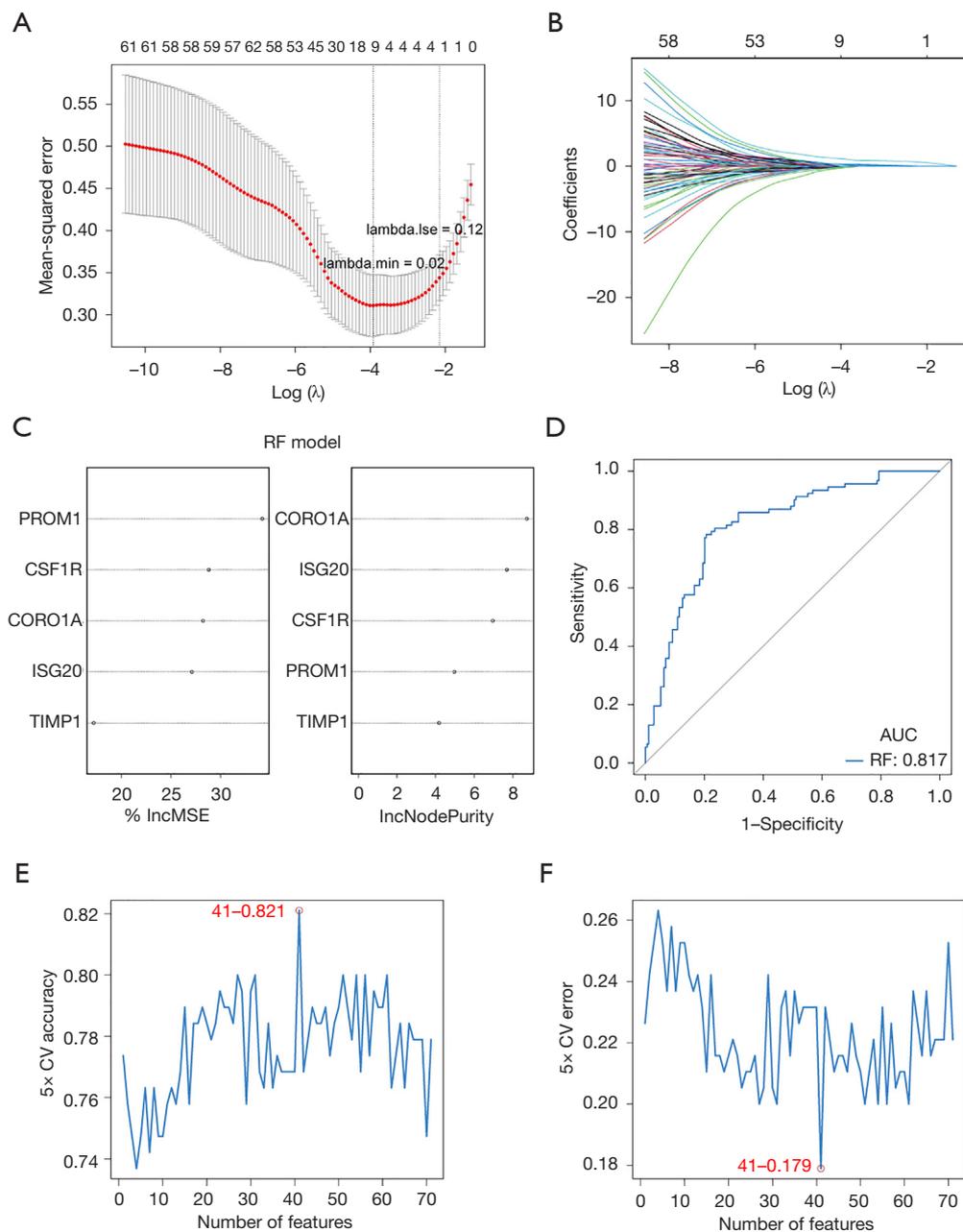
Figure 4 Selection of the hub gene signatures by machine learning. (A) Tuning parameter selection in the LASSO model. The mean-squared error is plotted against log (λ), where λ is the tuning parameter. Mean-squared error values are shown, with error bars representing SE. The dotted vertical lines are drawn at the optimal values by minimum criteria and 1-SE criteria. (B) LASSO coefficient profiles of key genes in RIF. The coefficients are plotted against log (λ). (C) Identification of the relative important variables via RF. The importance of variables was assessed based on an increase in mean squared error (%IncMSE) and increase in node purity (IncNodePurity), respectively. (D) Evaluation of the RF model through the ROC. The ROC of the RF model showed its good performance, with AUC of 0.817. (E,F) Identification of relative important variables by deleting SVM-generated eigenvectors in conjunction with 5-fold cross-validation. The accuracy (E) and error (F) of 5-fold cross-validation were plotted against number of features, respectively, which showed that the number of features was set at 41, with the highest accuracy (0.821) and lowest error (0.179). LASSO, least absolute shrinkage and selection operator; RIF, renal interstitial fibrosis; RF, random forest; ROC, receiver operating curve; SVM, support vector machine.
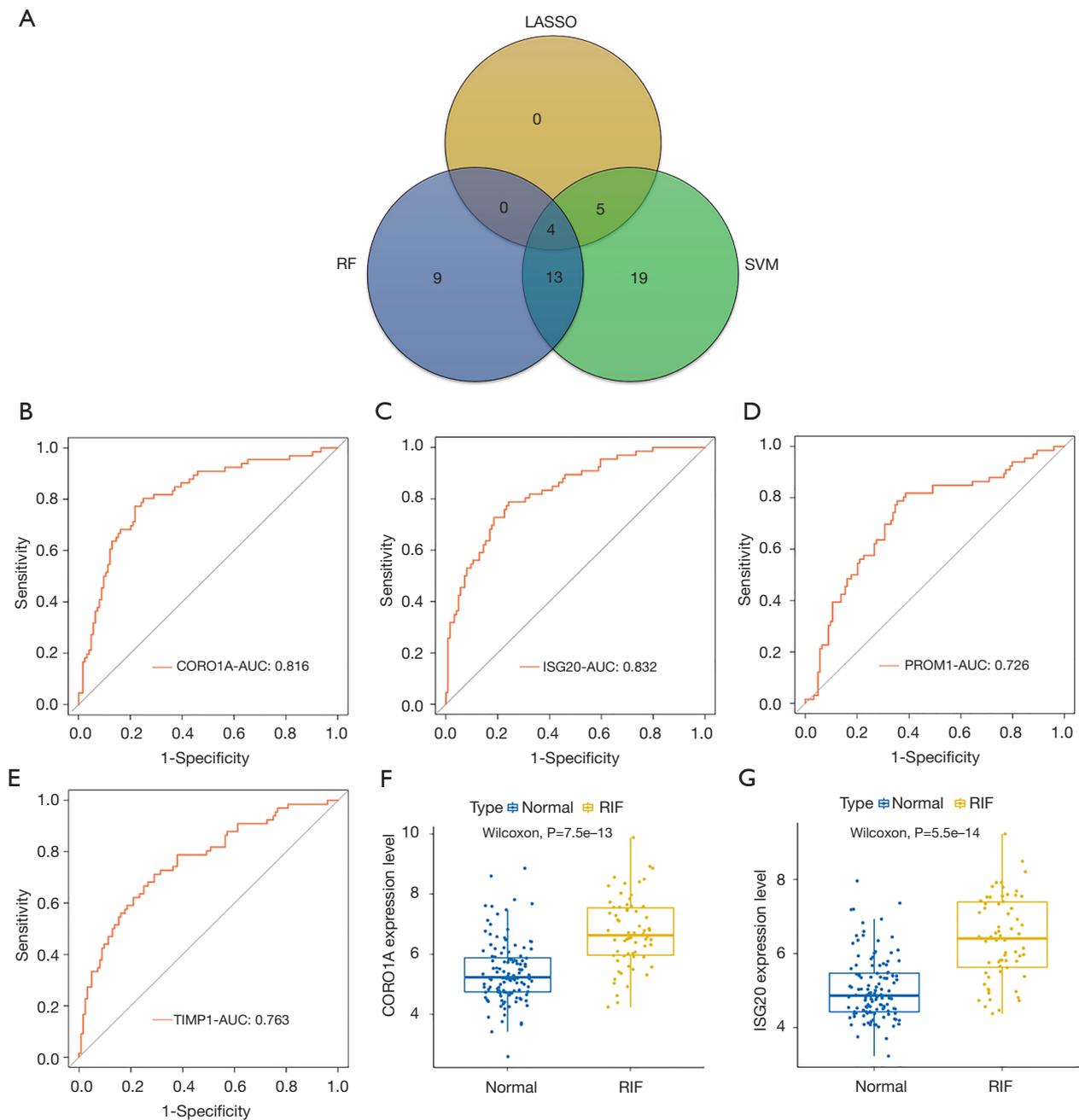
Page 10 of 16

Hu et al. Key biomarkers identification in renal interstitial fibrosis

Figure 5 Identification of *ISG20* and *CORO1A* as key biomarkers of RIF. (A) Venn diagrams of key biomarkers shared by LASSO, RF, and SVM-RFE. *PROM1*, *CORO1A*, *ISG20*, and *TIMP1* were identified as hub signatures in RIF selected by LASSO, RF, and SVM-RFE. (B-E) Evaluation of the diagnostic value of the hub signatures through the ROC. The AUCs of *CORO1A*, *ISG20*, *PROM1*, and *TIMP1* were 0.816, 0.832, 0.726, and 0.763, respectively. (F,G) Validation of the expression of diagnostic biomarkers in GSE22459 and GSE76882. The expression levels of *CORO1A* (F) and *ISG20* (G) were markedly elevated in RIF samples compared with normal samples. LASSO, least absolute shrinkage and selection operator; RF, random forest; SVM-RFE, support vector machine-recursive feature elimination; RIF, renal interstitial fibrosis; ROC, receiver operating curve; AUC, areas under the curve.
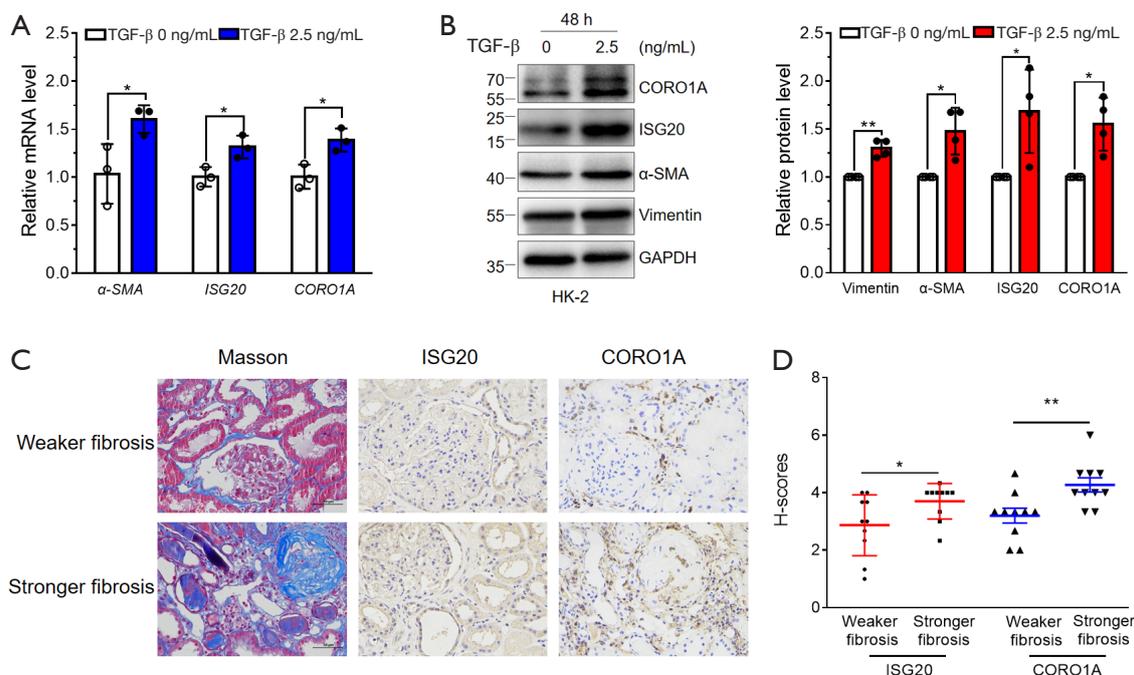
Figure 6 Validation of the expression of CORO1A and ISG20 in fibrotic tubular cells and renal tissue. (A) *CORO1A* and *ISG20* mRNA expression in TGF-β-stimulated HK-2 cells. HK-2 cells were stimulated with TGF-β (0 and 2.5 ng/mL) for 48 h. Relative mRNA level of *α-SMA*, *ISG20*, and *CORO1A* in the above cells was measured by qPCR. *α-SMA* was used as a fibrosis marker induced by TGF-β. Mean ± SDs were obtained from three technical replicates. Student's *t*-test (two-sided), *, P<0.05. (B) CORO1A and ISG20 protein expression in TGF-β-stimulated HK-2 cells. HK-2 cells were stimulated with TGF-β (0 and 2.5 ng/mL) for 48 h. Cell lysates from the above cells were immunoblotted against Vimentin, α-SMA, ISG20, CORO1A, and GAPDH. Vimentin and α-SMA were used as fibrosis markers induced by TGF-β, and GAPDH was used as a loading control. Mean ± SDs are depicted and P value was calculated using Student's *t*-test (two-sided) with data from four biological replicates. *, P<0.05; **, P<0.01. (C,D) CORO1A and ISG20 protein expression in fibrotic renal tissue. Renal tissues from CKD patients with different degree of fibrosis were stained for CORO1A and ISG20 by immunohistochemistry. Representative images of Masson staining and immunohistochemistry staining for CORO1A and ISG20 in renal tissues (C). Scale bar: 50 μm. The expression level of CORO1A and ISG20 (D) in renal tissues with different degrees of fibrosis. Student's *t*-test (two-sided), *, P<0.05; **, P<0.01. TGF-β, transforming growth factor-β; α-SMA, *α*-smooth muscle actin; CKD, chronic kidney disease.

1 T helper cells, monocytes, macrophages, NK T cells, immature DCs, activated DCs, plasmacytoid DCs, mast cells, and T follicular helper cells (*Figure 8C*). Based on the t-SNE results, the GEO cohort could be classified into two distinct clusters (RIF and normal) (*Figure 8D*), which indicates that DIIC may also serve as a biomarker in the diagnosis of RIF. We also calculated the correlations between DIICs and the two key biomarkers. However, the results showed that only a weak relationship existed between most DIIC and the two key genes. Further, CD56dim NK cells and plasmacytoid DCs were found to be positively correlated with *ISG20* and *CORO1A* to some degree (*Figure 8E*).

Discussion

RIF is a crucial and complex metabolic change in the late stages of CKD. Therefore, there is an urgent need to better understand the detailed mechanisms to develop novel strategies to diagnose and treat RIF. Based on datasets downloaded from the GEO database, *ISG20* and *CORO1A* were identified as key genes in RIF; such findings were further verified via biochemical experiments. Additionally, 29 and 25 chemicals were predicted as potential drugs for the treatment of RIF as they could reduce the expression of *ISG20* and *CORO1A*, respectively. According to the results of ssGSEA, most immune cells were differentially infiltrated in the RIF and normal samples, and the abundance of some
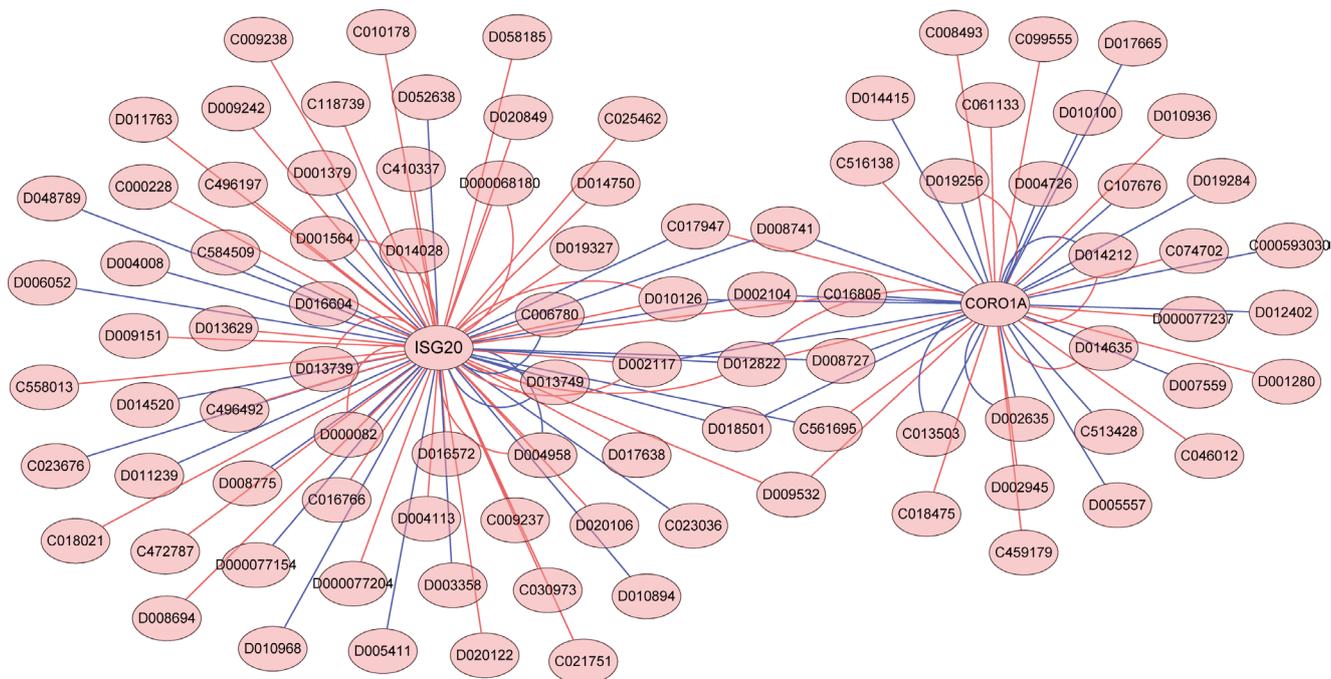
Page 12 of 16

Hu et al. Key biomarkers identification in renal interstitial fibrosis

Figure 7 Identification of potential drugs for the treatment of RIF. A chemical-diagnostic biomarker network was constructed based on potential chemicals targeting *ISG20* and *CORO1A*, and their interactions. The red and blue lines represent the respective increasing or decreasing effects of the chemicals (with chemical IDs in the ellipse) on the expression of target genes (with gene names in the ellipse). RIF, renal interstitial fibrosis.

immune cells was positively correlated with these two key genes.

A previous study using the same GEO datasets screened *CD2*, *CCL5*, and *CCR5* as potential therapeutic target genes in RIF by overlapping DEGs identified from GSE22459 and GSE76882 datasets and PPI network analysis (8). However, unlike the previous study, the candidate key genes of RIF in the present study were first identified by overlapping the RIF-related genes from WGCNA with DEGs from the two datasets. Thereafter, the hub genes related to RIF were selected by machine learning methods, including LASSO, RF, and SVM-RFE. Finally, ROC curves were generated to determine the diagnostic value of the hub gene signature. The two genes, *ISG20* and *CORO1A*, with better performance at distinguishing between RIF and normal samples, were ultimately identified. The differences in the analysis methods and procedures may account for the different hub genes screened. In the present study, the different mRNA and protein expression patterns of *ISG20* and *CORO1A* were further verified in fibrotic tubular cells stimulated with TGF-β and fibrotic renal tissues from patients with varying degrees of fibrosis. In accordance with

the results of bioinformatics analysis, the expression levels of *ISG20* and *CORO1A* were significantly higher in the fibrotic tubular cells; the more severe the fibrosis in renal tissues, the higher the expression level of the two hub genes.

Tryptophan aspartate-containing coat protein (*TACO*, also known as *CORO1A* or *coronin-1*), which is almost exclusively expressed in hematopoietic lineages, has been studied in the immunological field over the last few years and has been found to play an important role in T lymphocyte activation, CD4$^+$ effector/memory T cell differentiation, and innate immunity (9-11). Many reports have demonstrated the contribution of *CORO1A* to inflammation and immune-related diseases, including multiple microorganism infection (12-14), spontaneous osteoarthritis (15), severe combined immune deficiency (16), and cancer (17). Recently, *CORO1A* was identified as a hub gene in chronic classic cardiomyopathy (CCC), characterized by severe cardiac inflammation and myocardial fibrosis (18). However, the relationship between *CORO1A* and renal-related diseases and the underlying molecular mechanisms have not been reported. Nevertheless, our current study identified *CORO1A* as
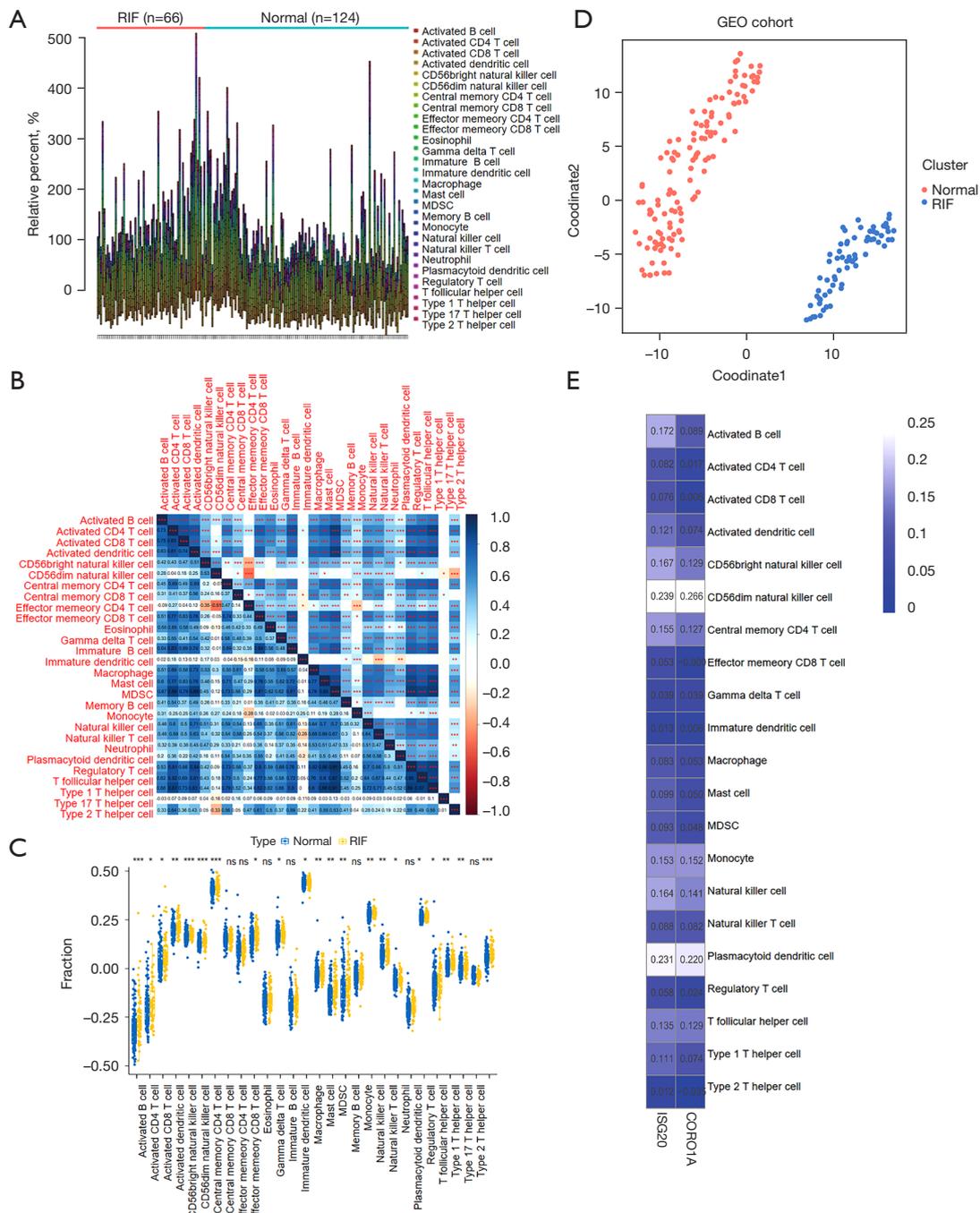
Figure 8 Distribution of immune cells in RIF. (A) The infiltration levels in RIF and normal samples. The ssGSEA algorithm in "GSVA" R package was employed to calculate the infiltration levels of 28 immune cell types in 66 samples with RIF and 124 normal samples from the GSE22459 and GSE76882 datasets. (B) Correlation heatmap of 28 types of immune cells. The numbers in the lower left quarter represent the correlation coefficient between row-defining immune cells and column-defining immune cells, while the statistical significance is highlighted in the upper right quarter. Pearson rank correlation test, *, P<0.05; **, P<0.01; ***, P<0.00. (C) Comparisons of immune cell infiltration between RIF and normal samples. Wilcoxon test, *, P<0.05; **, P<0.01; ***, P<0.001; ns, not significant. (D) Distribution of the cohort differentially infiltrated immune cells (DIICs). t-SNE method was applied to cluster and visualize the distribution of DIICs in the GEO cohort. (E) Correlations between DIICs and diagnostic biomarkers. The numbers in the frame represent the Spearman correlation coefficient between each DIIC and *ISG20* or *CORO1A*. RIF, renal interstitial fibrosis; DIIC, differentially infiltrated immune cell; GEO, Gene Expression Omnibus.

Page 14 of 16

Hu et al. Key biomarkers identification in renal interstitial fibrosis

a novel key gene that was positively correlated with the degree of RIF.

Interferon-stimulated exonuclease gene 20 (*ISG20*), an interferon-regulated gene, codes for a 20-kDa protein with 181 amino acids, induced by both type I (IFN-α/β) and type II (IFN-γ) IFNs (19). According to a previous report (20), ISG20 expression can be induced in renal mesangial cells by polyinosinic-polycytidylic acid [poly (I:C)], an authentic double-stranded RNA that mimics viral infections when applied to cells. Renal mesangial cells have been known to play an important role in the immune and inflammatory responses in the kidney. Additionally, the common immune response module (CRM) scores of seven genes, including *ISG20*, were significantly higher in kidney transplant patients with acute rejection (AR) and progressive interstitial fibrosis and tubular atrophy (pIFTA) than those without rejection response (21). Therefore, these findings may provide indirect evidence of the correlation between *ISG20* and renal fibrosis, a pathological change driven by various immune and inflammatory responses and served as a feature of renal transplantation failure. In addition to the findings of previous studies (20,21), our results imply that *ISG20* might act as a key gene in RIF. Thus, targeting this novel biomarker may be a promising therapeutic strategy for the treatment of RIF.

Wet experiments that identify drugs targeting specific genes are time-consuming and expensive. However, many drug-gene associations remain unobserved or unknown. The use of a database to predict unobserved drugs targeting a specific gene is an important and urgent task. By using the CTD database in the present study, 29 and 25 chemicals were identified to decrease the expression of *ISG20* and *CORO1A*, respectively, including prednisolone, a classic glucocorticoid widely used to delay the progression of renal inflammatory and fibrosis in clinical practice. Most importantly, seven chemicals were found to simultaneously select the two targets. Among them, chenodeoxycholic acid was recently reported to improve renal fibrosis, inflammation, and oxidative stress in rats fed high fructose (22). However, the exact targets and underlying mechanisms remain unknown. Therefore, our work offers potential opportunities and drug targets for the treatment of RIF.

The infiltration of inflammatory cells, including T cells, monocytes/macrophages, DCs, and mast cells, is a major cellular event in tubulointerstitial fibrosis (23). Despite the established correlation between sustained inflammation and fibrotic disease, the role of various inflammatory cells

is highly complex (24). The infiltration and activation of macrophages, T cells, and DCs contribute to fibrogenesis, while infiltrated mast cells tend to attenuate renal fibrosis induced by kidney injury (25,26). Recently, early-stage accumulation of B lymphocytes in the kidney was reported to accelerate monocyte/macrophage mobilization and infiltration, aggravating fibrosis (27). In our study, among the 28 types of immune cells, the infiltration of 21 immune cells was significantly different between RIF and normal samples. Notably, *ISG20* and *CORO1A* were positively correlated with CD56dim NK cells and plasmacytoid DCs to some degree. The literature evidence and our results imply that immune cell infiltration plays an important role in RIF and should be the focus of further studies.

As this study had limitations, further *in vivo* and *in vitro* experiments are needed to explore the function of these two genes and the underlying mechanism in RIF.

Conclusions

The hub genes, *ISG20* and *CORO1A*, were identified in RIF via bioinformatics analysis using the GEO database and biochemical experiments with fibrotic tubular cells and renal tissues. Further, the drugs targeting these two genes were predicted, and the immune cell infiltration pattern of RIF was further described, thereby providing useful information and directions to further investigate the roles of *ISG20* and *CORO1A* in RIF. Overall, our findings provide novel insights into the mechanisms and treatment of patients with RIF.

Acknowledgments

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at https://atm.amegroups.com/article/view/10.21037/atm-22-366/rc

Data Sharing Statement: Available at https://atm.amegroups.com/article/view/10.21037/atm-22-366/dss

Conflicts of Interest: All authors have completed the

ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-22-366/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All procedures performed in this study involving human participants were in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the ethics committee of the Second Affiliated Hospital of Soochow University (No. JD-HG-2021-52) and informed consent was taken from all the patients.

References

1. Ammirati AL. Chronic Kidney Disease. Rev Assoc Med Bras (1992) 2020;66 Suppl 1:s03-9.
2. Ruiz-Ortega M, Rayego-Mateos S, Lamas S, et al. Targeting the progression of chronic kidney disease. Nat Rev Nephrol 2020;16:269-88.
3. Ke B, Zhu N, Luo F, Xu Y, Fang X. Targeted inhibition of endoplasmic reticulum stress: New hope for renal fibrosis (Review). Mol Med Rep 2017;16:1014-20.
4. Humphreys BD. Mechanisms of Renal Fibrosis. Annu Rev Physiol 2018;80:309-26.
5. Wang W, Shen J, Qi C, et al. The key candidate genes in tubulointerstitial injury of chronic kidney diseases patients as determined by bioinformatic analysis. Cell Biochem Funct 2020;38:761-72.
6. Xu J, Lamouille S, Derynck R. TGF-beta-induced epithelial to mesenchymal transition. Cell Res 2009;19:156-72.
7. Yakirevich E, Resnick MB, Mangray S, et al. Oncogenic ALK Fusion in Rare and Aggressive Subtype of Colorectal Adenocarcinoma as a Potential Therapeutic Target. Clin Cancer Res 2016;22:3831-40.
8. Zhang C, Hu X, Qi F, et al. Identification of CD2, CCL5 and CCR5 as potential therapeutic target genes for renal interstitial fibrosis. Ann Transl Med 2019;7:454.
9. Shiow LR, Roadcap DW, Paris K, et al. The actin regulator coronin 1A is mutant in a thymic egress-deficient mouse strain and in a patient with severe combined immunodeficiency. Nat Immunol 2008;9:1307-15.
10. Kaminski S, Hermann-Kleiter N, Meisel M, et al. Coronin 1A is an essential regulator of the TGFβ receptor/SMAD3 signaling pathway in Th17 CD4(+) T cells. J Autoimmun 2011;37:198-208.
11. Pick R, Begandt D, Stocker TJ, et al. Coronin 1A, a novel player in integrin biology, controls neutrophil trafficking in innate immunity. Blood 2017;130:847-58.
12. Cohen JI. Primary Immunodeficiencies Associated with EBV Disease. Curr Top Microbiol Immunol 2015;390:241-65.
13. Fei X, Li Z, Yang D, et al. Neddylation of Coro1a determines the fate of multivesicular bodies and biogenesis of extracellular vesicles. J Extracell Vesicles 2021;10:e12153.
14. Heidari M, Pakdel A, Bakhtiarizadeh MR, et al. Integrated Analysis of lncRNAs, mRNAs, and TFs to Identify Regulatory Networks Underlying MAP Infection in Cattle. Front Genet 2021;12:668448.
15. Wang Y, Wu C, Tao J, et al. Differential proteomic analysis of tibial subchondral bone from male and female guinea pigs with spontaneous osteoarthritis. Exp Ther Med 2021;21:633.
16. Vignesh P, Rawat A, Kumrah R, et al. Clinical, Immunological, and Molecular Features of Severe Combined Immune Deficiency: A Multi-Institutional Experience From India. Front Immunol 2021;11:619146.
17. He Y, Jiang Z, Chen C, et al. Classification of triple-negative breast cancers based on Immunogenomic profiling. J Exp Clin Cancer Res 2018;37:327.
18. Wu J, Cao J, Fan Y, et al. Comprehensive analysis of miRNA-mRNA regulatory network and potential drugs in chronic chagasic cardiomyopathy across human and mouse. BMC Med Genomics 2021;14:283.
19. Zheng Z, Wang L, Pan J. Interferon-stimulated gene 20-kDa protein (ISG20) in infection and disease: Review and outlook. Intractable Rare Dis Res 2017;6:35-40.
20. Imaizumi T, Tanaka H, Mechti N, et al. Polyinosinic-polycytidylic acid induces the expression of interferon-stimulated gene 20 in mesangial cells. Nephron Exp Nephrol 2011;119:e40-8.
21. Sigdel TK, Bestard O, Tran TQ, et al. A Computational

Page 16 of 16

Hu et al. Key biomarkers identification in renal interstitial fibrosis

Gene Expression Score for Predicting Immune Injury in Renal Allografts. PLoS One 2015;10:e0138133.

22. Hu Z, Ren L, Wang C, et al. Effect of chenodeoxycholic acid on fibrosis, inflammation and oxidative stress in kidney in high-fructose-fed Wistar rats. Kidney Blood Press Res 2012;36:85-97.

23. Zeisberg M, Neilson EG. Mechanisms of tubulointerstitial fibrosis. J Am Soc Nephrol 2010;21:1819-34.

24. Andrade-Oliveira V, Foresto-Neto O, Watanabe IKM, et al. Inflammation in Renal Diseases: New and Old Players. Front Pharmacol 2019;10:1192.

25. Liu L, Kou P, Zeng Q, et al. CD4+ T Lymphocytes, especially Th2 cells, contribute to the progress of renal fibrosis. Am J Nephrol 2012;36:386-96.

26. Beghdadi W, Madjene LC, Claver J, et al. Mast cell chymase protects against renal fibrosis in murine unilateral ureteral obstruction. Kidney Int 2013;84:317-26.

27. Han H, Zhu J, Wang Y, et al. Renal recruitment of B lymphocytes exacerbates tubulointerstitial fibrosis by promoting monocyte mobilization and infiltration after unilateral ureteral obstruction. J Pathol 2017;241:80-90.
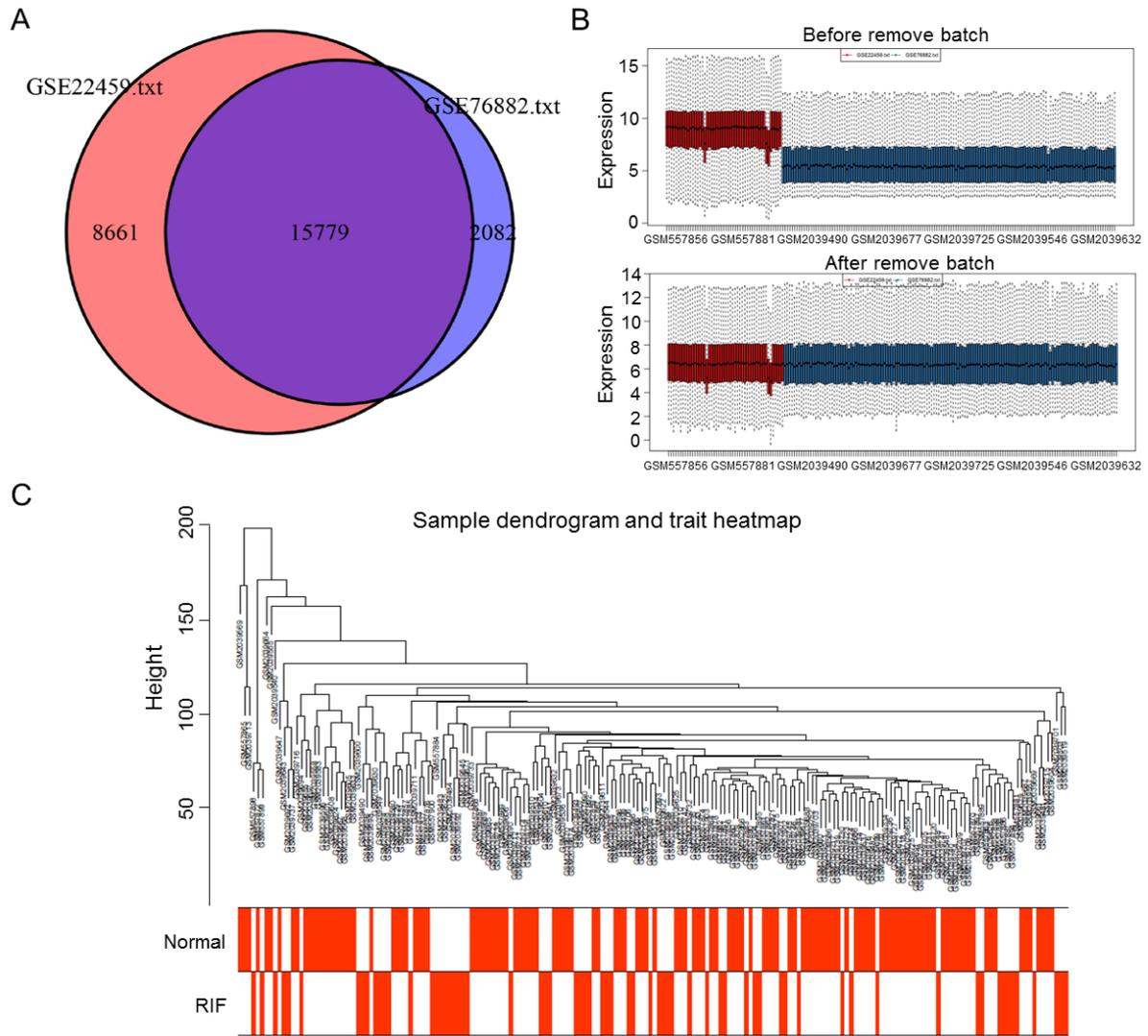
Figure S1 Removal of batch effect of GSE22459 and GSE76882 datasets. (A) Venn diagrams of combined data of GSE22459 and GSE76882. The pink and blue circles indicate GSE22459 and GSE76882 dataset, respectively. 15779 genes shared by the two datasets were chosen for downstream analysis. (B) Removal of batch effect. The Sangerbox online tool (http://sangerbox.com) was applied to remove the batch effect of the two datasets. (C) The construction of sample clustering tree map. A sample clustering tree map was constructed to detect and eliminate outliers.

Clustering of module eigengenes

Figure S2 Clustering of module eigengenes. The 29 modules obtained from co-expression network were merged into 18 modules with 0.2 as the MEDiss Thres indicated by the red line.

Figure S3 Heatmap of the infiltration levels of 28 immune cells in RIF and normal samples. Each row represents an immune cell and each column represents a sample. The red and green color of the tile indicates high or low expression, respectively.