scientific reports

Check for updates

OPEN

Identification of key candidate genes for IgA nephropathy using machine learning and statistics based bioinformatics models

Md. Al Mehedi Hasan[1], Md. Maniruzzaman[1,2] & Jungpil Shin[1]✉

Immunoglobulin-A-nephropathy (IgAN) is a kidney disease caused by the accumulation of IgAN deposits in the kidneys, which causes inflammation and damage to the kidney tissues. Various bioinformatics analysis-based approaches are widely used to predict novel candidate genes and pathways associated with IgAN. However, there is still some scope to clearly explore the molecular mechanisms and causes of IgAN development and progression. Therefore, the present study aimed to identify key candidate genes for IgAN using machine learning (ML) and statistics-based bioinformatics models. First, differentially expressed genes (DEGs) were identified using limma, and then enrichment analysis was performed on DEGs using DAVID. Protein-protein interaction (PPI) was constructed using STRING and Cytoscape was used to determine hub genes based on connectivity and hub modules based on MCODE scores and their associated genes from DEGs. Furthermore, ML-based algorithms, namely support vector machine (SVM), least absolute shrinkage and selection operator (LASSO), and partial least square discriminant analysis (PLS-DA) were applied to identify the discriminative genes of IgAN from DEGs. Finally, the key candidate genes (FOS, JUN, EGR1, FOSB, and DUSP1) were identified as overlapping genes among the selected hub genes, hub module genes, and discriminative genes from SVM, LASSO, and PLS-DA, respectively which can be used for the diagnosis and treatment of IgAN.

Immunoglobulin-A-nephropathy (IgAN) is one of the major public health problems. IgAN is also known as Berger's disease[1,2]. It is a kidney disease caused by the accumulation of IgAN deposits in the kidneys, which causes inflammation and damage to the kidney tissues. IgAN is the most common primary glomerulonephritis that can progress to renal failure worldwide[3,4]. IgAN is sometimes associated with different kinds of diseases such as heart disease[5,6], liver cirrhosis[6,7], coeliac disease[6,8], skin disease[6]. About, 20–47% of primary glomerular diesases are responsible for IgAN, which is mainly characterized by hypertension, hematuria, proteinuria, and failure of the renal[9,10]. About 20–40% of people with IgAN have end-stage renal disease after 10–20 years[11]. The overall prevalence of IgA nephropathy varies from regions to regions[12]. The highest prevalence's of IgAN are found in Asia region (especially, in China and Japan) and its prevalence has been diagrammatically increased over past three decades[13–15]. It is noted that the risks of deaths have been increased among patients with IgAN[16]. As a result, we need to know the molecular mechanism about the development and progression of IgAN in order to diagnose IgAN patients properly and decrease the death rate. However, molecular mechanism can be studied properly by knowing the key genes or biomarkers for the development and progression of IgAN. Despite numerous studies examining the molecular characteristics of IgAN, the mechanism underlying IgAN development and progression remains a challenging issue[15]. Therefore, it is urgent to propose an effective tool for determining potential or key candidate genes of IgAN in order to understand molecular mechanism of IgAN.

Bioinformatics analysis is a powerful approach for predicting molecular pathways and gene connections. This approach is widely used to predict novel candidate genes and pathways associated with different cancers like breast[17], gastric[18], cervical[19], and so on. Recently, this approach has increasingly revealed the molecular pathways underlying kidney disease[20,21]. Various studies were conducted for the identification of key hub genes for IgAN patients. Qian et al. suggested twenty-one hub genes as well as identified five key candidate genes which were strongly correlated with IgAN patients[10]. Zhang et al. investigated ten hub genes of IgAN and proposed four novel biomarkers that may be played an essential roles in the progression of IgAN and could be used as

[1]School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Fukushima 965-8580, Japan. [2]Statistics Discipline, Khulna University, Khulna 9208, Bangladesh. ✉email: jpshin@u-aizu.ac.jp

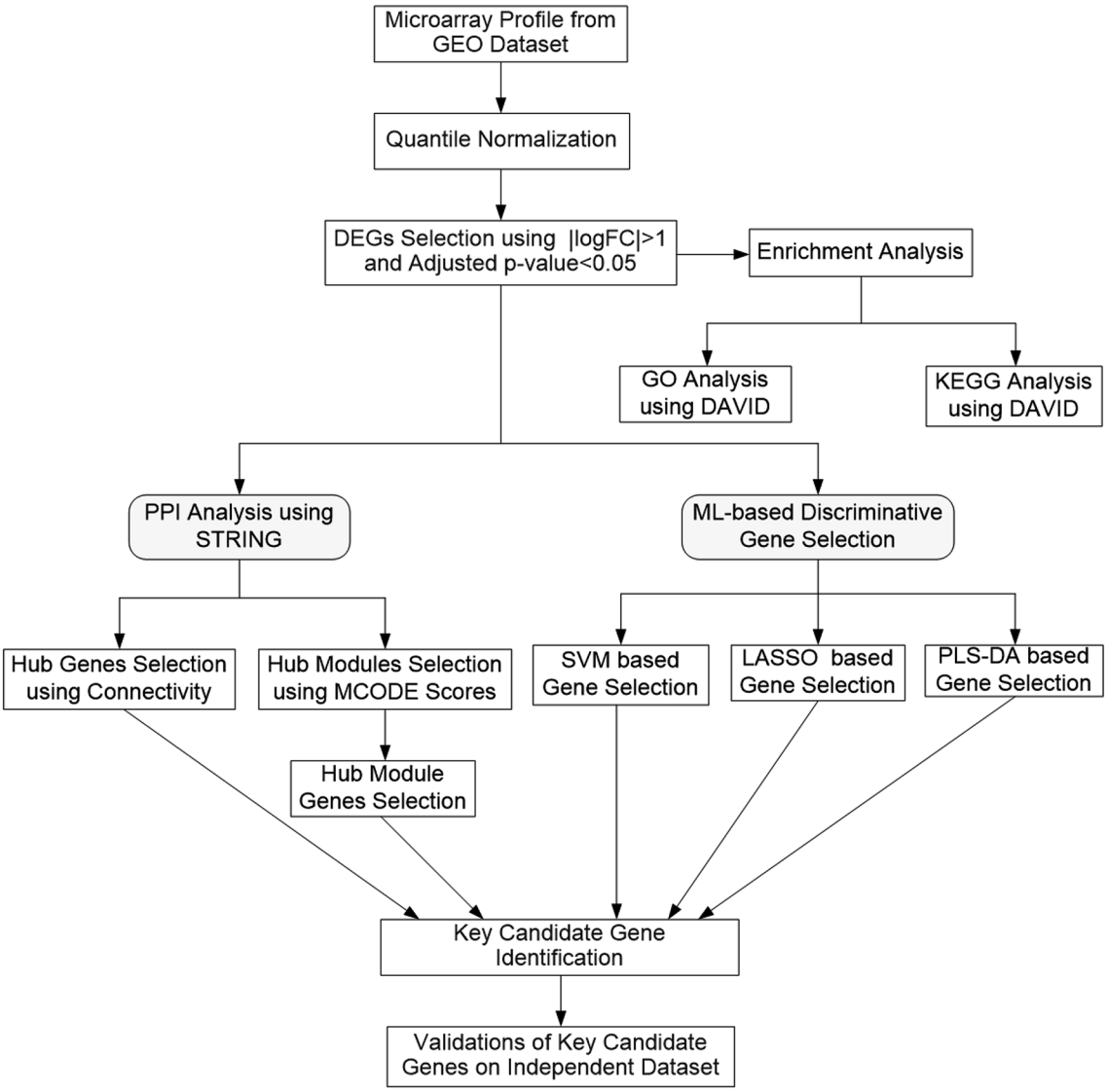

Figure 1. Flowchart of data preparation, processing, analysis, and validation.

potential biomarkers for IgAN diagnosis and treatment[20]. Chen et al. suggested six biomarkers that were also related to the pathogenesis of IgAN[22]. Chen et al. also suggested plausible new drugs (thapsigargin, ciclopirox, and ikarugamycin) for the treatment of IgAN[22]. All of these previous studies demonstrated key biomarkers of IgAN using bioinformatics analysis[10,20,22–25] and showed different gene sets as key candidate genes. All researcher identified their potential biomarkers or genes using only hub genes, determined by the degree of connectivity in the PPI netwrok. In recent years, machine learning (ML)-based techniques have gained more popularity to ease one of the important challenges associated with study of genetic data: extraction of meaningful genes[26–28]. Since the set of identified key genes by existing works are different, there is still some scope to identify genes more confidently using ML and statistics-based bioinformatics models.

In the current study, we selected one microarray gene expression (MGE) dataset from the Gene Expression Omnibus (GEO) database to identify the key candidate genes of IgAN. First, we identified DEGs for IgAN patients. Then, we used Database for Annotation, Visualization, and Integration Discovery (DAVID) to discover the functions of the DEGs and obtained Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses. Using the STRING database, we constructed a protein–protein interaction (PPI) network and identified hub genes from Cytoscape using the degree of connectivity as well as the most potential modules using Molecular Complex Detection (MCODE). We also identified the hub modules and their associated genes from the selected potential modules. We applied three ML-based algorithms to identify the significant genes for IgAN patients. The objective of this research was to determine the potential key candidate genes or biomarkers that can be used to diagnose and treat IgAN. Figure 1 summarized the data preparation, processing, analysis, and validation.

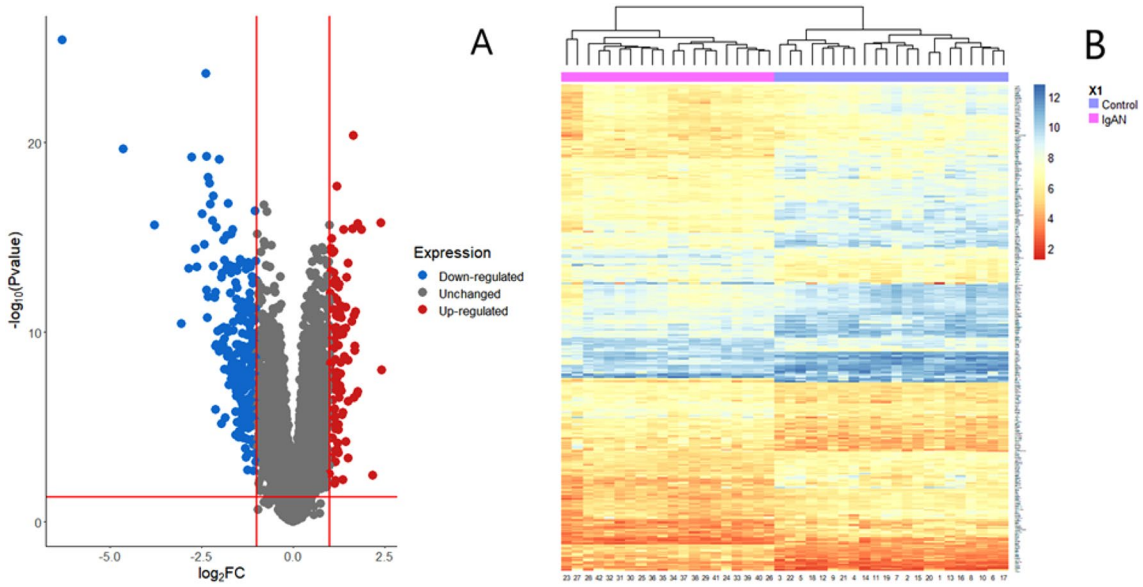

Figure 2. Identification and hierarchical clustering of DEGs for IgAN patients. (**A**) Volcano plot of DEGs which were generated using "ggplot2 version 3.3.6" package in R[63] (https://cran.r-project.org/package=ggplot2) . Dodger blue represents down-regulated, gray represents no significant genes, and fire brick represents up-regulated DEGs. (**B**) Heatmap of the DEGs for IgAN patients which were generated using "NMF" version 0.24.0 package in R[64] (https://cran.r-project.org/package=NMF). The horizontal axis shows the number of patients and the vertical axis shows DEGs.

Results

Experimental settings. For this experiment, the R-programming language version 4.1.2 was used for all statistical analysis. As the operating system, Windows 10 version 21H1 (build 19043.1151) 64 bit was used. In terms of hardware, an Intel (R) Core (TM) i5-10400 processor with 16 GB of RAM was used. In this study, we used three GEO datasets: GSE93798, GSE116626 and GSE35487. We selected the key candidate genes from the GSE93798 dataset. Another two independent test datasets: GSE116626 and GSE35487 were used for the validation of key candidate genes.

Identification of DEGs. Using the cutoff of adjusted p-value < 0.05 and $|\log FC| > 1$, a total of 348 DEGs were identified for IgAN patients. Among them, 107 genes were up-regulated and 241 genes were down-regulated. The volcano plot and heatmap of the DEGs for IgAN patients and healthy controls was presented in Fig. 2A,B.

Go term enrichment and KEGG pathway analysis. We imported the DEGs into the DAVID for the enrichment analysis of GO and KEGG pathways. To determine the significant GO terms and KEGG pathway, we considered the cutoff point of p-value < 0.05. In Table 1, the top five significant GO terms of DEGs for biological process (BP), cellular component (CC), and molecular function (MF) were presented. As in BP, the DEGs were significantly enriched in response to inflammatory, response to camp, cytokine-mediated signaling pathway, cellular response to lipopolysaccharide, and neutrophil chemotaxis. In the CC group, the DEGs were mainly enriched in extracellular exosome, region, space, collagen trimer, and blood microparticle. The MF group GO terms, including transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding, zinc ion binding, transmembrane transporter activity, and gextra cellular matrix structural constituent conferring tensile strength, were significantly enriched by DEGs.

The analysis of the KEGG pathway for DEGs was showed in Table 2. We observed that DEGs were mainly involved in glycine, serine and threonine metabolism, age-rage signaling pathway in diabetic complications, protein digestion and absorption, IL-17 signaling pathway, and osteoclast differentiation.

PPI network construction and hub gene selection. PPI networks TSV data file was obtained from STRING and imported to Cytoscape and built a PPI network with 206 nodes and 880 edges (see Fig. 3A). The hub genes were selected using a degree of connectivity > 18. Using this cutoff, we selected 19 hub genes which were shown in detail in Table 3.

Hub module and its associated gene selection. A total of 13 modules/clusters were built using MCODE with the cutoffs: degree = 2, cluster finding = haircut, nodes score = 0.2, K-score = 2, and max depth = 100. The MCODE scores ranged from 3 to 8.44. We selected two significant modules with cutoffs: MCODE scores ≥ 6 and number of nodes ≥ 6 for determining hub module genes (see Table 4). The corresponding PPI

	GO ID	Description	Count	*p*-value
BP	GO:0006954	Inflammatory response	32	8.16×10^{-13}
	GO:0051591	Response to camp	10	5.67×10^{-8}
	GO:0019221	Cytokine-mediated signaling pathway	21	3.78×10^{-7}
	GO:0071222	Cellular response to lipopolysaccharide	16	5.44×10^{-7}
	GO:0030593	Neutrophil chemotaxis	11	7.67×10^{-7}
CC	GO:0070062	Extracellular exosome	93	2.79×10^{-18}
	GO:0005576	Extracellular region	70	8.49×10^{-9}
	GO:0005615	Extracellular space	62	1.58×10^{-7}
	GO:0005581	Collagen trimer	12	3.03×10^{-7}
	GO:0072562	Blood microparticle	14	6.18×10^{-7}
MF	GO:0005201	Extracellular matrix structural constituent	15	8.57×10^{-8}
	GO:0001228	Transcriptional activator activity	27	1.18×10^{-7}
	GO:0008270	Zinc ion binding	36	1.39×10^{-6}
	GO:0022857	Transmembrane transporter activity	13	1.41×10^{-5}
	GO:0030020	Gextra cellular matrix structural constituent conferring tensile strength	7	5.84×10^{-5}

Table 1. GO analysis of DEGs in biological process, cellular component, and molecular function. Top five items were selected based on *p*-value. *GO* gene ontology, *BP* biological process, *CC* cellular component, *MF* molecular function.

Pathway ID	Description	Count	*p*-value
hsa00260	Glycine, serine and threonine metabolism	10	4.50×10^{-7}
hsa04933	Age-rage signaling pathway in diabetic complications	13	6.63×10^{-6}
hsa04974	Protein digestion and absorption	12	4.90×10^{-5}
hsa04657	IL-17 signaling pathway	10	5.45×10^{-4}
hsa04380	Osteoclast differentiation	11	0.0013

Table 2. KEGG pathway analysis of DEGs. Top five items were selected based on *p*-value.

network for module 1 and module 2 were showed in Fig. 3B,C. There were 10 genes in module 1 and 11 genes in module 2. We took the union of module 1 and module 2 and got 21 genes which were considered as hub module genes.

Analysis of discriminative gene selection using ML. *Identifying discriminative genes using SVM.* We applied SVM with RBF kernel on 348 DEGs and computed the classification accuracy for each gene. The gene selection procedure using SVM was already discussed in "Methods" section. The classification accuracy of each gene had sorted and were showed in Fig. 4. We selected 35 discriminative genes out of 348 DEGs whose classification accuracy was greater than 95.0%.

Identifying discriminative genes using LASSO. A total of 348 DEGs were identified between IgAN and control groups to fit LASSO-based logistic regression model. The next step was to determine the optimal values for lambda (λ= 0.008012) using 10-fold CV. Finally, 32 discriminative genes (SRPX2, LYL1, PCDH18, PPP1R10, DUSP1, EMP3, FPR3, NR1H4, C8ORF4, CD44, EGR1, FOSB, FOS, RNF186, DEPDC7, GSTA3, NETO2, CYP27B1, PCK1, C3AR1, CYSLTR1, JUN, TOP2A, CRTAM, CEBPD, LINC01279, SLC19A2, ZFP36, PTGS1, PLD6, FN1, KLF4) with no-zero coefficients were identified in discriminating IgAN and healthy control (see Fig. 5).

Identifying discriminative genes using PLS-DA. PLS-DA was adopted on 348 DEGs to determine the significant genes of IgAN patients. We selected 20 components. Among them, we took the first two PLS-DA components and visualized these two components, which were presented in Fig. 6A. The red points indicated the IgAN patients and the green points indicated the healthy controls (Fig. 6A). PLS-DA can be significantly differentiated IgAN patients from healthy controls. We selected the top 20 most important genes (FOSB, DUSP1, PCDH18, FOS, ZFP36, EGR1, RNF186, CEBPD, LYL1, JUN, CSRNP1, ERRFI1, CYP27B1, PPP1R10, DEPDC7, KLF4, COL1A2, SOX17, APOLD1, and ATF3) for IgAN patients, which were illustrated in Fig. 6B.

Key candidate genes selection. The key candidate genes were identified by overlapping genes according to five methods. Among them, three methods were ML-based algorithms (SVM, LASSO, and PLS-DA) for the

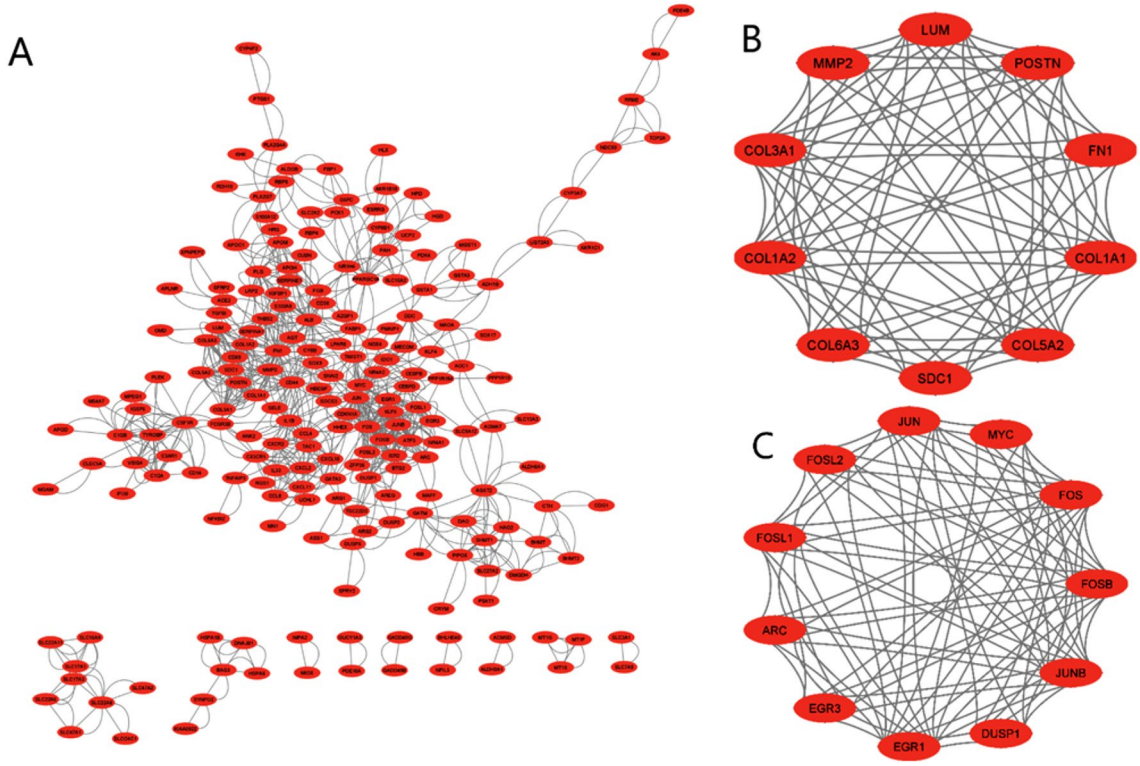

Figure 3. (**A**) PPI network of DEGs, (**B**) Module 1, and (**C**) Module 2. These three figures were generated by Cytoscape 3.9.1[54] (www.cytoscape.org).

SN	Gene	Degree	Betweenness	Closeness
1	FOS	50	0.113	0.314
2	JUN	44	0.164	0.326
3	FN1	38	0.113	0.321
4	ALB	34	0.190	0.330
5	IL1B	32	0.234	0.337
6	EGR1	32	0.012	0.280
7	JUNB	30	0.023	0.291
8	CD44	28	0.074	0.310
9	MMP2	28	0.033	0.288
10	MYC	26	0.076	0.315
11	FOSB	26	0.011	0.275
12	COL1A2	24	0.006	0.264
13	TYROBP	22	0.060	0.223
14	CSF1R	22	0.093	0.248
15	COL1A1	22	0.008	0.269
16	CCL4	20	0.062	0.303
17	ATF3	20	0.044	0.264
18	DUSP1	20	0.036	0.262
19	LUM	20	0.013	0.250

Table 3. List of 19 hub genes which were identified from PPI network based on degree of connectivity.

Cluster	Score	Nodes	Edges	Node IDs
1	8.44	10	76	COL5A2, POSTN, COL6A3, LUM, COL1A1, SDC1, COL3A1, MMP2, FN1, COL1A2
2	8.40	11	84	DUSP1, JUN, JUNB, EGR3, MYC, FOSL2, FOSB, FOSL1, EGR1, FOS, ARC

Table 4. Two modules selected from the PPI network. Score=density × no. of nodes.

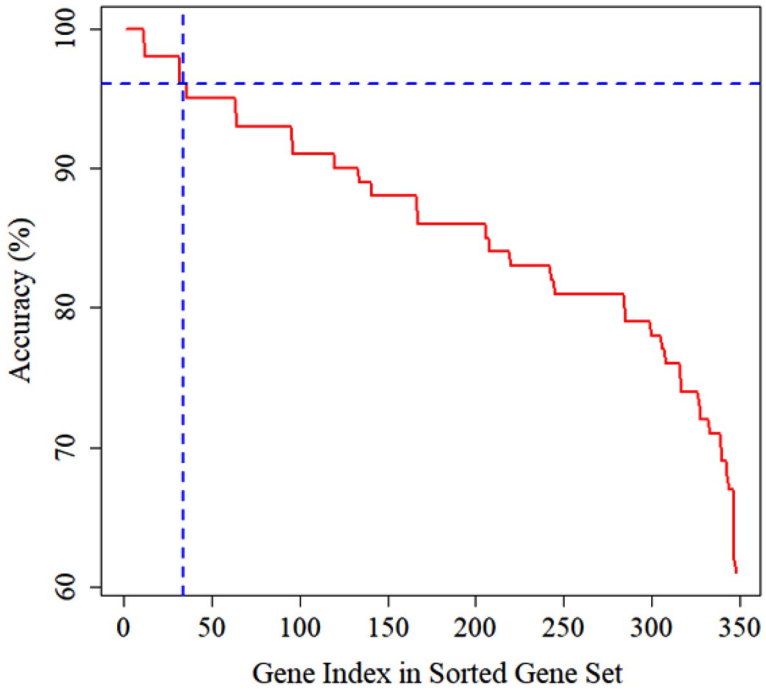

Figure 4. Classification accuracy of SVM for each gene.

identification of discriminative genes. The hub genes were identified using the degree of connectivity from the PPI network and hub module genes were from two significant modules. Five key candidate genes (FOS, JUN, EGR1, FOSB, and DUSP1) were selected, which were shown in Fig. 7A, and their PPI network analysis was also shown in Fig. 7B. These five key candidate genes and their probable significance in IgAN indicated that they could be novel therapeutic target genes. We observed that each key candidate gene was significantly differentiated IgAN patients from healthy controls (Fig. 8A–E). We also performed hierarchical clustering for each candidate gene, which was shown in Fig. 8F.

Validation or confirmation of key candidate genes. The GSE116626 and GSE35487 datasets were used for the validation of key candidate genes. We evaluated five key candidate genes on the basis of area under the curve (AUC), computed from the receiver operating characteristic curve (ROC). For ROC analysis of each gene, the class label (IgAN vs. healthy control) and gene expression labels need to be collected. First, we used leave-one-out cross-validation and employed a logistic regression (LR) model to classify the subjects as either IgAN or healthy controls. After fitting the LR model, we computed AUC values using "pROC" R-package[29].

The ROC curve of five key candidate genes for the GSE116626 dataset was presented in Fig. 9A–E. In GSE116626, the AUC values of five key candidate genes were as follows: FOS (AUC: 0.997, 95% CI 0.989–1.000, Fig. 9A), JUN (AUC: 0.890, 95% CI 0.807–0.973, Fig. 9B), EGR1 (AUC: 0.929, 95% CI 0.859–0.998, Fig. 9C), FOSB (AUC: 0.959, 95% CI 0.910–1.000, Fig. 9D), DUSP1 (AUC: 0.937, 95% CI 0.875–0.999, Fig. 9E). The hierarchical clustering for each key candidate gene was shown in Fig. 9F.

Similarly, the ROC curve of five key candidate genes for the GSE35487 dataset was presented in Fig. 10A–E. We observed that the AUC values of five key candidate genes were greater than 0.900. The AUC values of these five key candidate genes were as follows: FOS (AUC: 0.993, 95% CI 0.975–1.000, Fig. 10A), JUN (AUC: 0.980, 95% CI 0.941–1.000, Fig. 10B), EGR1 (AUC: 0.967, 95% CI 0.900–1.000, Fig. 10C), FOSB (AUC: 1.000, 95% CI 0.980–1.000, Fig. 10D), DUSP1 (AUC: 0.967, 95% CI 0.900–1.000, Fig. 10E). The hierarchical clustering for each key candidate gene was shown in Fig. 10F. Finally, we recommended that the five key candidate genes (FOS, JUN, EGR1, FOSB, and DUSP1) may be considered as potential genes or key candidate genes for IgAN. Therefore, our findings were validated for both GSE116626, and GSE35487 datasets.

Discussion

In this study, we evaluated the GSE93798 dataset from GEO database to filter DEGs for IgAN patients and determine the key candidate genes. We identified 348 DEGs (up-regulated: 107 and down-regulated: 241) from GSE93798 that can be easily differentiated IgAN patients from healthy controls (Fig. 2A–B). To validate the pathogenetic process of DEGs, we did gene functional enrichment analysis of DEGs using DAVID. We considered the top five GO terms for BPs, MFs, and CCs, as well as KEGG pathways, that were statistically significantly associated with IgAN patients through DEGs. According to the GO analysis for BP, the DEGs were statistically significantly associated with inflammatory, camp, and cellular responses to lipopolysaccharide, cytokine-mediated signaling pathway, and neutrophil chemotaxis. Among them, some previous studies found response to

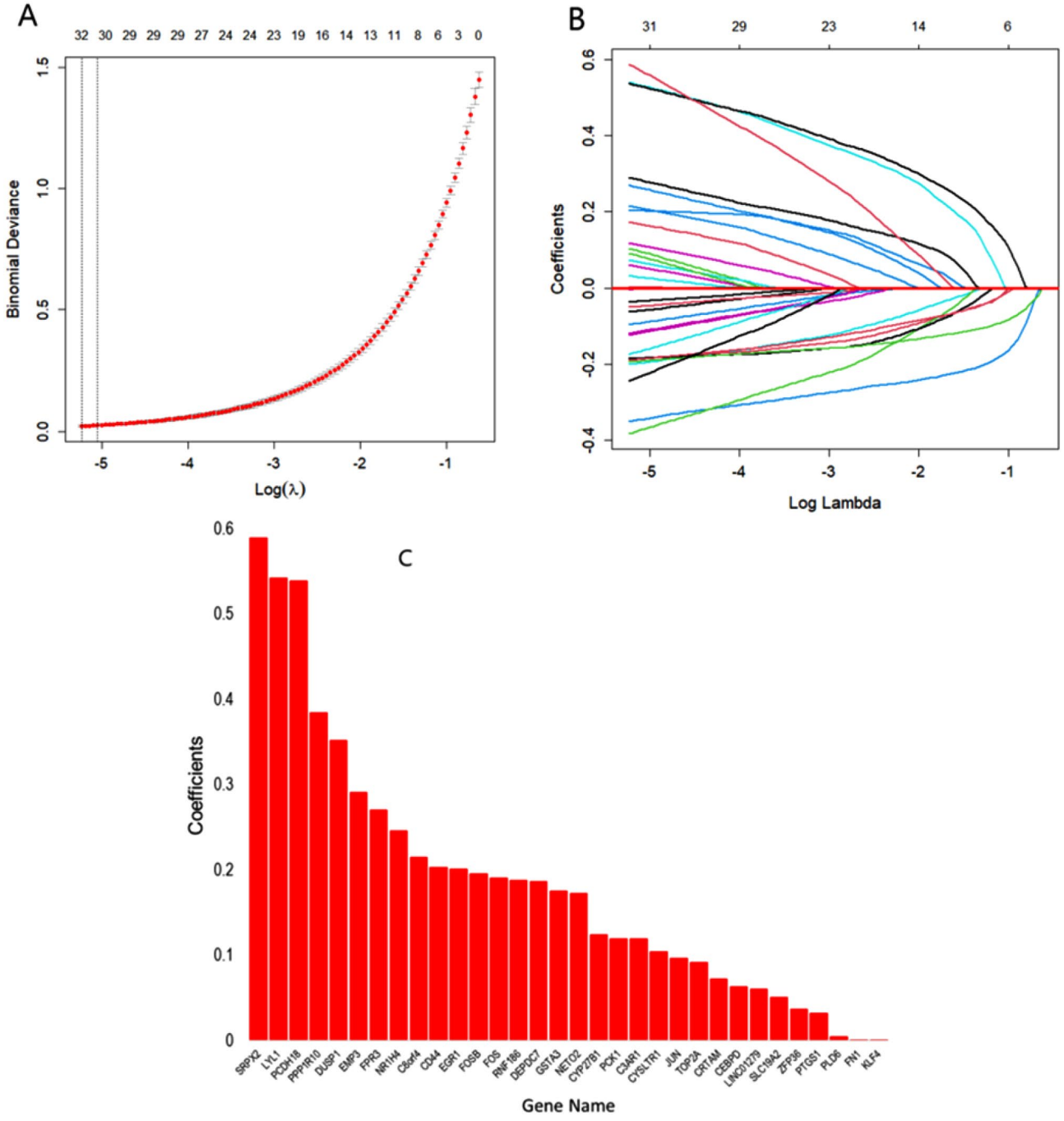

Figure 5. Discriminative gene selected using LASSO-based model by 10 CV: (**A**) A coefficient profile plot was generated against the log (λ) sequence. (**B**) 32 discriminative genes were selected for IgAN. (**C**) Contribution of 32 discriminative genes for IgAN patients.

inflammatory[10,21,30], response to camp[10,21,22,31], and cellular response to lipopolysaccharide[10,21] as highly significant GO terms.

In case of CCs, the top five GO terms were significantly associated with DEGs for IgAN patients, which we got in this study were consistent with previous studies such as extracellular exosome[10,21,30], extracellular region[10,21,30], extracellular space[10,21,30], collagen trimer[21,22], and blood microparticle[10,21,22]. For MFs, the three MFs supported by previous studies were extracellular matrix structural constituents[22], transcriptional activator activity, RNA polymerase II transcription regulatory region[10,20–22], and zinc ion binding[30]. In KEGG pathway analysis, our findings were closely related with previous studies. They showed that glycine, serine and threonine metabolism[10,22], age-rage signaling pathway in diabetic complications[22,32], protein digestion and absorption[21,22,30], IL-17 signaling pathway[22,32], and osteoclast differentiation[10,21,22,30,32] were significant pathways for DEGs.

The 348 DEGs were imported to STRING and visualized their PPI network with 206 nodes and 880 edges using Cytoscape. On the basis of degree of connectivity>18, we selected 19 hubs genes from the PPI network,

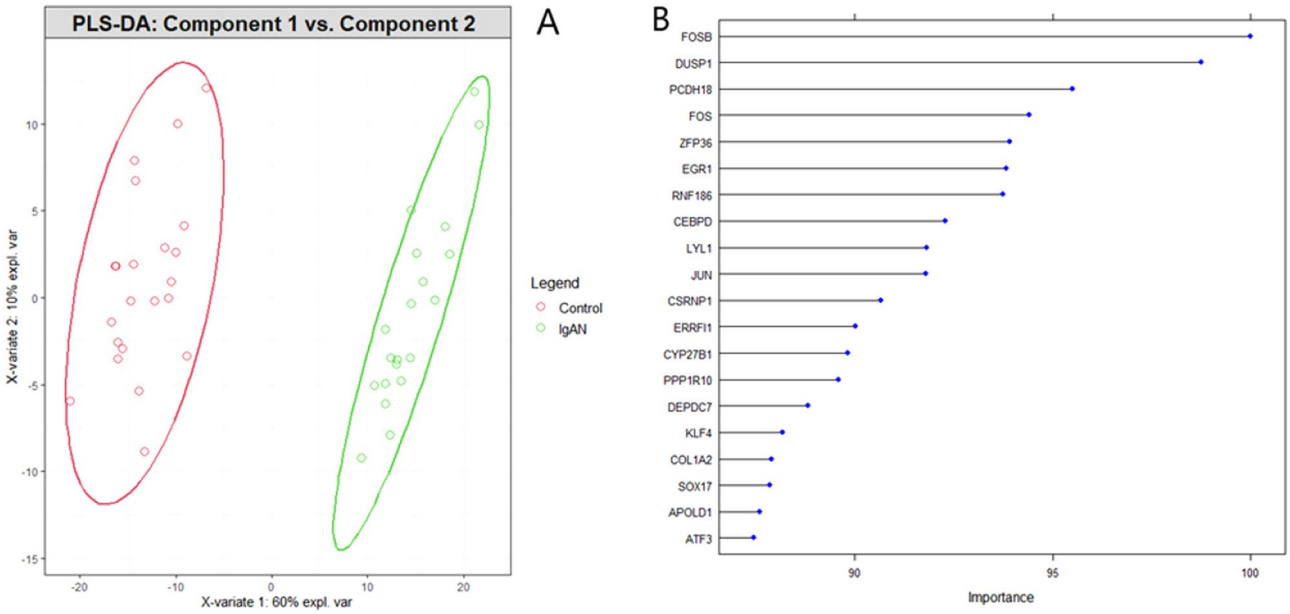

Figure 6. PLS-DA for DEGs: (**A**) Component 1 vs. Component 2. The red points indicate IgAN patients and the green points indicate healthy control; (**B**) Importance of top 20 discriminative genes for IgAN.

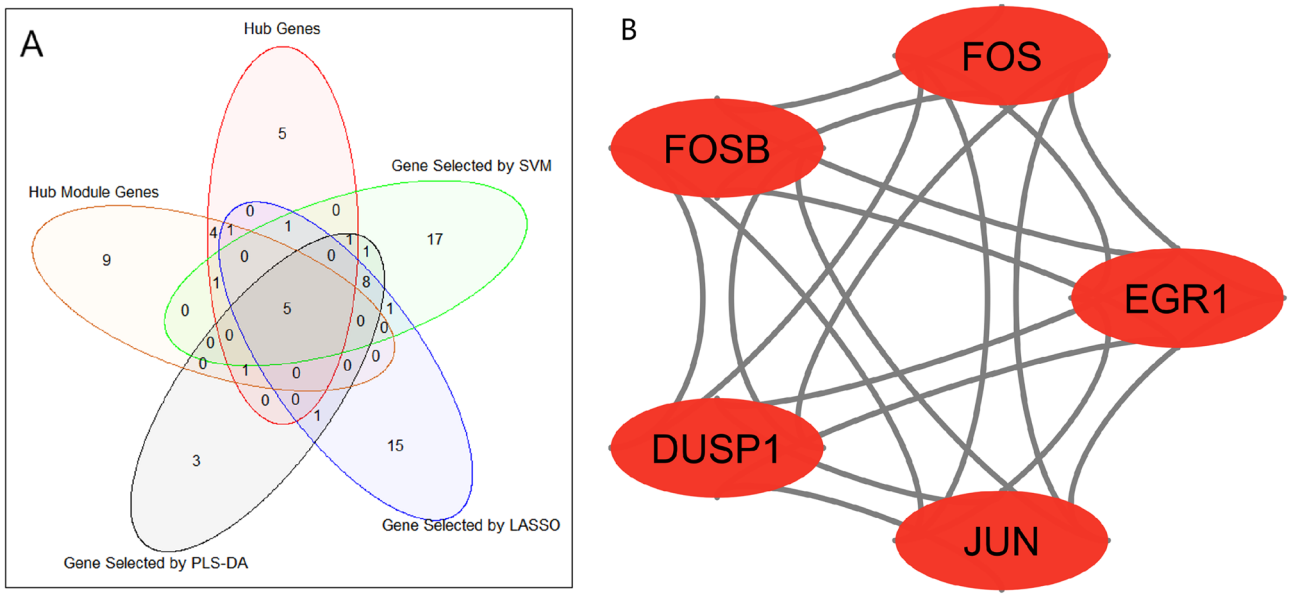

Figure 7. Identification and PPI analysis of key hub genes for IgAN patients. (**A**) Key candidate genes identification from hub module genes, computed from Cytohubba, SVM, LASSO, and PLS-DA. (**B**) PPI analysis of key five candidate genes.

which were showed in Table 3. Two significant modules were selected using MCODE with the cutoff points: MCODE scores $\geq$ 6 and number of nodes $\geq$ 6. The first module had 10 nodes and 11 nodes were in module 2, which were presented in Table 4 and their PPI network were also presented in Fig. 3B,C. Furthermore, we selected 21 hub module genes by taking the union of module 1 and module 2. To identify the discriminative genes, we applied three ML-based algorithms (SVM, LASSO, and PLS-DA) on 348 DEGs. We selected 35 discriminative genes using SVM (see in Fig. 4), 32 discriminative genes using LASSO (see in Fig. 5C, and 20 discriminative genes using PLS-DA (see in Fig. 6B). We identified five key candidate genes (FOS, JUN, EGR1, FOSB, and DUSP1) from the hub genes, hub module genes, and discriminative genes selected by SVM, LASSO, and PLS-DA (see Fig. 7A) and their PPI network were showed in Fig. 7B. We observed that each key candidate gene could be easily differentiated IgAN patients from healthy controls (Fig. 8A–E). The hierarchical clustering of the key candidate genes revealed that they were able to completely separate IgAN patients from healthy controls (Fig. 8F).

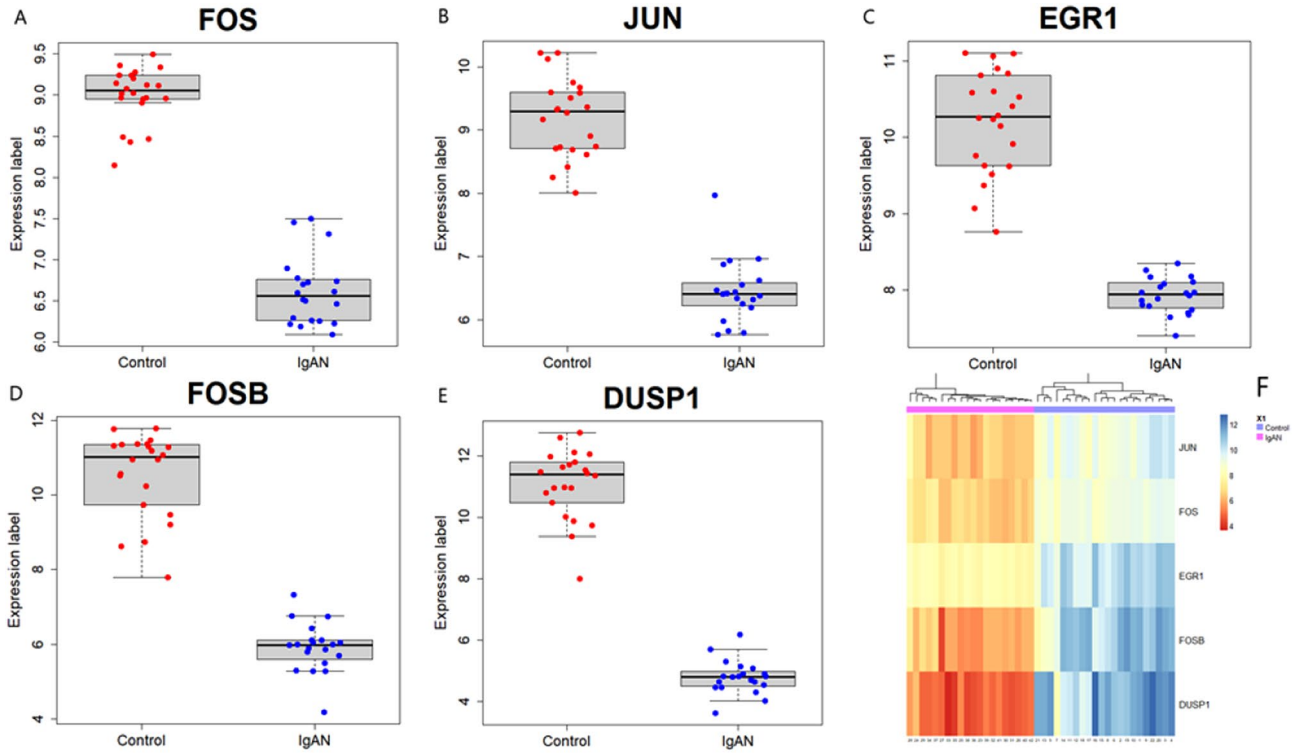

Figure 8. Boxplot of five key candidate genes as (**A**) FOS, (**B**) JUN, (**C**) EGR1, (**D**) FOSB, (**E**) DUSP1 for IgAN patients, and (**F**) Heatmap of the five key candidate genes in renal tissue samples which were generated using "NMF" version 0.24.0 package in R[64] (https://cran.r-project.org/package=NMF).

FOS is a component of activator protein 1 (AP-1) transcription factors[33] that controls the expression of genes involved in cell growth, death, inflammation, and differentiation[30,34,35]. FOS was significantly linked with DNA damage, telomere injury-related aging markers, and neutrophil activation, which also controlled IgAN initiation and evolution[36,37]. A study revealed that FOS was related to the disappearance of podocyte foot processes[38]. Our findings showed that FOS was strongly associated/correlated with IgAN, which was consistent with the previous studies[10,20,22,30,32,36,39]. JUN plays a crucial role in IgAN. It is also an AP-1 transcription factors and one of the most potential factors for IgAN. A study revealed that AP-1 was strongly associated with IgAN[15]. Our study also revealed that JUN was also a potential biomarker for IgAN, which was supported by the previous studies[10,22,30,32,36].

EGR1 is a zinc-finger protein that plays an important role in cell growth and proliferation. It promotes the epithelial-mesenchymal transition that contributes to diabetic kidney disease[39]. In rhabdomyosarcoma, EGR1 overexpression reduces cell proliferation, motility, and anchorage-independent growth[40]. In our study, EGR1 was one of the top five key biomarkers and significantly associated with IgAN, which was also supported by previous studies[10,20,30]. FOSB is one of the members of the FOS gene family and can be overexpressed in numerous diseases such as IgAN, mesangial proliferation, lupus nephropathy, and so on. Our study reported that FOSB was also a significant biomarker for IgAN. One of the DEGs was DUSP1, a gene linked to fibrosis[20]. DUSP1 is involved in both the human biological response to stress and the negative regulation of cell growth[41]. For hypertensive patients, angiotensin-1-7 increased DUSP1, which reduced fibrosis in resistant arterioles and end-stage organ damage[42]. Our study also reported that DUSP1 was a potential biomarker for IgAN, which was consistent with previous study[43].

In light of the above mentioned approach, we identified five key candidate genes (FOS, JUN, EGR1, FOSB, and DUSP1) that can easily be differentiated IgAN patients from healthy controls. Therefore, our study suggested that FOS, JUN, EGR1, FOSB, and DUSP1 may function as key biomarkers for the detection and diagnosis of IgAN. These five key candidate genes may play an important role in the development of IgAN and act as potential candidate molecular targets for the diagnosis and treatment of IgAN. This research will be helpful to the readers who will be interested in determining the correlated pathway of IgAN. However, more research into the processes of these genes in IgAN is required.

In the future, we will try to implement our proposed system for the identification of key candidate ncRNA for IgAN and compared our findings with previous studies[44–48]. Furthermore, we will adopt more ML-based and deep learning-based algorithms to identify the potential key candidate genes.

Methods

Microarray dataset. In this study, we used three publicly available GEO datasets with accession numbers: GSE93798[49], GSE116626[50] and GSE35487[51], which came from renal biopsies and one can easily be downloaded from the GEO database (www.ncbi.nlm.nih.gov/geo/). The GSE93798 dataset was used to determine the key

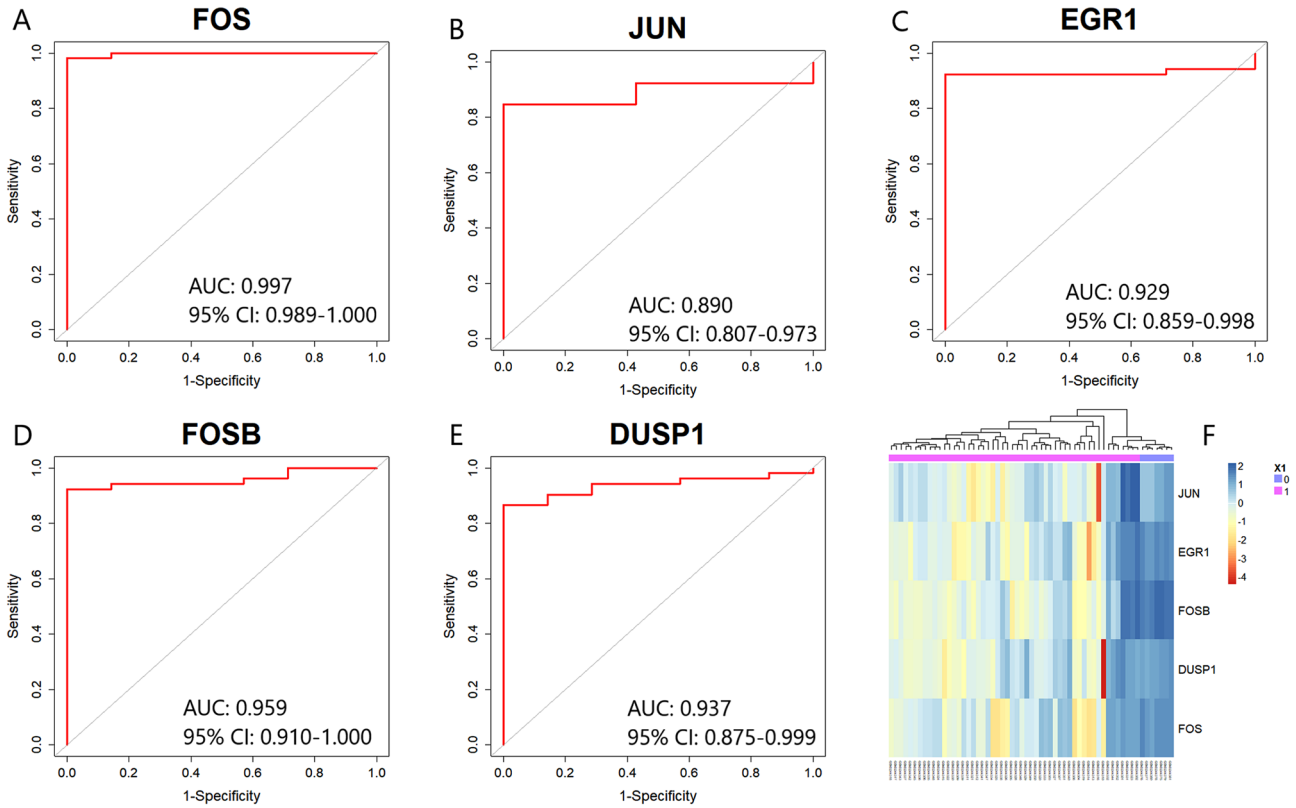

Figure 9. Validation of the five key candidate genes using ROC curves which were generated by pROC package with version 1.18.0 in R[29] (https://cran.r-project.org/package=pROC) and heatmap for GSE116626 dataset. (**A**) FOS (**B**) JUN (**C**), EGR1 (**D**) FOSB (**E**) DUSP1 (**F**) Heatmap of the five key candidate genes in renal tissue samples which were generated using "NMF" version 0.24.0 package in R[64] (https://cran.r-project.org/package=NMF). *CI* confidence interval.

candidate genes. The GSE93798 dataset was based on GPL22945 platform [*HG* − *U*133_*Plus*_2] and included 42 subjects, with 20 IgAN patients and 22 healthy controls. Another two independent datasets: GSE116626 and GSE35487 were used for the validation of key candidate genes. The GSE116626 dataset was based on GPL14951 platform and consisted of 52 IgAN patients and 7 healthy controls. On the other hand, the GSE35487 dataset was based on the GPL96 platform and composed of 25 IgAN patients and 6 healthy controls. Although these datasets were taken from the publicly available GEO repository, being the Human data, all methods were performed in accordance with the relevant guidelines and regulations.

Identification of DEGs. Using the platform GPL22945, the probe matrix was merged with our gene series matrix by Affymetrix ID and no genes were removed from our database. The DEGs between IgAN patients and healthy controls were identified using the limma package[52] in R software with version 4.1.2 (https://cran.r-project.org/). The DEGs were selected using the following cutoffs: adjusted probability value (p-value) < 0.05 and $|\log FC| > 1$. Where, FC is the fold change. The DEGs between IgAN and healthy control subjects were analyzed using hierarchical clustering.

Enrichment analysis of DEGs. The DEGs and top key candidate genes were both selected for GO and KEGG pathway analysis[53]. With these DEGs and top key candidate genes, GO term and KEGG enrichment analysis were obtained using DAVID version 6.8 tools (david.ncifcrf.gov) and a p-value < 0.05 was chosen as the cut-off criteria.

PPI network analysis and hub gene identification. We constructed an integrated network among selected DEGs. The STRING version 11.5 online based software (www.string-db.org) was used to make the network[21]. We set a confidence score to > 0.70 and a maximum number of interactors to 0 as a cutoff value to build the interaction of DEGs. Then, export the string interaction file and save it in TSV format. We visualized the PPI network on Cystoscape version 3.9.1[54]. To identify the hub genes, we set the degree of connectivity > 18 as a cutoff value.

Hub module and its gene identification. MCODE was used to visualize the significant nodes and also partition the network into different modules with degree cut-off = 2, cluster finding = haircut, node score cut-off = 0.2, K-score = 2, and maximum depth = 100, respectively. To select the most significant modules using

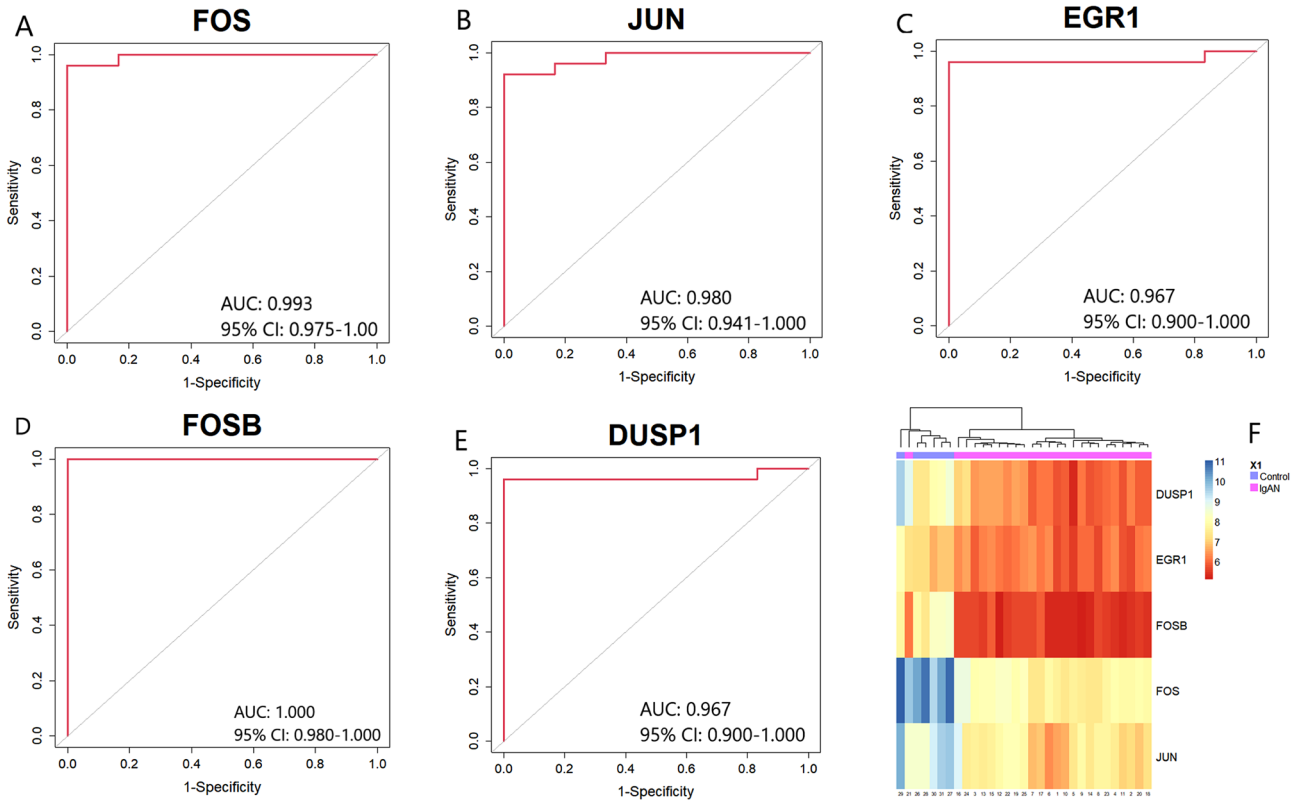

Figure 10. Validation of the five key candidate genes using ROC curves which were generated by pROC package with version 1.18.0 in R[29] (https://cran.r-project.org/package=pROC) and heatmap for GSE35487 dataset. (**A**) FOS (**B**) JUN (**C**), EGR1 (**D**) FOSB (**E**) DUSP1 (**F**) Heatmap of the five key candidate genes in renal tissue samples which were generated using "NMF" version 0.24.0 package in R[64] (https://cran.r-project.org/package=NMF).

MCODE, we set the cutoff values as follows: MCODE scores ≥ 6 and number of nodes ≥ 6, respectively. After selecting the significant module, we selected the hub module using the following formula:

$$\text{Hub Module Genes} = \bigcup_{i=1}^{m} \text{Genes from Module}_i \tag{1}$$

where, m is the number of significant modules. The corresponding genes were considered as hub module genes.

ML-based discriminative gene selection. After identifying DEGs, we have adopted three supervised ML algorithms as support vector machine (SVM), least absolute shrinkage and selection operator (LASSO), partial least squares discriminant analysis (PLS-DA) to identify the discriminative genes of IgAN. The brief descriptions of these algorithms are summarized as follows:

Support vector machine. SVM[55,56] is one of the most popular supervised ML algorithms. The aimed of SVM is to determine a hyperplane in a high dimensional space that can easily classified the groups as IgAN patients and healthy controls which needs to solve the following constraint problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2}$$

Subject to

$$\sum_{i=1}^{n} {y_i}^T \alpha_i = 1, 0 \leq \alpha_i \leq C, i = 1, \ldots, n \;\&\; \forall\, i = 1, 2, 3, \ldots, n \tag{3}$$

The final discriminate function takes the following form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) + b \tag{4}$$

where, b is the bias terms.

In this research, we have used radial basis kernel which is defined as follows:

$$K(x_i,\ x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{5}$$

There were some additional parameters in SVM with RBF kernel, such as cost (C) and gamma (γ), called hyper parameters. These hyperparameters were tuned using the grid search method and chose the hyperparameters that provided the highest classification accuracy. In this study, we used SVM as discriminative gene selection algorithm. We will identify the most discriminative genes from a set of DEGs for IgAN patients based on the following steps:

Step 1: Take 80% of the dataset for the training set and 20% of the dataset for the test set.
Step 2: Choose one gene from a list of 348 DEGs.
Step 3: Trained SVM model on the training dataset.
Step 4: Calculate the classification accuracy for this feature.
Step 5: Repeat Step 1 to **Step 4** into five times.
Step 6: Calculate the average of the classification accuracy.
Step 7: Repeat **Step 1** to **Step 6** for all (348) genes.
Step 8: Sort the classification accuracy from the largest to smallest.
Step 9: Select the genes that will produce more than 95.0% classification accuracy.

LASSO. LASSO is a supervised learning that is widely used both in biomarker selection and classification problems. We trained a logistic LASSO-based regression model on 348 DEGs to identify the discriminative genes of IgAN using the "*glmnet*" package in R with version 4.1.2[27,57]. To select the optimal parameters, we adopted a 10-fold cross-validation protocol, and the partial likelihood deviance met the minimum criteria. The genes with non-zero coefficients of the LASSO-based logistic regression model are selected as discriminative genes, and we remove the genes with zero coefficients of the LASSO-based model from our analysis.

PLS-DA. PLS-DA is one of the most popular supervised ML algorithms. It is widely used not only in dimension reduction algorithms such as PCA, but also in gene selection[58–60] and classification[61,62]. We utilized PLS-DA while the response variable takes a categorical variable, for example, "1" for yes and "0" for no. It is similar to logistic regression. In this study, we used PLS-DA as a gene selection method to identify the discriminative genes for IgAN patients using the *"mixOmics"* package in R.

Key candidate genes identification. To identify the key candidate genes and avoid the missing the important genes, we identified the discriminative genes using three ML-based methods (SVM, LASSO, and PLS-DA), the hub genes using the degree of connectivity from PPI network, and hub module genes from significant modules. We identified the key candidate genes using the following formula:

$$\text{Key Candidate Genes} = \bigcap_{i=1}^{k} \text{Identification Methods}_i \tag{6}$$

where, k is the number of potential gene identification methods.

Data availability

The datasets generated and/or analysed during the current study are available in the Gene Expression Omnibus (GEO) repository with accession numbers: GSE93798, GSE116626 and GSE35487. Using these accession numbers, one can easily download these datasets from the following link: www.ncbi.nlm.nih.gov/geo/.

Received: 17 May 2022; Accepted: 8 August 2022
Published online: 17 August 2022

References

1. Boully, A. *et al.* A brain problem with listeria monocytogenes. *Lancet. Infect. Dis.* **22**, 296. https://doi.org/10.1016/S1473-3099(21)00683-6 (2022).
2. Berger, J. & N, H. Les depots intercapillaires d'iga-igg. *J. Urol. Nephrol.* **74**, 694–695 (1968).
3. D'amico, G. The commonest glomerulonephritis in the world: Iga nephropathy. *Q. J. Med.* **64**, 709–727. https://doi.org/10.1093/oxfordjournals.qjmed.a068143 (1987).
4. Lai, K. N. *et al.* Iga nephropathy. *Nat. Rev. Dis. Primers.* **2**, 1–20. https://doi.org/10.1038/nrdp.2016.1 (2016).
5. Jarrick, S. *et al.* Immunoglobulin a nephropathy and ischemic heart disease: A nationwide population-based cohort study. *BMC Nephrol.* **22**, 1–8. https://doi.org/10.1186/s12882-021-02353-7 (2021).
6. Mustonen, J. & Pasternack, A. Associated diseases in iga nephropathy. In *IgA nephropathy*, 47–65, https://doi.org/10.1007/978-1-4613-2039-5_5 (Springer, 1987).
7. Kalambokis, G., Christou, L., Stefanou, D., Arkoumani, E. & Tsianos, E. V. Association of liver cirrhosis related iga nephropathy with portal hypertension. *World J. Gastroenterol.* **13**, 5783–5786. https://doi.org/10.3748/wjg.v13.i43.5783 (2007).
8. Habura, I. *et al.* Iga nephropathy associated with coeliac disease. *Cent. Eur. J. Immunol.* **44**, 106–108. https://doi.org/10.5114/ceji.2019.84021 (2019).
9. Wyatt, R. J. & Julian, B. A. Iga nephropathy. *N. Engl. J. Med.* **368**, 2402–2414. https://doi.org/10.1056/nejmra1206793 (2013).
10. Qian, W., Xiaoyi, W. & Zi, Y. Screening and bioinformatics analysis of iga nephropathy gene based on geo databases. *Biomed. Res. Int.* **1–7**, 2019. https://doi.org/10.1155/2019/8794013 (2019).

11. Fellström, B. C. *et al.* Targeted-release budesonide versus placebo in patients with iga nephropathy (nefigan): A double-blind, randomised, placebo-controlled phase 2b trial. *Lancet* **389**, 2117–2127. https://doi.org/10.1016/S0140-6736(17)30550-0 (2017).
12. Konieczny, A. *et al.* Clinical and histopathological factors influencing iga nephropathy outcome. *Diagnostics* **11**, 1764. https://doi.org/10.3390/diagnostics11101764 (2021).
13. Woo, K.-T. *et al.* Global evolutionary trend of the prevalence of primary glomerulonephritis over the past three decades. *Nephron Clin. Pract.* **116**, c337–c346. https://doi.org/10.1159/000319594 (2010).
14. Schena, F. P. & Nistor, I. Epidemiology of iga nephropathy: A global perspective. *In Semin. Nephrol.* **38**, 435–442. https://doi.org/10.1016/j.semnephrol.2018.05.013 (2018) (**Elsevier**).
15. Wang, J. & Cao, J. Gene expression analysis in tubule interstitial compartments reveals candidate agents for iga nephropathy. *Kidney Blood Press. Res.* **39**, 361–368. https://doi.org/10.1159/000355814 (2014).
16. Jarrick, S. *et al.* Mortality in iga nephropathy: A nationwide population-based cohort study. *J. Am. Soc. Nephrol.* **30**, 866–876. https://doi.org/10.1681/ASN.2018101017 (2019).
17. Rahman, M. *et al.* Identification of potential long non-coding rna candidates that contribute to triple-negative breast cancer in humans through computational approach. *Int. J. Mol. Sci.* **22**, 12359–12373. https://doi.org/10.3390/ijms222212359 (2021).
18. Hossain, M. T. *et al.* Identification of circrna biomarker for gastric cancer through integrated analysis. *Front. Mol. Biosci.* **9**, 1–13. https://doi.org/10.3389/fmolb.2022.857320 (2022).
19. Reza, M. S. *et al.* Bioinformatics screening of potential biomarkers from mrna expression profiles to discover drug targets and agents for cervical cancer. *Int. J. Mol. Sci.* **23**, 3968–3989. https://doi.org/10.3390/ijms23073968 (2022).
20. Zhang, D. *et al.* Integrated bioinformatics analysis reveals novel hub genes closely associated with pathological mechanisms of immunoglobulin a nephropathy. *Exp. Ther. Med.* **18**, 1235–1245. https://doi.org/10.3892/etm.2019.7686 (2019).
21. Jiang, X., Xu, Z., Du, Y. & Chen, H. Bioinformatics analysis reveals novel hub gene pathways associated with iga nephropathy. *Eur. J. Med. Res.* **25**, 1–11. https://doi.org/10.1186/s40001-020-00441-2 (2020).
22. Chen, X. & Sun, M. Identification of key genes, pathways and potential therapeutic agents for iga nephropathy using an integrated bioinformatics analysis. *J. Renin Angiotensin Aldosterone Syst.* **21**, 1–9. https://doi.org/10.1177/1470320320919635 (2020).
23. Tan, K. *et al.* Genome-wide analysis of micrornas expression profiling in patients with primary iga nephropathy. *Genome* **56**, 161–169. https://doi.org/10.1139/gen-2012-0159 (2013).
24. Wei, S.-Y., Guo, S., Feng, B., Ning, S.-W. & Du, X.-Y. Identification of mirna-mrna network and immune-related gene signatures in iga nephropathy by integrated bioinformatics analysis. *BMC Nephrol.* **22**, 1–15. https://doi.org/10.1186/s12882-021-02606-5 (2021).
25. Wang, W. *et al.* The key candidate genes in tubulointerstitial injury of chronic kidney diseases patients as determined by bioinformatic analysis. *Cell Biochem. Funct.* **38**, 761–772. https://doi.org/10.1002/cbf.3545 (2020).
26. Qing, J.-B., Song, W.-Z., Li, C.-Q. & Li, Y.-F. The diagnostic and predictive significance of immune-related genes and immune characteristics in the occurrence and progression of iga nephropathy. *J. Immunol. Res.* **1–20**, 2022. https://doi.org/10.1155/2022/9284204 (2022).
27. Yu, S.-H. *et al.* Lasso and bioinformatics analysis in the identification of key genes for prognostic genes of gynecologic cancer. *J. Pers. Med.* **11**, 1177. https://doi.org/10.3390/jpm11111177 (2021).
28. Basith, S., Hasan, M. M., Lee, G., Wei, L. & Manavalan, B. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. *Brief. Bioinform.* **22**, bbab252. https://doi.org/10.1093/bib/bbab252 (2021).
29. Robin, X. *et al.* proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics* **12**, 1–8. https://doi.org/10.1186/1471-2105-12-77 (2011).
30. Noor, F., Saleem, M. H., Aslam, M. F., Ahmad, A. & Aslam, S. Construction of mirna-mrna network for the identification of key biological markers and their associated pathways in iga nephropathy by employing the integrated bioinformatics analysis. *Saudi J. Biol. Sci.* **28**, 4938–4945. https://doi.org/10.1016/j.sjbs.2021.06.079 (2021).
31. Bai, Y., Li, Y., Xi, Y. & Ma, C. Identification and validation of glomerulotubular crosstalk genes mediating iga nephropathy by integrated bioinformatics. *BMC Nephrol.* **23**, 1–11. https://doi.org/10.1186/s12882-022-02779-7 (2022).
32. Zhou, X., Wang, N., Zhang, Y. & Yu, P. Expression of ccl2, fos, and jun may help to distinguish patients with iga nephropathy from healthy controls. *Front. Physiol.* **13**, 840890. https://doi.org/10.3389/fphys.2022.840890 (2022).
33. Zenz, R. *et al.* Activator protein 1 (fos/jun) functions in inflammatory bone and skin disease. *Arthritis Res. Ther.* **10**, 1–10. https://doi.org/10.1186/ar2338 (2008).
34. Durchdewald, M., Angel, P. & Hess, J. The transcription factor fos, a janus-type regulator in health and disease. *Histol. Histopathol.*https://doi.org/10.14670/hh-24.1451 *(2009).*
35. Hess, J., Angel, P. & Schorpp-Kistner, M. Ap-1 subunits: quarrel and harmony among siblings. *J. Cell Sci.* **117**, 5965–5973. https://doi.org/10.1242/jcs.01589 (2004).
36. Hu, S.-L. *et al.* Identification of key genes and pathways in iga nephropathy using bioinformatics analysis. *Medicine***99**, 1–6. https://doi.org/10.1097/FMD.0000000000021372 (2020).
37. Jiang, H. *et al.* Functional networks of aging markers in the glomeruli of iga nephropathy: A new therapeutic opportunity. *Oncotarget***7**, 33616–33626. https://doi.org/10.18632/oncotarget.9033 (2016).
38. Park, H. J., Kim, J. W., Cho, B.-S. & Chung, J.-H. Association of fos-like antigen 1 promoter polymorphism with podocyte foot process effacement in immunoglobulin a nephropathy patients. *J. Clin. Lab. Anal.* **28**, 391–397. https://doi.org/10.1002/jcla.21699 (2014).
39. Hu, F. *et al.* Early growth response 1 (egr1) is a transcriptional activator of nox4 in oxidative stress of diabetic kidney disease. *J. Diabetes Res.* **1–10**, 2018. https://doi.org/10.1155/2018/3405695 (2018).
40. Mohamad, T., Kazim, N., Adhikari, A. & Davie, J. K. Egr1 interacts with tbx2 and functions as a tumor suppressor in rhabdomyosarcoma. *Oncotarget***9**, 18084–18098. https://doi.org/10.18632/oncotarget.24726 (2018).
41. Jianping, W. *et al.* Pos-374 identifying dusp-1 and fosb as hub genes in immunoglobulin a nephropathy by wgcna and degs screening and validation. *Kidney Int. Rep.* **7**, S169. https://doi.org/10.1016/j.ekir.2022.01.396 (2022).
42. Carver, K. A., Smith, T. L., Gallagher, P. E. & Tallant, E. A. Angiotensin-(1–7) prevents angiotensin ii-induced fibrosis in cremaster microvessels. *Microcirculation* **22**, 19–27. https://doi.org/10.1111/micc.12159 (2015).
43. Hammer, M. *et al.* Dual specificity phosphatase 1 (dusp1) regulates a subset of lps-induced genes and protects mice from lethal endotoxin shock. *J. Exp. Med.* **203**, 15–20. https://doi.org/10.1084/jem.20051753 (2006).
44. Chen, X., Yan, C. C., Zhang, X. & You, Z.-H. Long non-coding rnas and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **18**, 558–576. https://doi.org/10.1093/bib/bbw060 (2017).
45. Wang, C.-C., Han, C.-D., Zhao, Q. & Chen, X. Circular rnas and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **22**, bbab286. https://doi.org/10.1093/bib/bbab286 (2021).
46. Chen, X., Xie, D., Zhao, Q. & You, Z.-H. Micrornas and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **20**, 515–539. https://doi.org/10.1093/bib/bbx130 (2019).
47. Liu, W. *et al.* Nscgrn: A network structure control method for gene regulatory network inference. *Brief. Bioinform.*https://doi.org/10.1007/s12539-021-00478-9 *(2022).*
48. Liu, W. *et al.* Inferring gene regulatory networks using the improved markov blanket discovery algorithm. *Interdiscip. Sci.* **14**, 168–181. https://doi.org/10.1007/s12539-021-00478-9 (2022).

49. Liu, P. *et al.* Transcriptomic and proteomic profiling provides insight into mesangial cell function in iga nephropathy. *J. Am. Soc. Nephrol.* **28**, 2961–2972. https://doi.org/10.1681/ASN.2016101103 (2017).
50. Cox, S. N. *et al.* Formalin-fixed paraffin-embedded renal biopsy tissues: An underexploited biospecimen resource for gene expression profiling in iga nephropathy. *Sci. Rep.* **10**, 1–14. https://doi.org/10.1038/s41598-020-72026-2 (2020).
51. Reich, H. N. *et al.* A molecular signature of proteinuria in glomerulonephritis. *PLoS ONE* **5**, e13451–e13462. https://doi.org/10.1371/journal.pone.0013451 (2010).
52. Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* **43**, 1–13. https://doi.org/10.1093/nar/gkv007 (2015).
53. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. Kegg: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551. https://doi.org/10.1093/nar/gkaa970 (2021).
54. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504. https://doi.org/10.1101/gr.1239303 (2003).
55. Hasan, M. A. M., Nasser, M., Pal, B. & Ahmad, S. Support vector machine and random forest modeling for intrusion detection system (ids). *J. Intell. Learn. Syst. Appl.* **2014**. https://doi.org/10.4236/jilsa.2014.61005 (2014).
56. Jan, S. U., Lee, Y.-D., Shin, J. & Koo, I. Sensor fault classification based on support vector machine and statistical time-domain features. *IEEE Access* **5**, 8682–8690. https://doi.org/10.1109/ACCESS.2017.2705644 (2017).
57. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **58**, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x (1996).
58. Gutkin, M., Shamir, R. & Dror, G. Slimpls: A method for feature selection in gene expression-based disease classification. *PLoS ONE* **4**, e6416. https://doi.org/10.1371/journal.pone.0006416 (2009).
59. Christin, C. *et al.* A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Mol. Cell. Proteomics* **12**, 263–276. https://doi.org/10.1074/mcp.M112.022566 (2013).
60. Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K. & Narasimhan, G. So you think you can pls-da?. *BMC Bioinf.* **21**, 1–10. https://doi.org/10.1186/s12859-019-3310-7 (2020).
61. Lee, L. C., Liong, C.-Y. & Jemain, A. A. Partial least squares-discriminant analysis (pls-da) for classification of high-dimensional (hd) data: A review of contemporary practice strategies and knowledge gaps. *Analyst* **143**, 3526–3539. https://doi.org/10.1039/C8AN00599K (2018).
62. Gold, K. M., Townsend, P. A., Herrmann, I. & Gevens, A. J. Investigating potato late blight physiological differences across potato cultivars with spectroscopy and machine learning. *Plant Sci.* **295**, 110316. https://doi.org/10.1016/j.plantsci.2019.110316 (2020).
63. Wickham, H. et al. ggplot2: Create elegant data visualisations using the grammar of graphics (3.3. 6)[computer software], https://cran.r-project.org/package=ggplot2 (2022).
64. Gaujoux, R. & Seoighe, C. Nmf: Algorithms and framework for nonnegative matrix factorization (nmf). R Package Version 0.20 **6**, http://CRAN.R-project.org/package=NMF (2015).

Author contributions

All listed authors participated meaningfully in the study, and they have seen and approved the submission of this manuscript. Conceptualization, M.A.MH.; Methodology, M.A.M.H, M. M.; Data collection and curation: M.A.M.H, M.M, J.S.; Interpreted and analyzed the data, M.A.M.H, M.M, J.S.; Writing—original draft preparation, M.A.M.H, M.M.; Writing—review and editing, M.A.M.H, M.M, J.S.; Supervision, J.S., M.A.M.H.; Project administration and funding, J.S.

Funding

This work was supported by the Graduate School Research Fund of The University of Aizu, Japan.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.S.

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

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