Zeng Xiaoxi (Orcid ID: 0000-0001-9946-4291)

A network-based variable selection approach for identification of modules and biomarker genes associated with end-stage kidney disease

Xiaoxi Zeng[1,2,3#], Chunyang Li[1,3#], Yi Li[4], Haopeng Yu[1,3], Ping Fu[2,3*], Hyokyoung G. Hong[5*], Wei Zhang[1,3]

1. West China Biomedical Big Data Center, West China School of Medicine (West China Hospital), Sichuan University, Chengdu, China
2. Division of Nephrology, Kidney Research Institute, West China Hospital, Sichuan University, Chengdu, China
3. Medical Big Data Center, Sichuan University, Chengdu, China
4. Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, USA
5. Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA

# These authors contributed equally to this work.

Running title: Identification of biomarker genes for ESKD

***Corresponding Authors**

Ping Fu
Division of Nephrology, Kidney Research Institute, West China Hospital
Sichuan University
37 Guo Xue Xiang
Chengdu, 610041, China
Tel: 86-28-85422335
E-mail: fupinghx@scu.edu.cn

Hyokyoung G. Hong
Department of Statistics and Probability
Michigan State University
619 Red Cedar
East Lansing, MI 48824, USA
Tel: 1-517-432-1485
Fax: 1-517-432-1405
E-mail: hhong@msu.edu

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/nep.13655

A network-based variable selection approach for identification of modules and biomarker genes associated with end-stage kidney disease

Xiaoxi Zeng[1,2,3#], Chunyang Li[1,3#], Yi Li[4], Haopeng Yu[1,3], Ping Fu[2,3*], Hyokyoung G. Hong[5*], Wei Zhang[1,3]

1. West China Biomedical Big Data Center, West China School of Medicine (West China Hospital), Sichuan University, Chengdu, China
2. Division of Nephrology, Kidney Research Institute, West China Hospital, Sichuan University, Chengdu, China
3. Medical Big Data Center, Sichuan University, Chengdu, China
4. Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, USA
5. Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA

# These authors contributed equally to this work.

Running title: Identification of biomarker genes for ESKD

***Corresponding Authors**

Ping Fu
Division of Nephrology, Kidney Research Institute, West China Hospital
Sichuan University
37 Guo Xue Xiang
Chengdu, 610041, China
Tel: 86-28-85422335
E-mail: fupinghx@scu.edu.cn

Hyokyoung G. Hong
Department of Statistics and Probability
Michigan State University
619 Red Cedar
East Lansing, MI 48824, USA
Tel: 1-517-432-1485
Fax: 1-517-432-1405

E-mail: hhong@msu.edu

ABSTRACTS:

Aims: Intervention for end-stage kidney disease (ESKD), which is associated with adverse prognoses and major economic burdens, is challenging due to its complex pathogenesis. The study was performed to identify biomarker genes and molecular mechanisms for ESKD by bioinformatics approach.

Methods: Using the Gene Expression Omnibus (GEO) dataset GSE37171, this study identified pathways and genomic biomarkers associated with ESKD via a multi-stage knowledge discovery process, including identification of modules of genes by weighted gene co-expression network analysis (WGCNA), discovery of important involved pathways by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses, selection of differentially expressed genes (DEGs) by the empirical Bayes method, and screening biomarker genes by the least absolute shrinkage and selection operator (Lasso) logistic regression. The results were validated using GSE70528, an independent testing dataset.

Results: Three clinically important gene modules associated with ESKD, were identified by WGCNA. Within these modules, GO and KEGG enrichment analyses revealed important biological pathways involved in ESKD, including TGF-β and Wnt

signaling, RNA-splicing, autophagy, and chromatin and histone modification. Furthermore, Lasso logistic regression was conducted to identify five final genes, namely, CNOT8, MST4, PPP2CB, PCSK7 and RBBP4, that are differentially expressed and associated with ESKD. The accuracy of the final model in distinguishing the ESKD cases and controls was 96.8% and 91.7% in the training and validation datasets, respectively.

Conclusions: Network-based variable selection approaches can identify biological pathways and biomarker genes associated with ESKD. The findings may inform more in-depth follow-up research and effective therapy.

Key words: End-stage kidney disease; Computational biology; Machine learning; Genetic transcription; Biomarkers

INTRODUCTION

Chronic kidney disease (CKD), with a prevalence of 10-15% worldwide [1] and the projected 5th leading cause of death by 2040 [2], has emerged as a major epidemic. CKD may eventually progress to end-stage kidney disease (ESKD), which is associated with markedly increased mortality and adverse complications, such as cardiovascular events, anemia, bone mineral disorders and frequent hospitalizations [1, 3]. In addition, ESKD consumed 7.1% of the overall Medicare claims in US [4] and 2–3% of the total health care expenditure in other developed countries [5], and has posed considerable economic burden on developing countries [3]. The etiology for the progression of CKD to ESKD is multifactorial, involving pathophysiological pathways of fibrosis, inflammation, oxidative stress, and mitochondrial damage, among others [1, 6, 7].

Despite the availability of renal replacement therapy, it remains challenging to develop effective therapeutic interventions due to the complex pathogenesis of ESKD. Understanding the mechanisms that govern the progression from CKD to ESKD is key in disease intervention, and identification of reliable molecular mechanisms for the CKD progression remains an enduring theme in renal research. Advances in molecular biology, genomics, and computational statistics have accelerated the work of finding novel disease-related genomic and molecular factors that will shed light on effective treatments. Since co-expression patterns of genes provide useful information of the underlying cellular processes, we conducted weighted gene co-expression network analysis (WGCNA) to identify highly correlated gene expression modules, which are related to a number of physical, behavioral, and disease traits [8-10]. In WGCNA, the connection of each pair of genes is weighted by a "soft" threshold, yielding more robust results than unweighted networks [8]. In addition, variable selection approaches may detect meaningful phenotype-genotype relationships [11]. Therefore, the combined use of WGCNA and variable selection may help detect novel genes associated with a number of diseases [12-15] and construct predictive models.

Given the fact that kidney biopsy in ESKD patients was generally not safe, this paper seeks to discover featured biomarkers in ESKD patients, gene expression profiles of peripheral blood cells (PBCs) from a previously published cohort of ESKD patients [16], and normal controls from the Gene Expression Omnibus (GEO) via WGCNA and the least absolute shrinkage and selection operator (Lasso) logistic regression. The findings might provide new knowledge of the pathophysiological alterations of ESKD on the molecular level, and suggest therapeutic targets for future research and clinical intervention.

MATERIALS AND METHODS

Microarray data selection

Microarray data were retrieved from the Gene Expression Omnibus (GEO) database of the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/geo/). Dataset GSE37171 was used as the training set

[16]. The original study was of case-control design, enrolling 75 ESKD patients and 20 normal controls. Dataset GSE70528 [17], with 4 ESKD patients and 8 controls, was used as the validating set. The gene expression profiles of both datasets were analyzed on Affymetrix Human Genome U133 Plus 2.0 Array (GPL570 [HG-U133_Plus_2]).

Microarray data preprocessing

The diagram of the study is shown in Figure 1. We conducted background adjustment, quantile normalization, log-transformation and summarization of the raw data by using the GC-robust multi-array analysis (GC-RMA), which uses probe sequence information to estimate probe affinity to non-specific binding [18, 19]. As only 40% of genes were expressed in most tissues, filtering by variance was performed with a cutoff at 50%, the default value in GC-RMA [20]. The probes were then mapped to gene symbols according to the annotation files. For genes with multiple matching probes, the mean values of the probes were used as the genes' expression levels. The "ComBat" function in the SVA package was used to remove the batch effects between the two datasets [21].

WGCNA

In WGCNA, each gene pair is assigned a connection weight via "soft" thresholding. Specifically, the pairwise correlation between two co-expressed genes, say, s_{ij}, is transformed to be a connection weight, a_{ij}, through a power function: $a_{ij} = |s_{ij}|^{\beta}$. Here, $\beta > 0$ is the soft thresholding parameter and is chosen to ensure a good scale-free topology fit (eg. Index R^2 larger than 0.80) and a large number of connections [8]. We implemented "blockwiseModules" function in WGCNA package for network construction and module detection in a group-wise manner. Firstly, genes were grouped by using K-means clustering. For each group of genes, the topological overlap was calculated and genes were further clustered by average linkage hierarchical clustering. The minimum module size was set to include at least 30 genes [22, 23]. The connectivity (or degree) of genes in the modules were obtained by the

"fundamentalNetworkConcepts" function [22, 23].

Identification of clinically important modules and enrichment analysis

Clinically important modules were identified according to the correlations between clinical traits and module eigengenes (MEs), which were considered as the first principal component for each gene module [23]. Additionally, the module significance (MS) was measured by the average gene significance (GS), defined as the average of the negative log p-values for individual genes' correlations within each module [23].

The "clusterProfiler" package was used to conduct Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for genes in clinically important modules based on the hypergeometric distribution [24]. The cutoff criteria were p-value <0.01 and q- <0.2.

Biomarker genes selection

Candidate genes were pre-selected according to the following criteria: (1) DEGs between ESKD and the controls screened by empirical Bayes methods using the "limma" package [25] have adjusted p-value <0.05 and $|\log_2 FC(\text{fold change})|>0.585$ (corresponding to $|FC|>1.5$)); (2) High module membership (MM), defined as the correlation of gene expression with MEs [23], e.g. $|MM|\geq 0.8$, and significant correlation with ESKD (Pearson's correlation $|r|\geq 0.4$); (3) Hub genes with the highest 1% connectivity (degree) in the modules with clinical significance (or 2% if the number of genes are much smaller than the others); (4) Genes with node degree ≥2 in the protein-protein interactions (PPI) network calculated by CytoHubba plugin in Cytoscape [26]. The original PPI was constructed by the Search Tool for the Retrieval of Interacting Genes (STRING) database and visualized in Cytoscape.

The "glmnet" package, which performs the Lasso logistic regression, was used to conduct variable selection on the training set. The tuning parameters in the models were chosen via 5-fold cross-validation [27]. The predictive performance of the selected

biomarkers was further validated using the testing dataset GSE70528. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated in both training and testing datasets. To assess the overall differentiation accuracy, the Area under the Receiver-operating characteristic (ROC) curve (AUC) with 95% confidence interval (CI) was obtained by using the "pROC" package [28]. All statistical tests were performed with R (version 3.5.1).

Results:

WGCNA

After data preprocessing, the expressions of 12,007 genes were obtained from 75 ESKD patients and 20 controls in the training GSE37171 dataset. The soft threshold was used in the adjacency function, wherein β is set at 11 to ensure the scale-free topology fitting (R^2 =0.857) with an acceptable number of connections (mean k=370) (Figure 2). WGCNA identified a total of 10 modules; see Figure 3 A. The turquoise module contained the most genes (n=5,441), followed by the blue (n=1009), brown (n=939), yellow (n=453), green (398), red (n=158), black (n=66), pink (n=52), and magenta (n=44) modules. Within these mentioned-above modules, expression profiles of genes were highly correlated [8]. Although the gray module contained 3,447 genes, it featured a heterogeneous expression pattern that could not be clustered to any specific modules. Hence, it was excluded from further analysis. The measure of over-expression (eg. module eigengene) was also obtained. Figure 3 B and Figure 3 C demonstrate the topological overlap matrix based on 1,000 randomly selected genes, and the adjacency of eigengenes, the first principal component in each module, with red being highly related and blue being not related in the heatmap plot.

Identification of gene modules with clinical significance

The Pearson correlations between the module eigengenes (MEs) and clinical traits were computed. The brown, turquoise, and blue modules had absolute r-values greater than 0.4 and p-values < 0.05 with ESKD (Figure 4 A). These three modules also had the highest MS (Figure 4 B), which denoted association of the genes in the

modules with ESKD. In addition, Figure 4 C demonstrates that genes with higher MM (correlation with modules' eigengenes) were likely to have higher GS (correlation with ESKD) as shown in the blue, brown and turquoise modules. Thus, we considered the brown, turquoise and blue modules for further analysis.

Functional enrichment analysis of clinically significant modules

A total of 939 genes in the brown module, 5,441 genes in the turquoise module, and 1,009 genes in the blue module were included in the GO and KEGG pathway enrichment analyses. Figure 5 presents the first 5 enriched results for each module. The blue module was associated RNA complex biogenesis, and covalent chromatin and histone modification, the brown module was associated with the biological function of erythrocyte systems, and the turquoise module was associated with RNA splicing and processing. KEGG analysis has further showed that these modules were enriched in several pathways, such as ubiquitin mediated proteolysis, autophagy, spliceosome, endocrine resistance and mitophagy.

Selection of biomarkers

A total of 2,975 DEGs in the training set were identified using the empirical Bayes method (Figure 4 D). Applying the criteria of the connectivity and the correlation with ESKD and MM had pre-selected 64 candidate genes (Figure S1), whose PPI network is shown in Figure (Figure 6). These genes were enriched in several GO-BP terms (eg. proteasome-mediated ubiquitin-dependent protein catabolic process and histone methylation) and KEGG pathways (eg. spliceosome, Wnt and TGF-β signaling pathways) (Figure 6).

To construct a model to differentiate ESKD patients and normal controls, we applied the Lasso logistic regression analysis on the 64 candidate genes in the training set (identified by WGCNA and DEGs approach) and selected 5 genes: CNOT8, MST4 and PPP2CB (in the turquoise module), PCSK7 (in the brown module), and RBBP4 (in the blue module); see Table 1. The model correctly identified 74 out of 75 ESKD patients and 18 out of 20 healthy controls, with an accuracy of 96.8% and

AUC of 0.943 (95%CI: 0.875 -1.000). Other diagnostic metrics are listed in Table 2.

The 5- gene model was validated using the GSE70528 dataset. The accuracy of the classifier was 91.7% (with only one misclassification). The AUC of the classifier for the validation set was 0.900 (95%CI: 0.704- 1.000).

Discussion

Since interactions among a cell's numerous constituents (ie. DNA, RNA, et al) play an important role in regulating biological functions [9], network based approaches, including WGCNA, have been widely used in kidney research and other biomedical studies [12-14, 29] to capture altered molecular networks and pathways. Once driver modules or pathways are identified, it is crucial to pinpoint key genes for in-depth pathophysiological study and intervention target identification. Variable selection has become a routinely used method for identifying relevant biomarkers [11], and Lasso regression is widely used for binary classifications [14].

The combined use of co-expression network-based analysis (WGCA) and variable selection techniques can be powerful for identifying novel biomarkers and developing predictive models for ESKD risks. Indeed, the predictive model based on the five selected differentially expressed genes showed a strong discrimination power to classify ESKD and controls in the validation dataset.

The modules that were enriched by GO terms and KEGG pathways highlighted several biological processes that might be closely associated with ESKD, such as the RNA-splicing related process and autophagy pathway for the turquoise module, chromatin and histone modification for the blue module, erythrocyte differentiation/homeostasis, endocrine resistance and mitophagy for the brown module, and protein catabolic process, histone methylation, TGF-β and Wnt signaling pathways for the candidate genes.

These findings confirm several previously reported pathways in ESKD. For example, the TGF-β pathway is actively involved in the progression of CKD via fibrogenesis, apoptosis, epithelial-to-mesenchymal transition (EMT), and inflammation [30]. Moreover, dysregulation of Wnt/β-catenin is associated with renal

fibrosis after injury, possibly via regulated effects on the downstream mediators implicated in kidney fibrosis, such as fibronectin, Snail1, matrix metalloproteinase-7, and hepatocyte growth factor [31]. However, the role of autophagy in CKD remains controversial. In some CKD models, autophagy protects against renal disease progression, but can also be profibrotic in different conditions [7]. Specifically, mitophagy was selected by KEGG pathway analysis in the current study, confirming the involvement of autophagy at the mitochondria in kidney disease [32].

Our results showed some novel ESKD mechanisms. The epigenetics-related process, including histone modification and covalent chromatin modification, was highlighted in the network-based analysis. A growing number of studies have revealed the role of histone modification in CKD. For example, histone H3 lysine methylation was involved in the TGF-β1-induced expression of ECM-associated genes (ie. connective tissue growth factor, collagen-α1, and plasminogen activator inhibitor-1) [33], while administration of an histone deacetylase inhibitor could ameliorate renal fibrosis via modulating TGF-β and epidermal growth factor receptor signaling [34]. Though not being fully validated for clinical use, epigenetic mechanisms do provide insights for CKD intervention targets.

The GO analysis detected the involvement of the turquoise module of RNA-splicing-related biological process. RNA splicing is a tightly controlled process, during which spliceosome removes introns and joins exons to generate a mature mRNA molecule, providing transcriptional plasticity and protein variability [35, 36]. Alternative splicing plays important roles in various diseases [37, 38], such as neurological and muscle diseases, and cancers. It has been reported that the standard isoform of the cell surface protein CD44 is associated with fibrotic disease [39]. In human renal proximal tubular epithelial cells, nuclear hyaluronidase 2 (HYAL2), regulated by bone morphogenetic protein 7, could displace serine-arginine-rich splicing factor 5 from CD44 pre-mRNA and the early spliceosome to promote the accumulation of the antifibrotic CD44 isoform CD44v7/8 at the cell surface via alternative splicing [39].

These findings indicate that RNA splicing could play a role in the development of

CKD and ESKD-related complications. However, the results need to be validated through *in vivo* studies. Furthermore, it is noteworthy that RNA splicing is regulated by chromatin structure and histone modifications [40]. These two processes were both identified in our study, indicating the potential crosstalk between epigenetic changes and RNA splicing in the pathogenesis of kidney diseases.

The identification of potential pathways and biomarker genes that may be involved in ESKD might facilitate in-depth analysis for mechanisms. Protein phosphatase 2 catalytic subunit beta (PPP2CB) encodes the catalytic subunit beta isoform of the protein phosphatase 2A (PP2A), one of the most abundant serine/threonine phosphatases that catalyze the dephosphorylation of phosphoproteins with an important role in the regulation of numerous intracellular processes [41, 42]. The role of PP2A in the renal fibrosis has been previously reported. The catalytic subunit of PP2A was found to be positively correlated with extracellular matrix accumulation in the unilateral ureteral obstruction model [43]. Previous studies suggested that the inhibition of PP2A could prevent endothelial-to-mesenchymal transition and renal fibrosis [43, 44]. However, the role of differentially expressed PPP2CB in the peripheral blood in the development of ESKD has not been reported yet.

Meanwhile, mammalian sterile-20-like kinase 4 (MST4), residing within the striatin interacting phosphatase and kinase (STRIPAK) complex along with PP2A and other proteins [45], was selected in our study. MST4 is ubiquitously distributed at low levels, with high expression in the thymus, placenta and peripheral blood leukocytes [46]. The STRIPAK PP2A complex mediates the kinase activity of Hpo/MST in the Hippo signaling pathway, which has a key role in organ size control and the pathogenesis of several diseases [47, 48]. For example, the STRIPAK complex regulates autophagosome transport in neurons [49] and is associated with heart disease, diabetes, autism, and cancers [50]. Other than being a component of STRIPAK complexes, MST4 mediates cell growth and transformation via extracellular signal-regulated protein kinase (ERK) signaling pathways in *in vitro* studies [46], regulates epithelial-mesenchymal transition (EMT) in hepatocellular carcinoma [51], regulates TLR signaling and inflammatory responses via direct phosphorylation of the adaptor

TRAF6 and ameliorates experimental septic shock [52], and can be stimulated by epidermal growth factor receptor ligands to contribute to prostate cancer progression [53].

Of the five featured biomarkers, two were closely associated with the epigenetic pathways. One is retinoblastoma binding protein (RBBP4), a component of several complexes, such as the nucleosome remodeling and deacetylase (NuRD) complex [54] and polycomb repressive complex 2 [55]. It can interact with partners like histone H3 [54], through which RBBP4 plays a major role in the regulation of chromatin remodeling, histone modification and gene expression, and is implicated in various conditions, including cancer [56], sensitivity to cancer therapy [57], autoimmune exocrinopathy [58], age-related memory loss [59], and human pluripotent stem cell maintenance [60]. Additionally, RBBP4 can regulate the efficiency of importin α/β-mediated nuclear import, which is associated with cellular senescence [61]. The other is the CCR4-NOT transcription complex subunit 8 (CNOT8), which, along with CNOT7 encodes the Caf1 catalytic subunit of Ccr4-Not deadenylase complex. The complex participates in the transcription and histone modification in the nucleus and mRNA turn over through degradation of eukaryotic mRNA by removing the poly(A) tail of mRNA in the cytoplasm, and, thus, is essential for accurate gene expression [62, 63]. Additionally, Caf1 interacts with BTG/TOB proteins and other substrates to regulate cell cycle [64] and proliferation [65].

The remaining biomarker was protein convertase subtilisin/kexin type 7 (PCSK7). Previous studies revealed that PCSK7 was associated with lipids (especially triglyceride) [66], insulin sensitivity via interaction with dietary carbohydrate [67], liver cirrhosis in hereditary hemochromatosis [68], and soluble transferrin receptor (sTfR) levels, a marker for iron deficiency and erythropoietic activity [69]. The roles of the featured biomarkers in CKD and ESKD are worthy of further investigation, since the conditions and pathways that are associated with CKD progression have not been fully studied.

This study has some limitations. Firstly, alterations of gene expression might be influenced by fluctuations in peripheral blood [16]. However, since kidney biopsy is

generally not safe and feasible in ESKD patients, the gene expressions in PBCs may hint at potential mechanisms. Secondly, the sample size of the validation set was small. Further studies are required to validate the obtained results.

In summary, WGCNA was performed to identify three modules of genes closely associated with ESKD and potential pathophysiological processes. Lasso regression has further identified 5 predictive genes. These findings might inform in-depth research and clinical interventions for ESKD patients.

Acknowledgements:

This work was funded by Science and Technology Department of Sichuan Province, China (2016HH0069, 2016FZ0108), and Chengdu Science and Technology Bureau (2015-RK00-00252-ZF). We appreciate the important suggestions from Associated Professor Liang Ma on data interpretation.

Conflict of interest:

The authors have no competing interests to declare.

References:

1 Levin A, Tonelli M, Bonventre J*, et al.*: Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. *Lancet*. 2017; **390**: 1888-917.

2 Foreman KJ, Marquez N, Dolgert A*, et al.*: Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. *Lancet*. 2018; **392**: 2052-90.

3 Jha V, Garcia-Garcia G, Iseki K*, et al.*: Chronic kidney disease: global dimension and perspectives. *Lancet*. 2013; **382**: 260-72.

4 Chapter 9: Healthcare Expenditures for Persons With ESRD. *American Journal of Kidney Diseases*. 2018; **71**: S433-S40.

5 Jha V, Wang AY, Wang H: The impact of CKD identification in large countries: the burden of illness. *Nephrol Dial Transplant*. 2012; **27 Suppl 3**: iii32-8.

6 Galvan DL, Green NH, Danesh FR: The hallmarks of mitochondrial dysfunction in chronic kidney disease. *Kidney Int*. 2017; **92**: 1051-57.

7 Deng X, Xie Y, Zhang A: Advance of autophagy in chronic kidney diseases. *Ren Fail*. 2017; **39**: 306-13.

8 Zhang B, Horvath S: A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005; **4**: Article17.

9 Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S: Weighted gene coexpression network analysis: state of the art. *J Biopharm Stat*. 2010; **20**: 281-300.

10 Wang B, He L, Miao W*, et al.*: Identification of key genes associated with Schmid-type metaphyseal chondrodysplasia based on microarray data. *Int J Mol Med*. 2017; **39**: 1428-36.

11 Lin E, Lane HY: Machine learning and systems genomics approaches for multi-omics data. *Biomark Res*. 2017; **5**: 2.

12 Chen Y, Bi F, An Y, Yang Q: Coexpression network analysis identified Kruppel-like factor 6 (KLF6) association with chemosensitivity in ovarian cancer. *J Cell Biochem*. 2018.

13 Lu X, Deng Y, Huang L, Feng B, Liao B: A co-expression modules based gene selection for cancer recognition. *J Theor Biol*. 2014; **362**: 75-82.

14 Ma X, Tao R, Li L*, et al.*: Identification of a 5microRNA signature and hub miRNAmRNA interactions associated with pancreatic cancer. *Oncol Rep*. 2018.

15 Ma X, Tao R, Li L*, et al.*: Identification of a 5microRNA signature and hub miRNAmRNA interactions associated with pancreatic cancer. *Oncol Rep*. 2019; **41**: 292-300.

16 Scherer A, Gunther OP, Balshaw RF*, et al.*: Alteration of human blood cell transcriptome in uremia. *BMC Med Genomics*. 2013; **6**: 23.

17 Kitajima S, Iwata Y, Furuichi K*, et al.*: Messenger RNA expression profile of sleep-related genes in peripheral blood cells in patients with chronic kidney disease. *Clin Exp Nephrol*. 2016; **20**: 218-25.

18 Zhijin (Jean) Wu RI. Descrption of gcrma package [updated 2018 Oct 30; cited 2019 Apr 19]. Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/gcrma/inst/doc/gcrma2.0.pdf.

19 Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat Assoc*. 2004; **99**: 909-17.

20 Gentleman R, Carey V, Huber FH. genefilter: methods for filtering genes from high-throughput Experiments [updated 2019 Apr 16; cited 2019 Apr 19]. Available from: https://bioconductor.org/packages/release/bioc/manuals/genefilter/man/genefilter.pdf.

21 Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012; **28**: 882-83.

22 Langfelder P, Horvath S. WGCNA: Weighted Correlation Network Analysis [updated 2019 Apr 11; cited 2019 Apr 19]. Available from: https://cran.r-project.org/web/packages/WGCNA/WGCNA.pdf.

23 Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008; **9**: 559.

24 Yu G, Wang LG, Han Y, He QY: clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*. 2012; **16**: 284-7.

25 Ritchie ME, Phipson B, Wu D*, et al.*: limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015; **43**: e47.

26 Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY: cytoHubba: identifying hub objects and

sub-networks from complex interactome. *BMC Syst Biol*. 2014; **8 Suppl 4**: S11.

27 Friedman J, Hastie T, Tibshirani R: Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010; **33**: 1-22.

28 Robin X, Turck N, Hainard A*, et al.*: pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011; **12**: 77.

29 Wang Z, Chen X, Zhang D, Cao Y, Zhang L, Tang W: PYCARD Gene Plays a Key Role in Rapidly Progressive Glomerulonephritis: Results of a Weighted Gene Co-Expression Network Analysis. *Am J Nephrol*. 2018; **48**: 193-204.

30 Ju W, Smith S, Kretzler M: Genomic biomarkers for chronic kidney disease. *Transl Res*. 2012; **159**: 290-302.

31 Zuo Y, Liu Y: New insights into the role and mechanism of Wnt/beta-catenin signalling in kidney fibrosis. *Nephrology (Carlton)*. 2018; **23 Suppl 4**: 38-43.

32 Kimura T, Isaka Y, Yoshimori T: Autophagy and kidney inflammation. *Autophagy*. 2017; **13**: 997-1003.

33 Sun G, Reddy MA, Yuan H, Lanting L, Kato M, Natarajan R: Epigenetic histone methylation modulates fibrotic gene expression. *J Am Soc Nephrol*. 2010; **21**: 2069-80.

34 Liu N, He S, Ma L*, et al.*: Blocking the class I histone deacetylase ameliorates renal fibrosis and inhibits renal fibroblast activation via modulating TGF-beta and EGFR signaling. *PLoS One*. 2013; **8**: e54001.

35 Urbanski LM, Leclair N, Anczukow O: Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip Rev RNA*. 2018; **9**: e1476.

36 Wang ET, Sandberg R, Luo S*, et al.*: Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; **456**: 470-6.

37 Tang JY, Lee JC, Hou MF*, et al.*: Alternative splicing for diseases, cancers, drugs, and databases. *ScientificWorldJournal*. 2013; **2013**: 703568.

38 Scotti MM, Swanson MS: RNA mis-splicing in disease. *Nat Rev Genet*. 2016; **17**: 19-32.

39 Midgley AC, Oltean S, Hascall V*, et al.*: Nuclear hyaluronidase 2 drives alternative splicing of CD44 pre-mRNA to determine profibrotic or antifibrotic cell phenotype. *Sci Signal*. 2017; **10**.

40 Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T: Epigenetics in alternative pre-mRNA splicing. *Cell*. 2011; **144**: 16-26.

41 Barford D: Molecular mechanisms of the protein serine/threonine phosphatases. *Trends Biochem Sci*. 1996; **21**: 407-12.

42 Lin Q, Buckler ESt, Muse SV, Walker JC: Molecular evolution of type 1 serine/threonine protein phosphatases. *Mol Phylogenet Evol*. 1999; **12**: 57-66.

43 Hou T, Xiao Z, Li Y*, et al.*: Norcantharidin inhibits renal interstitial fibrosis by downregulating PP2Ac expression. *Am J Transl Res*. 2015; **7**: 2199-211.

44 Deng Y, Guo Y, Liu P*, et al.*: Blocking protein phosphatase 2A signaling prevents endothelial-to-mesenchymal transition and renal fibrosis: a peptide-based drug therapy. *Sci Rep*. 2016;

6: 19821.

45 Kean MJ, Ceccarelli DF, Goudreault M*, et al.*: Structure-function analysis of core STRIPAK Proteins: a signaling complex implicated in Golgi polarization. *J Biol Chem*. 2011; **286**: 25065-75.

46 Lin J-L, Chen H-C, Fang H-I, Robinson D, Kung H-J, Shih H-M: MST4, a new Ste20-related kinase that mediates cell growth and transformation via modulating ERK pathway. *Oncogene*. 2001; **20**: 6559.

47 Zheng Y, Liu B, Wang L, Lei H, Pulgar Prieto KD, Pan D: Homeostatic Control of Hpo/MST Kinase Activity through Autophosphorylation-Dependent Recruitment of the STRIPAK PP2A Phosphatase Complex. *Cell Rep*. 2017; **21**: 3612-23.

48 Bae SJ, Ni L, Osinski A, Tomchick DR, Brautigam CA, Luo X: SAV1 promotes Hippo kinase activation through antagonizing the PP2A phosphatase STRIPAK. *Elife*. 2017; **6**.

49 Neisch AL, Neufeld TP, Hays TS: A STRIPAK complex mediates axonal transport of autophagosomes and dense core vesicles through PP2A regulation. *J Cell Biol*. 2017; **216**: 441-61.

50 Hwang J, Pallas DC: STRIPAK complexes: structure, biological function, and involvement in human diseases. *Int J Biochem Cell Biol*. 2014; **47**: 118-48.

51 Lin ZH, Wang L, Zhang JB*, et al.*: MST4 promotes hepatocellular carcinoma epithelial-mesenchymal transition and metastasis via activation of the p-ERK pathway. *Int J Oncol*. 2014; **45**: 629-40.

52 Jiao S, Zhang Z, Li C*, et al.*: The kinase MST4 limits inflammatory responses through direct phosphorylation of the adaptor TRAF6. *Nat Immunol*. 2015; **16**: 246-57.

53 Sung V, Luo W, Qian D, Lee I, Jallal B, Gishizky M: The Ste20 kinase MST4 plays a role in prostate cancer progression. *Cancer Res*. 2003; **63**: 3356-63.

54 Millard CJ, Varma N, Saleh A*, et al.*: The structure of the core NuRD repression complex provides insights into its interaction with chromatin. *Elife*. 2016; **5**: e13941.

55 Sun A, Li F, Liu Z*, et al.*: Structural and biochemical insights into human zinc finger protein AEBP2 reveals interactions with RBBP4. *Protein Cell*. 2018; **9**: 738-42.

56 Li L, Tang J, Zhang B*, et al.*: Epigenetic modification of MiR-429 promotes liver tumour-initiating cell properties by targeting Rb binding protein 4. *Gut*. 2015; **64**: 156-67.

57 Kitange GJ, Mladek AC, Schroeder MA*, et al.*: Retinoblastoma Binding Protein 4 Modulates Temozolomide Sensitivity in Glioblastoma by Regulating DNA Repair Proteins. *Cell Rep*. 2016; **14**: 2587-98.

58 Ishimaru N, Arakaki R, Yoshida S, Yamada A, Noji S, Hayashi Y: Expression of the retinoblastoma protein RbAp48 in exocrine glands leads to Sjogren's syndrome-like autoimmune exocrinopathy. *J Exp Med*. 2008; **205**: 2915-27.

59 Kosmidis S, Polyzos A, Harvey L*, et al.*: RbAp48 Protein Is a Critical Component of GPR158/OCN Signaling and Ameliorates Age-Related Memory Loss. *Cell Rep*. 2018; **25**: 959-73.e6.

60 O'Connor MD, Wederell E, Robertson G*, et al.*: Retinoblastoma-binding proteins 4 and 9 are important for human pluripotent stem cell maintenance. *Exp Hematol*. 2011; **39**: 866-79 e1.

61 Tsujii A, Miyamoto Y, Moriyama T*, et al.*: Retinoblastoma-binding Protein 4-regulated Classical Nuclear Transport Is Involved in Cellular Senescence. *J Biol Chem*. 2015; **290**: 29375-88.

62 Yamashita A, Chang TC, Yamashita Y*, et al.*: Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nat Struct Mol Biol*. 2005; **12**: 1054-63.

63 Sandler H, Kreth J, Timmers HT, Stoecklin G: Not1 mediates recruitment of the deadenylase Caf1 to mRNAs targeted for degradation by tristetraprolin. *Nucleic Acids Res*. 2011; **39**: 4373-86.

64 Morel AP, Sentis S, Bianchin C*, et al.*: BTG2 antiproliferative protein interacts with the human CCR4 complex existing in vivo in three cell-cycle-regulated forms. *J Cell Sci*. 2003; **116**: 2929-36.

65 Doidge R, Mittal S, Aslam A, Winkler GS: The anti-proliferative activity of BTG/TOB proteins is mediated via the Caf1a (CNOT7) and Caf1b (CNOT8) deadenylase subunits of the Ccr4-not complex. *PLoS One*. 2012; **7**: e51331.

66 Kurano M, Tsukamoto K, Kamitsuji S*, et al.*: Genome-wide association study of serum lipids confirms previously reported associations as well as new associations of common SNPs within PCSK7 gene with triglyceride. *J Hum Genet*. 2016; **61**: 427-33.

67 Huang T, Huang J, Qi Q*, et al.*: PCSK7 genotype modifies effect of a weight-loss diet on 2-year changes of insulin resistance: the POUNDS LOST trial. *Diabetes Care*. 2015; **38**: 439-44.

68 Stickel F, Buch S, Zoller H*, et al.*: Evaluation of genome-wide loci of iron metabolism in hereditary hemochromatosis identifies PCSK7 as a host risk factor of liver cirrhosis. *Hum Mol Genet*. 2014; **23**: 3883-90.

69 Oexle K, Ried JS, Hicks AA*, et al.*: Novel association to the proprotein convertase PCSK7 gene locus revealed by analysing soluble transferrin receptor (sTfR) levels. *Hum Mol Genet*. 2011; **20**: 1042-7.

Supporting information legend:

Figure S1. Selection of candidate genes. Connectivity: genes with highest connectivity in the blue, brown, and turquoise modules identified by WGCNA; Module membership: genes with high module membership within the blue, brown, and turquoise modules identified by WGCNA; Gene significance: genes with significant association with ESKD in the blue, brown, and turquoise modules identified by WGCNA; DEGs & PPI degree: DEGs identified by empirical Bayes methods and with node degree≧2 in the protein-protein interaction network. WGCNA, weighted gene co-expression network; ESKD, end-stage kidney disease; DEGs, differentially expressed genes; PPI, protein-protein interaction.

Figure legends:

Figure 1. Diagram of the study.

Figure 2. The scale free topology fitting index (R^2) and the mean connectivity for different soft thresholds (power).

Figure 3. Modules in the co-expression network. A. Clustering dendrogram of genes and modules identified by WGCNA. Branches of the hierarchical clustering tree, denoted by different colors, correspond to the modules consisting of the genes whose expression profiles are highly correlated [8], and the non-module genes are color-coded as grey. B. TOM plot of randomly selected 1,000 genes. The color code from yellow to red denotes the strength of correlations between genes. Dark squares along the diagonal correspond to the modules. C. Heatmap plot of the adjacencies of modules' eigengenes. Red color denotes high adjacency between modules' eigengenes, the first principal component for each gene module, while blue color denotes low adjacency. WGCNA, weighted gene co-expression network; TOM, topological overlap matrix.

Figure 4. Selection of candidate biomarker genes. A. Correlation matrix of MEs and clinical traits. Rows correspond to module's eigengenes, and columns correspond to clinical traits. Pearson's r and p-value are presented in the cells. The different shades of color indicate the correlation strength: from blue (not significantly correlated) to red (highly significantly correlated). B. Gene significance of ESKD across modules. The height of bars represents the average $-\log_{10}$ (p-value) for individual genes' correlations with ESKD in the modules. C a-c. Scatterplots of gene significance versus module membership for ESKD associated modules (blue, brown, turquoise), along with correlations and p-values indicated. D. The heatmaps of DEGs between ESKD and the normal controls. The x-axis and y-axis present samples and DEGs, respectively. MEs, module eigengenes; ESKD, end-stage kidney disease; DEGs, differentially expressed genes.

Figure 5. Functional enrichment of genes in the turquoise module, the blue module, the brown module, and the candidate genes. The length of bars shows gene ratio, and the colors indicate the value of -$\log_{10}$ (p-value) for enrichment analysis.

Figure 6. Protein-protein interaction network of candidate genes among the turquoise, blue and brown modules.

Tables:

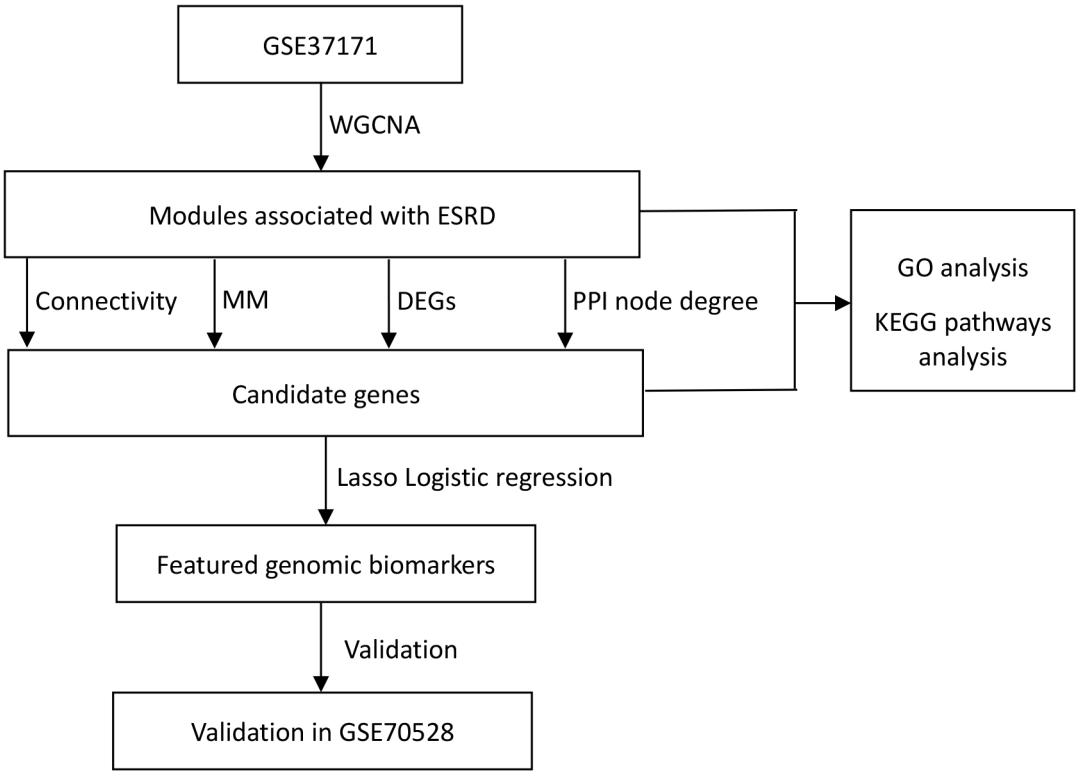
Table 1. Genes selected by the Lasso logistic regression, with the estimated coefficients and odds ratio.

Gene	Coefficient	Odds Ratio
CNOT8	-0.758	0.47
MST4	-0.872	0.42
PPP2CB	-0.290	0.75
PCSK7	-2.402	0.09
RBBP4	-1.266	0.28

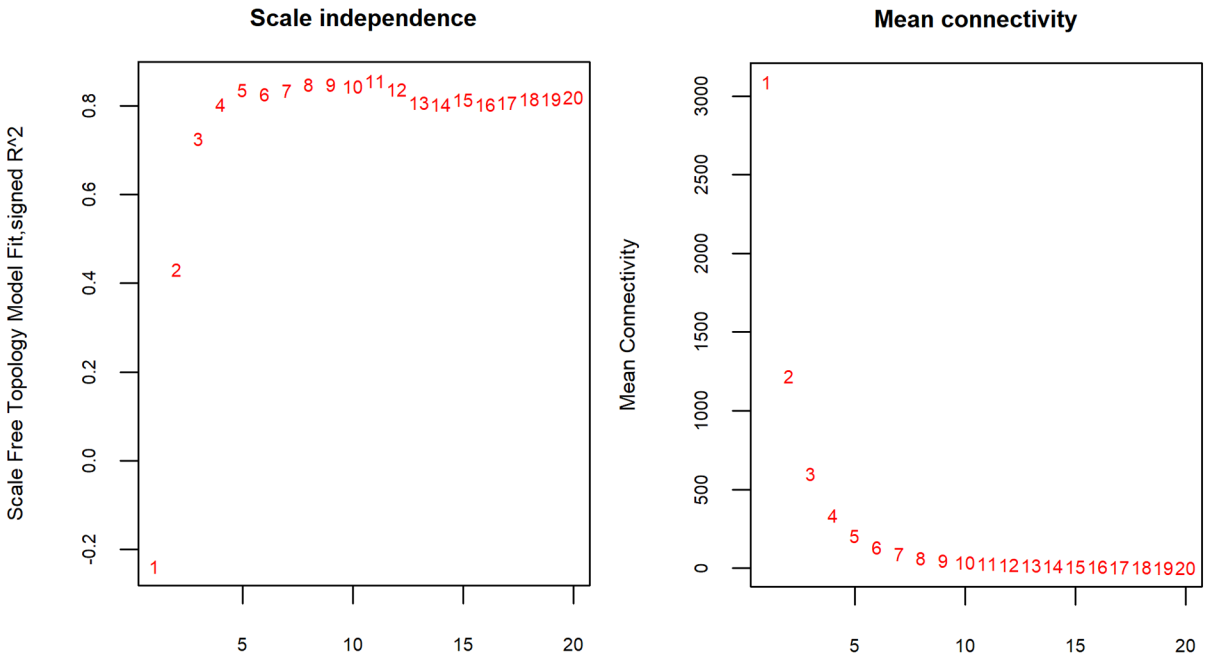
Table 2. Performance of the classifier obtained by Lasso logistic regression analysis in the training and validation sets.

	Sensitivity	Specificity	PPV	NPV	AUC
Training	0.989	0.900	0.979	0.947	0.943
Validation	1.000	0.875	0.800	1.000	0.900

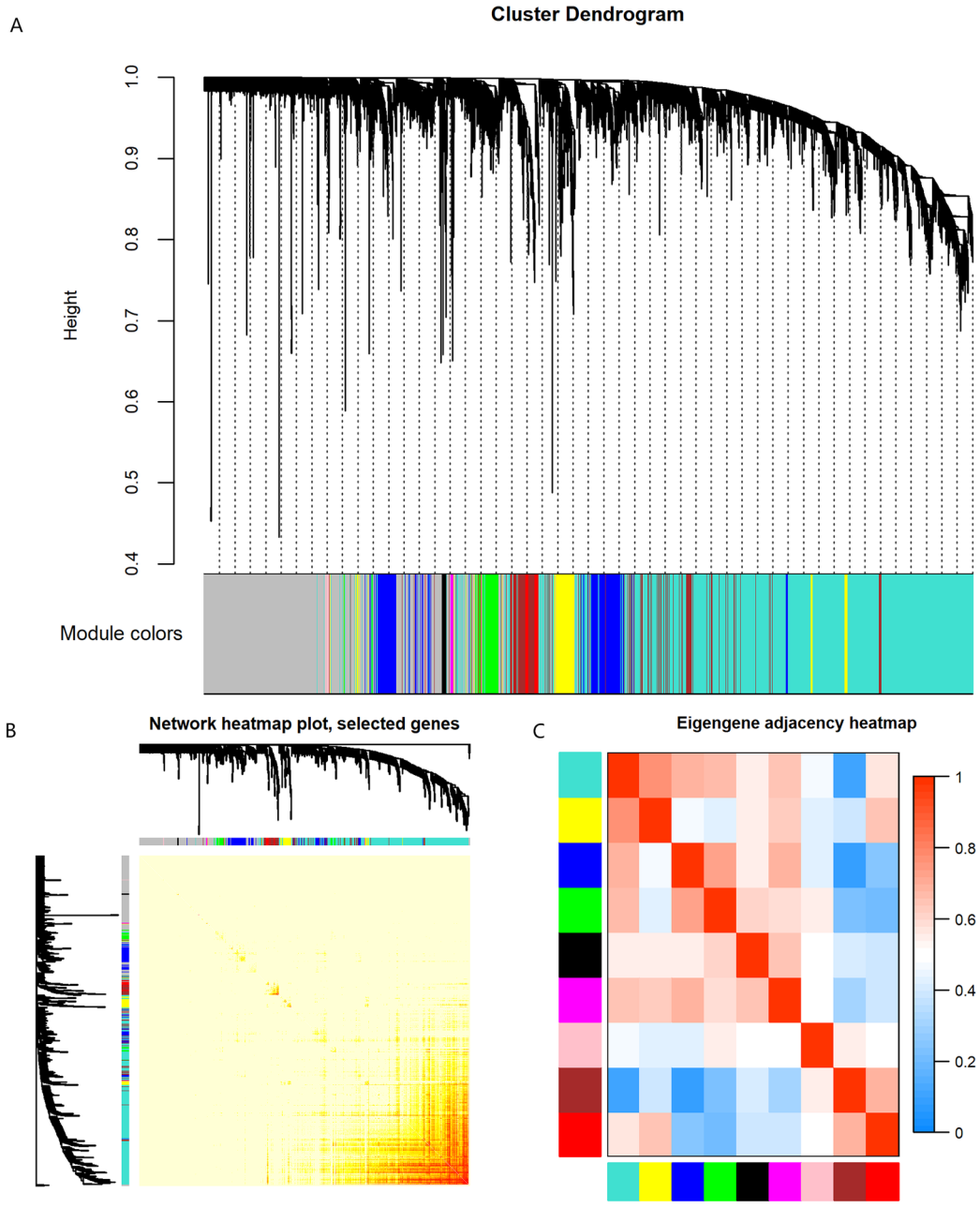
Abbreviations: PPV, Positive predictive value; NPV, Negative predictive value; AUC, area under the curve.



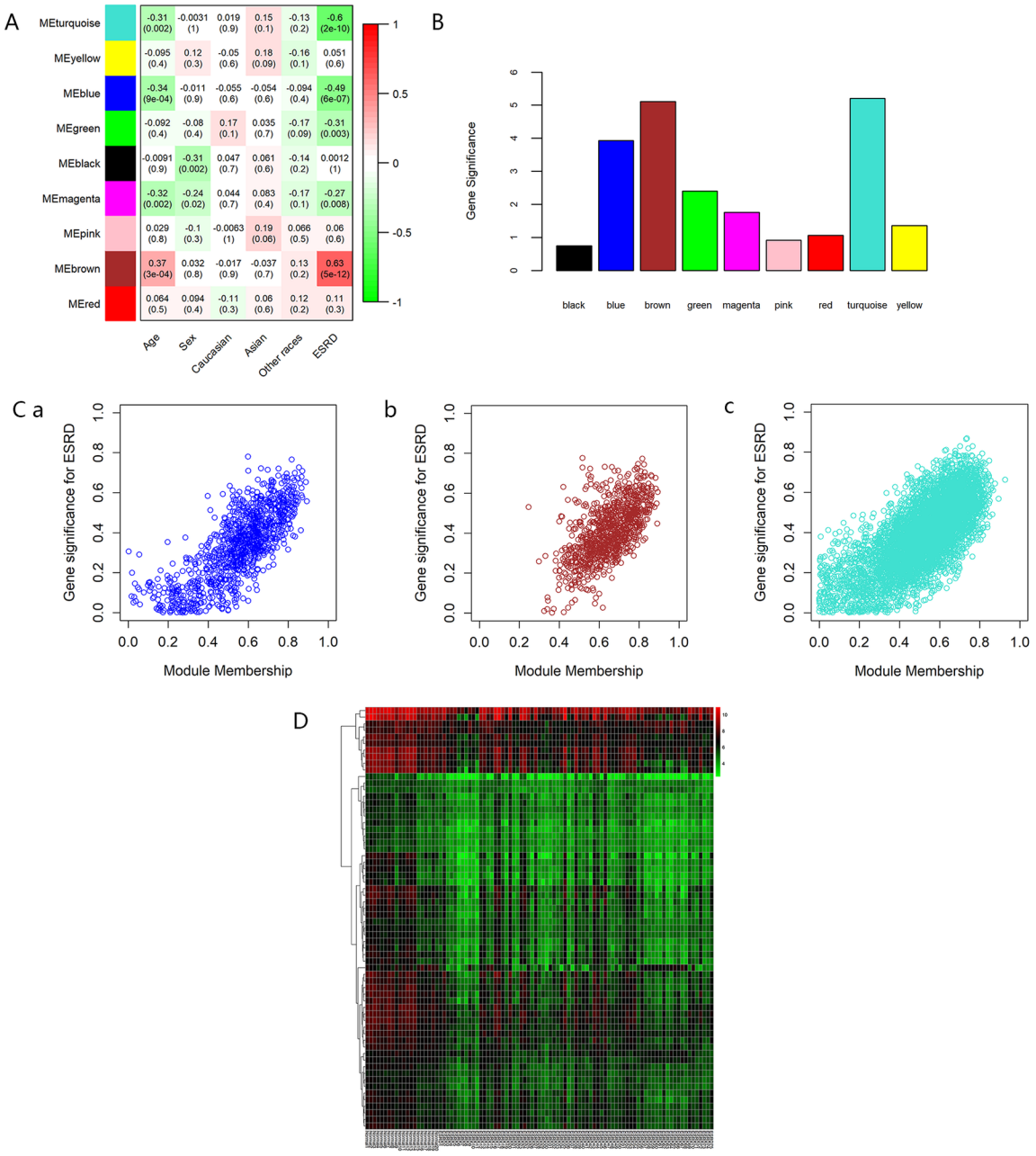
NEP_13655_Figure 1.tif



NEP_13655_Figure 2.tif

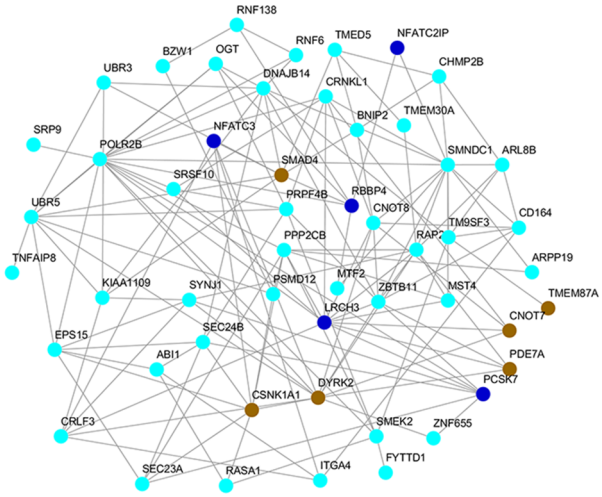


NEP_13655_Figure 3.tif



NEP_13655_Figure 4.tif

NEP_13655_Figure 5.tif



NEP_13655_Figure 6.tif