fortune Journals

Research Article

JOURNAL OF BIOINFORMATICS AND SYSTEMS BIOLOGY

ISSN: 2688-5107

Gene Cluster Expression Index for IgA Nephropathy

Aibing Rao[1+], Weiyi Guo[2+], Hong Cheng[2*], Jun Xiao[3], Yan Zeng[3], Qiuyue Li[3], Jing Zhou[3*]

Abstract

Background and objectives: Whole transcriptome gene expression data from microarray and next-generation sequencing (NGS) have made possible for large-scale data analysis for biomarker discovery and pathogenesis. IgA nephropathy (IgAN) is the most popular nephritis in eastern Asia and in the world, but its pathogenesis has not been fully elucidated. We hereby study the gene cluster expression abnormality of IgAN compared to other kidney diseases and healthy controls using tissue microarray data.

Methods: Publicly available needle biopsy tissue gene expression data sets for IgAN, other kidney diseases, and healthy control (HC) were used to develop the gene cluster expression analysis for nine representative pathways (clusters). By combining single variate prediction using ROC (receiver operating characteristic) and keyword search of gene function, nine gene clusters were heuristically determined. The gene expression status (up, down, normal) of a given gene was firstly determined by ROC, and then the gene cluster expression status, called the gene cluster expression index (GCEI), was determined by the percentages of abnormally expressed members. At last IgAN risks were assessed in the spectrum of GCEIs.

Results: Samples were classified into normal (GCEI=0) or abnormal (GCEI=1) for nine predefined gene clusters respectively, and were classified into one of 10 combinatory cGCEI groups. The percentages of IgAN differed dramatically among the groups. The percentages in the abnormal groups (GCEI=1) ranged from 40% to 70% while the percentages were from 16% to 30% in the normal groups (GCEI=0) in both training and testing data sets. The percentages of IgAN corresponding to cGCEIs trended up when the number of abnormal clusters went from 0 to 9, starting from 8.06% to 88.24%.

Conclusions: The binary GCEI is proposed to indicate whether a cluster of genes is normal (0) or abnormal (1) in terms of gene expression, and the categorical combinatory cGCEI represents the number of abnormal gene clusters. They are highly correlated to different IgAN risks so that they can be used as novel disease molecular sub-typing tools for future IgA nephropathy management and treatments.

Keywords: Gene cluster; Nephropathy; Transcriptome;

Abbreviations AUC: area under the curve; DN: diabetic nephropathy; FSGS: focal segmental glomerulosclerosis; GCEI: gene cluster expression index; cGCEI: combinatory gene cluster expression index; HC: healthy control; IgA: Immunoglobin A; IgAN: IgA nephropathy; FPR: false positive rate; MCD: minimal change disease; GN: glomerulonephritis; ROC: receiver operating characteristic; SLE: systemic lupus erythematosus; TPR: true positive rate.

Affiliation:

[1]Shenzhen Luwei (Biomanifold) Biotechnology Limited, Shenzhen, China

[2]Division of Nephrology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

[3]Department of Nephrology, The First Affiliated Hospital of Nanchang University, Nanchang, China

***Corresponding author:**
Hong Cheng, Division of Nephrology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

Jing Zhou, Department of Nephrology, The First Affiliated Hospital of Nanchang University, Nanchang, China

+: These authors contributed equally to this work.

Citation: Aibing Rao, Weiyi Guo, Hong Cheng, Jun Xiao, Yan Zeng, Qiuyue Li, Jing Zhou. Gene Cluster Expression Index for IgA Nephropathy. Journal of Bioinformatics and Systems Biology. 7 (2024): 196-211.

Received: August 23, 2024
Accepted: August 29, 2024
Published: October 08, 2024

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

Introduction

Immunoglobin A (IgA) nephropathy (IgAN), also called Berger disease, is one of the most prevalent kidney diseases in the world and is more frequent in Asian populations. The pathogenesis has not been fully elucidated and the current understanding is the popular multi-hit theory. Due to hereditary mutations or somatic factors, abnormal B-lymphocyte secretes abnormal Galactose-deficient IgA1 (Gd-IgA). Respiratory tract infections or gut infections give rise to a great amount of Gd-IgA1, targeted by glycan-specific IgG antibodies, and they together form IgA immune complexes. The abnormal IgA immune complexes are hardly cleared by the immune system, via circulation they slowly deposit in the mesangial region of the kidney, triggering local inflammation and leading to irreversible injury [1]. Needle biopsy pathologic inspection is the gold standard of IgAN diagnosis. In the past decade, a tremendous effort has been put in the search of new biomarkers and transcriptomic analysis was predicted to be an important direction for the personalized treatment and management of IgAN in the future [2]. Molecular profiling has been applied to find proteinuria signature [3], metabolic pathways [4], gene activations [5], inflammatory response transcriptional network [6], prediction of drug positioning [7], epidermal growth factor biomarkers [8], mesangial cell function in IgAN [9], etc. These research results have shown the complexity of the IgAN disease and other kidney diseases in the spectrum of gene expression, hence it is important to develop new analysis techniques from the system biology point of view by analyzing gene clusters. In this research, we applied the framework of gene cluster expression abnormality analysis developed for lung cancer molecular sub-typing [10] to IgAN.

Materials and Methods:

Training and testing data sets:

The training data set was stacked from three public data sets downloaded from Gene Expression Omnibus (GEO): GSE35489 [3], GSE104954 [4] and GSE115857, and the testing data set was constructed from four GEO sets: GSE116626 [5], GSE37460 [6, 7], GSE69438 [8] and GSE93798 [9]. These data sets are microarray data of kidney tissues from different platforms and the population consists of IgA nephropathy (IgAN), healthy control (HC), and other kidney diseases (collectively denoted as KD) such as diabetic nephropathy (DN), focal segmental glomerulosclerosis (FSGS), minimal change disease (MCD), glomerulonephritis (GN), systemic lupus erythematosus (SLE), etc., as described in Table 1. Since the analysis is typically formulated as a binary classification problem, the IgAN group is indexed as 1 (positive) and all the other groups (HC+KD) is indexed as 0. In summary of Table 1, the training data consists of 105 IgAN (positive) and 207 negative objects, while the testing data set has 101 positive and 133 negative objects. In order to stack the GEO data sets, for each GEO, at first, a normalization procedure was applied to each probe and then to each sample. The normalization is a linear map: (Q25, Q75) —> (0, 1), where Q25, Q75 are the 25th and the 75th percentile of a data vector; second, an average was taken with the normalized values of the probes mapped to the same gene and assigned to the gene; third, genes annotated by gencode. v22.annotation (https://www.encodeproject.org/files/gencode.v22.annotation/) as "protein-coding type" were used for the analysis. Moreover, genes missing in one of the seven data sets were omitted. Hence the training and the testing data sets contained 10459 common genes.

Table 1: The number of sample types in the training and the testing GEO data sets. HC: healthy control; IgAN: IgA nephropathy; CKD: chronicle kidney disease; DN: diabetic nephropathy; FSGS: focal segmental glomerulosclerosis; GN: glomerulonephritis; MCD: minimal change disease; SLE: systemic lupus erythematosus; etc.: all other kidney diseases.

GEON	HC	IgAN	CKD	DN	FSGS	GN	MCD	SLE	etc.
GSE115857	7	55	0	0	1	11	12	0	0
GSE35489	6	25	0	0	0	0	0	0	0
GSE104954	21	25	0	17	13	39	17	32	31
GSE116626	7	52	22	0	0	0	0	0	0
GSE37460	27	27	0	0	0	0	0	0	15
GSE93798	22	20	0	0	0	0	0	0	0
GSE69438	0	2	6	1	11	4	1	16	1

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

The gene clusters determined by ROC and keyword search

The samples in the training and the testing data sets were flagged as 1 for IgAN or 0 for HC+KD. For each gene, the receiver operating characteristic (ROC) method [11] was applied for a binary classification with respect to the flag (n1 = 105, n0 = 207). All genes with AUC equal to 0.5 (a random classification) were omitted. Next nine keywords INFECTI, CELL CYCLE, CYTOSKELETON, DNA REPAIR, GTPASE, IMMUNE, MITOCHON, PODOCYTE, and WNT were used to case-insensitively search the gene function annotations in Uniprot (https://www.uniprot.org/) database. Note that the first keyword INFECTI was designated so that both infection and infectious could be matched, MITOCHON was designated for the same reason. The matched genes were included as the candidate cluster members correspond to each of the nine clusters (pathways/functions): infection, cell cycle, cytoskeleton, DNA repair and chromosome remodeling, GTPase, immune system, mitochondria, podocyte, and Wnt signaling pathway. Note that nine clusters were selected because they might be related to IgAN pathogenesis [1, 12], and any other possible cluster is also applicable in the following. Each gene in a cluster was analyzed using ROC according to the procedure as described in Rao [10]. In more detail, in order to plot the correct ROC for a gene, first decide if the gene is up- or down-regulated. When the average expression of the IgAN group is greater than or equal to that of HC+KD, it's up-regulated, or else it is down-regulated. To plot ROC for a down-regulated gene, negate its expression by using the negative expression. Using the cutoff derived by ROC, which was set as the threshold at the point on the ROC curve point closest to the left-top corner of the ROC unit box. TPR (true positive rate) is the percent of IgAN samples with expression greater than or equal to the cutoff while FPR (false positive rate) is the percent of IgAN samples with expression less than the cutoff. Samples of expression greater than or equal to the cutoff were assigned prediction value 1, or 0 otherwise. Let $P\delta = TPR - FPR$, which is a reasonable measure of the prediction power of the gene. The range of $P\delta$ is (0, 1) where 1 is the perfect prediction and 0 is the random prediction (no power). A small $P\delta$ might also be due to the experiment noise hence all genes with $P\delta < 5\%$ were considered to be of no prediction power and were labeled as normal, and all the other genes were labeled as abnormal.

Gene cluster expression index (GCEI) and combinatory GCEI (cGCEI)

The population statistics were collected at the cluster level. Given a cluster, for each sample, the percentage of abnormally expressed genes within the cluster was calculated. ROC using the percentage to predict the IgAN index was plotted and a cutoff was also determined. A sample with a percentage greater or equal to the cutoff is labeled as 1, or 0 otherwise, called the gene cluster expression index (GCEI) corresponding to the cluster.

Putting all nine GCEI together, each sample was assigned a binary string of length 9, corresponding to the GCEI of nine clusters in the order of INFECTI, CELL CYCLE, CYTOSKELETON, DNA REPAIR, GTPASE, IMMUNE, MITOCHON, PODOCYTE, and WNT. For example, 100001000 represents that the infection (the first digit is 1) and the immune system (the sixth digit is 1) cluster are abnormally expressed and the rest are normal, while 000000000 and 111111111 are two extremes of none and all of the clusters are abnormal. Since there are 512 possible signatures and a limited number of samples (n=312), the population statistics within each signature would be too sparse, and therefore the signatures were collapsed into 10 groups by counting the number of 1's (abnormal ones), denoted as S_0, S_1, S_2,…, S_9. For example, S_0 represents signature 000000000 where none of the clusters are abnormal, and S_9 represents signature 111111111 where all of the clusters are abnormal, and S_1 contains 9 signatures with only one 1 appearing in any position, namely, 100000000, 010000000,…, 000000001, where only one of the clusters is abnormal, and so on. In more detail, 100000000 stands for the case when the first cluster, INFECTI, the infection pathway is abnormal, and 110000000 stands for the case when the first and the second clusters, INFECTI, CELL CYCLE, are abnormal, all other clusters are normal, etc. The percentages of IgAN objects were collected and compared for S_0, S_1, S_2,…,S_9. The collapsed GCEI is another index, called combinatory GCEI (cGCEI), with values from 0 to 9.

IgAN risk assessment concerning GCEI status and cGCEI value

IgAN risk is defined as the percentage of IgAN objects within a classification. For each gene cluster, a sample has a GCEI with two statuses: 1 (abnormal) and 0 (normal). The percentages of IgAN objects were collected and compared for two groups of each cluster. Moreover, the percentages of IgAN objects within each of the 10 cGCEI groups were collected.

Validation: Applying the various cutoffs obtained from the training phase as described in the above, the samples in the testing data set were labeled and the cluster GCEIs were also determined, and similar population statistics were collected and compared.

Data analysis and software: The statistics and the plots were implemented in R scripts. The ROC analysis was based on R package ROCR.

Results

Nine gene clusters

For any given cluster as defined by one keyword of INFECTI, CELL CYCLE, CYTOSKELETON,

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

DNA REPAIR, GTPASE, IMMUNE, MITOCHON, PODOCYTE, and WNT , a member gene is added if it satisfies 3 conditions: (1). Its Uniprot annotation contains the keyword; (2). AUC > 0.5; (3) Pδ ≥ 5%. Take the IMMUNE cluster as an example, 120 genes in the testing data set were included, of which 35 were up-regulated and 85 were down-regulated. Table 2 listed the top nine Down and the top nine Up genes in the decreasing order of Pδ, here only nine were selected just for demonstrating purpose to limit the table size. The corresponding ROC curves for the nine Down genes were plotted in Figure 1 (Top-Left). At the top is CEBPB (CCAAT enhancer binding protein beta), whose univariate ROC by using its expression alone gave rise to AUC = 0.61 with a sensitivity of 62% (TPR=0.62) , a specificity of 64% (1-FPR=0.64), and Cutoff = 0.33. The cutoff split 312 samples into two groups: in the group with an expression less than 0.33, there were 65 out of 105 IgAN objects (TPR = 61.90%) and 71 out of 207 HC+KD objects (FPR = 34.30%), so that Pδ = 27.60%. A summary of nine clusters is presented in Table 3 (Columns 1-4), it shows that there were 156, 145, 89, 57, 123, 120, 230, 16, and 60 member genes in the cluster order of INFECTI, CELL CYCLE, CYTOSKELETON, DNA REPAIR, GTPASE, IMMUNE, MITOCHON, PODOCYTE, and WNT respectively, and there were 27, 15, 12, 7, 10, 35, 19, 1, and 7 up-regulated genes in the same order. Interestingly most of the cluster member genes were down-regulated. In summary, all samples in the training data set were assigned an expression prediction label (0-normal, 1-abnormal) for any member genes by the corresponding cutoff derived from its ROC.

GCEI assignment

The goal is to assign each sample a GCEI value of 0 or 1 for a given cluster so that GCEI = 1 means a cluster was abnormal in terms of member expressions for the sample. Given a cluster, for any sample, the percentage of gene members of prediction = 1 was calculated and designated as a new feature. Applying ROC again to predict the IgAN flag using the new feature, another cutoff was derived. If the percent of the abnormally expressed member genes of a sample goes beyond the cutoff, the sample is called abnormal concerning the cluster, and label it as GCEI=1, otherwise label it as GCEI=0. The ROC curves of the nine clusters are presented in Figure 2 and the results are shown in Table 3 (Columns 5-8). The cutoffs are between 45% and 56% for the nine clusters, which are close to the member majority voting strategy of cutoff as 50%, but much more subtle. The AUCs range from 0.68 to 0.80. Applying the cutoffs accordingly, each sample was assigned a GCEI value for each cluster.

Table 2: ROC Results of the top nine member genes in cluster IMMUNE using the expression as the predictor and the percent of IgAN flags in the two groups by the cutoff. AUC: area under the curve; FPR: false positive rate; TPR: true positive rate; Expr.: regulation directions, Up- IgAN group has higher expression than HC+KD, Down - Otherwise; Pδ = TPR − FPR.

GENE	AUC	Cutoff	Expr.	FPR	TPR	P_δ
CEBPB	0.61	0.33	Down	34.3	61.9	27.6
TRIM10	0.6	0.26	Down	35.27	62.86	27.59
NFKBIA	0.6	0.49	Down	39.13	64.76	25.63
TFEB	0.6	0.2	Down	32.85	57.14	24.29
RIPK2	0.61	0.46	Down	40.58	64.76	24.18
LRRC19	0.59	0.43	Down	37.68	60.95	23.27
FGA	0.57	0.49	Down	45.89	68.57	22.68
POLR3E	0.58	0.5	Down	41.55	62.86	21.31
SERTAD3	0.61	0.45	Down	41.55	62.86	21.31
DDX3Y	0.51	0.18	Up	53.62	71.43	17.81
CXCL9	0.51	0.23	Up	56.04	69.52	13.48
CD48	0.55	0.39	Up	43.96	57.14	13.18
CLEC10A	0.55	0.45	Up	41.06	53.33	12.27
TRAF3IP3	0.54	0.49	Up	43.96	55.24	11.28
NFATC4	0.52	0.62	Up	39.13	49.52	10.39
LST1	0.52	0.46	Up	43	53.33	10.33
C3	0.53	0.45	Up	45.89	56.19	10.3
CEACAM3	0.52	0.55	Up	45.41	55.24	9.83

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

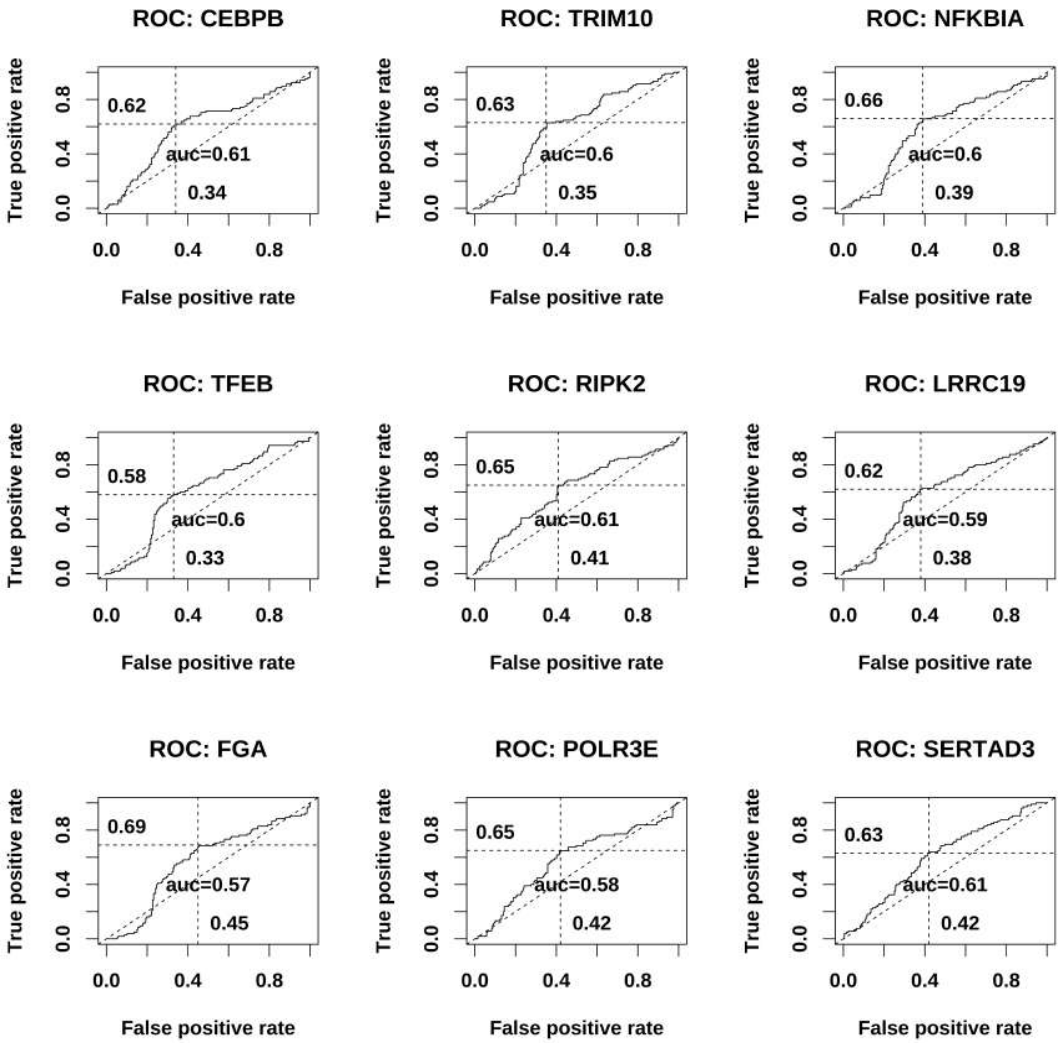

Figure 1: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster IMMUNE. Only top 9 genes in the decreasing order of Pδ are displayed.

Table 3: Cluster sizes and ROC results using the percentages of abnormally expressed members as the predictor of IgAN flags. n (Up/Down): number of up/downregulated member genes; AUC: area under the curve; FPR: false positive rate; TPR: true positive rate.

Cluster	Size	n (Up)	n (Down)	AUC	FPR	TPR	Cutoff(%)
INFECTI	156	27	129	0.75	0.25	0.69	48
CELL CYCLE	145	15	130	0.72	0.38	0.69	47
CYTOSKELETON	89	12	77	0.74	0.34	0.68	48
DNA REPAIR	57	7	50	0.7	0.22	0.56	51
GTPASE	123	10	113	0.69	0.37	0.64	45
IMMUNE	120	35	85	0.8	0.24	0.71	50
MITOCHON	230	19	211	0.69	0.35	0.65	45
PODOCYTE	16	1	15	0.68	0.34	0.57	56
WNT	60	7	53	0.74	0.27	0.64	48

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

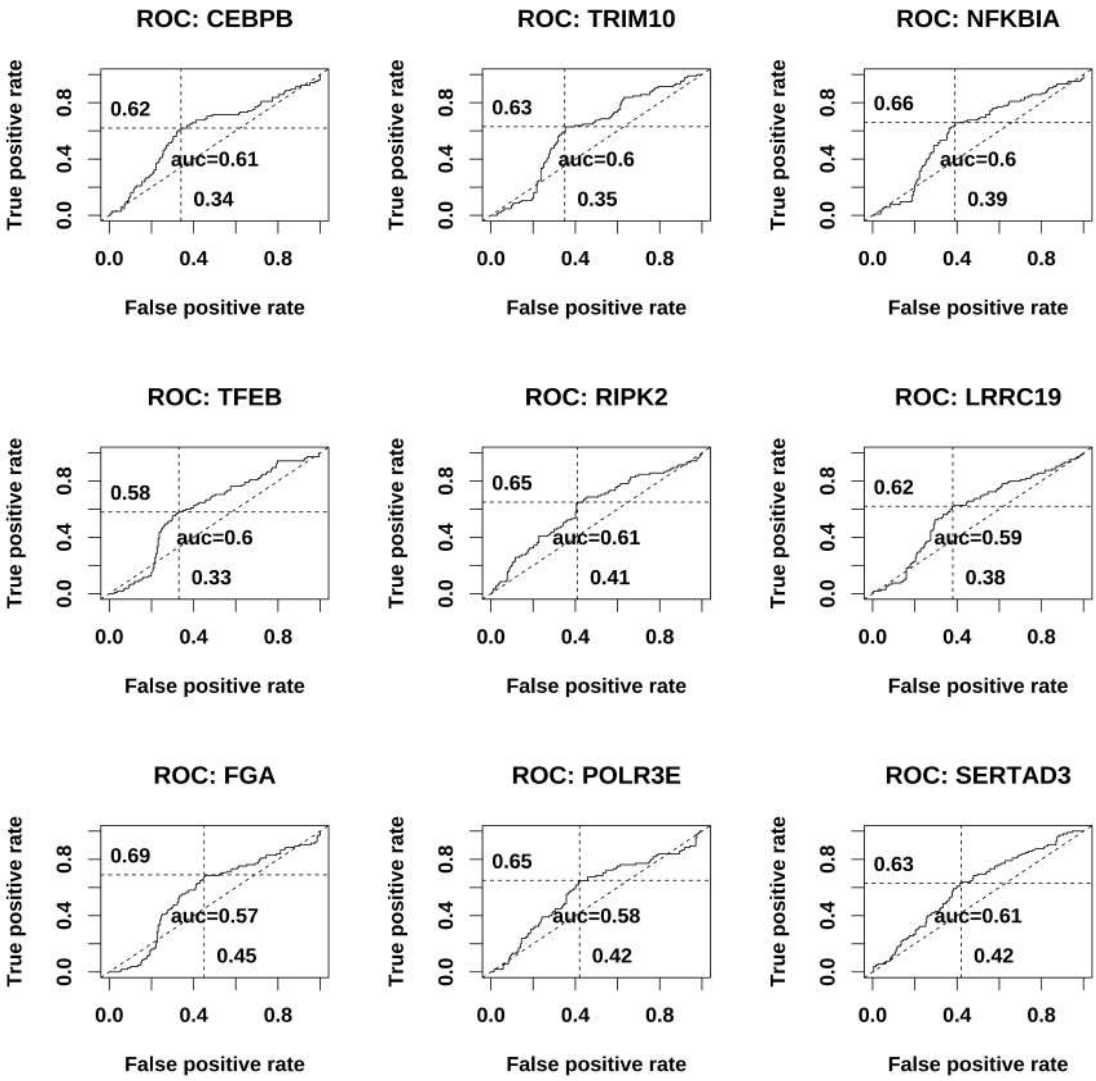

Figure 2: Receiver operating characteristic (ROC) using the percentages of abnormally expressed members as the predictor of IgAN flags for each gene cluster.

Risk assessment concerning GCEI and cGCEI

Given a GCEI of a cluster, the percentages of IgAN samples (with flag 1), called the IgAN risks, were calculated for GCEI value 0 and 1 respectively and are presented in Figure 3 as circle-marked lines. It shows that the risks for the abnormal groups of GCEI=1 range from 46% to 60% and those of GCEI=0 range from 16% to 25%. Therefore, a sample with any one of the abnormally expressed gene clusters (GCEI=1) has 2-3 times of IgAN risk, compared to the corresponding group of GCEI=0. On the other hand, by concatenating 9 atomic GCEI into a 9-digit binary string in the given order and then collapsing the 9-digit signatures into 10 groups represented by S_k, k = 0, 1,…, 9, the risk profiles are presented in Table 4. The training columns show that for group S_0 where all clusters have GCEI=0, there were a total of 79 samples within which 15 are IgAN so the IgAN risk is 18.99%, when S_k goes up from S_1 to S_9, the IgAN risk goes up with some hiccups. A sudden jump happens at S_7 where the risk is 61.9% which is more than doubled to the previous 26.32%, until 78% for S_9. Therefore, the more number of abnormal clusters with GCEI=1, the higher percent of the IgAN flags. In summary, a higher count of abnormal clusters indicates a higher likelihood a sample is IgAN.

Validation with the testing data set

As shown in Table 1, the testing data set was constructed from the bottom four GEO sets. It has 234 samples consisting of 56 HC samples, 101 IgAN samples, and 77 other KD samples. Again, the testing data set was also flagged with IgAN as 1 and HC+KD as 0 for a binary classification. The single gene cutoffs obtained at the training phase were applied first, then population statistics were collected at the cluster level to define the percentage of abnormally expressed genes in each cluster, and at last the same population cutoffs obtained at the training phase were applied to derive GCEI

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

Table 4: Percentages of IgAN flags (risks) within the sample groups with collapsed 9-digit signa- tures. S_k: sample group with k clusters of GCEI=1; n_1: number of samples with IgAN flag=1; n_t: group size; cGCEI: combinatory gene cluster expression index. As the number of abnormal clusters (GCEI=1) goes up, the corresponding group risk also goes up. At one extreme when all nine clusters are normal, the risks are 18.99%, 8.06% for the training and the testing respectively, while at the other extreme when all nine clusters are abnormal, the risks are 78%, 88.24% respectively.

cCGEI	Examplar Sig.	Training			Testing		
		n_1	n_t	Risk (%)	n_1	n_t	Risk (%)
S_0	000000000	15	79	18.99	5	62	8.06
S_1	100000000,010000000	5	37	13.51	7	23	30.43
S_2	110000000,100010000	6	27	22.22	6	16	37.5
S_3	111000000,100010001	5	20	25	4	12	33.33
S_4	111100000,100010011	5	27	18.52	4	14	28.57
S_5	111110000,100010111	3	15	20	3	7	42.86
S_6	111111000,100011111	5	19	26.32	5	12	41.67
S_7	111111100,100111111	13	21	61.9	16	20	80
S_8	111111110,101111111	9	17	52.94	21	34	61.76
S_9	111111111	39	50	78	30	34	88.24

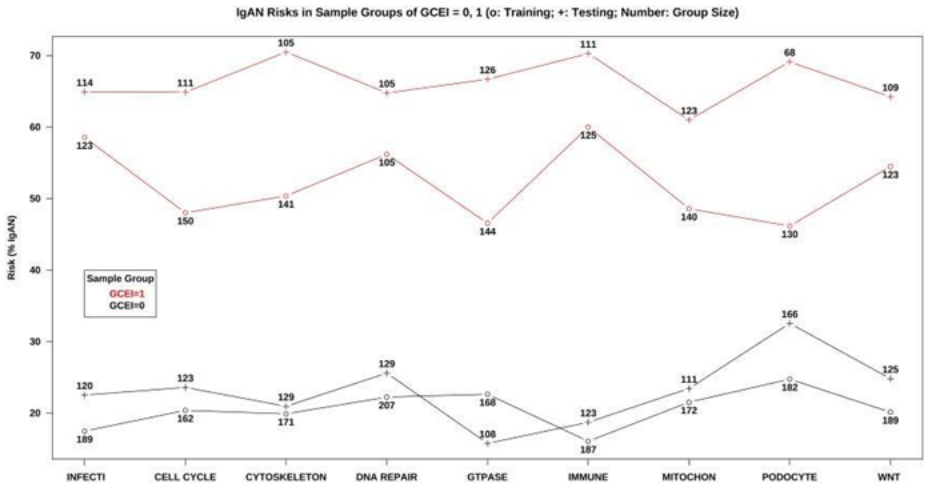

Figure 3: Percentages of IgAN flags (IgAN risks) within the sample groups of abnormal or normal clusters (GCEI=1 or 0). For each cluster, in the training phase, the risks of the abnormal group (GCEI=1) range from 46% to 60% while it is from 16% to 25% for the normal group (GCEI=0). Applying the same cutoffs to the testing data set, the abnormal group (GCEI=1) range from 60% to 70% while it is from 16% to 30% for the normal group (GCEI=0).

status. Afterward, the IgAN risks were calculated for sample groups defined by GCEI or cGCEI values respectively. The cluster-level GCEI group risks are also presented in Figure 3 as the cross-marked lines, in contrast to the circle-marked lines from the training set. It shows that the abnormal groups (GCEI=1) range from 60% to 70% while it is from 16% to 30% for the normal groups (GCEI=0), comparable to the training phase. On the other hand, for the collapsed cGCEI, Table 4 shows the risk profiles side by side with the training phase. It also shows an upward trend of the risks when S_k goes from S0 with merely 8.06% to S_9 with 88.24%, similar to the training phase.

Discussion

The binary GCEI and the categorical combinatory cGCEI proposed in this research provide tools for IgAN risk assessment and molecular sub-typing. It is when a large portion of the member genes become abnormally expressed that a cluster is labeled as GCEI = 1 (abnormal), moreover, the more the number of abnormal clusters, the larger the cGCEI, hence the higher the IgAN risk. Nine clusters have been intentionally selected due to their relevance to IgAN pathogenesis in literature. The genes included in a cluster contain the specified keywords in the annotations and genes usually play multiple roles, hence the functions of member

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

genes are not mutually exclusive among clusters. On the other hand, the selection here may be incomplete due to the complexity of IgAN pathogenesis and the same procedure is also applicable to any other cluster selection. We review the functions of some top genes found in this research. Table 5 lists the top 20 down-regulated and the top 20 up-regulated genes from all clusters in the decreasing order of Pδ respectively. The ROC results of each univariate gene and the population summary with the cutoffs are also shown. As for the individual cluster, the ROCs for the top nine genes are presented in supplemental section (Figure 4 to Figure 11). Among the top 20 down-regulated genes, half of them are related to immunity and infection responses, including ARC, UAP1, ARIH1, CEBPB, SENP2, TRIM10, NFKB1A, LRRC19, TFEB, and RIPK2. ARC is specifically expressed in skin-migratory dendritic cells and regulates fast dendritic cell migration and T-cell activation [14]; UAP1 (UDP-N-acetylglucosamine pyrophosphorylase 1), is a metabolic enzyme, participating innate immune response to virus infection by mediating IRF3 pyrophosphorylation and leading to type I interferon responses [15], hence UAP1 down-regulation in IgAN kidneys may reduce type I interferon responses and thereby may enhance tissue damage and autoimmunity; ARIH1 (ariadne RBR E3 ubiquitin protein ligase 1) was shown to enhance cellular antiviral responses [16] and promoted STING (stimulator of interferon response CGAMP interactor)-mediated T-cell activation in tumor [17], hence ARIH1 down regulation in IgAN may reduce antivirus and interferon responses; Another gene TRIM10 (tripartite motif containing 10), on one hand, also promotes type I interferon as an E3 ligase by enhancing STING1 aggregation [18], on the other hand, independent of its E3 ligase activity, TRIM10 restricts the IFN-I/JAK/STAT signaling pathway to suppress type I interferon responses [18]. Therefore TRIM10 plays a double-role in IFN1 regulation depending on whether its E3 ligase function is activated; CEBPB (CCAAT enhancer binding protein β), is a transcription factor regulating the expression of genes involved in immune and inflammatory responses, while SENP2 (SUMO specific peptidase 2) controls adipogenesis and stabilizes CEBPB via desumoylation [20], and a study showed that SENP2 restrains the generation of pathogenic Th17 cells [21], therefore SENP2 and CEBPB down-regulations in IgAN may indicate abnormal immune responses and pathogenic T-cell generation; NFKB1A (NFκB inhibitor alpha) is triggered by inflammatory responses and its repression in IgAN results in non-canonical NFκB activation [22], elevating kidney inflammation and injury [23]; LRRC19 (leucine rich repeat containing 19) also activates NFκB and induces pro-inflammatory cytokines in kidney[24]; Moreover, RIPK2 (receptor interacting Serine/Threonine kinase 2) activates TCR-induced NFκB [25]; TFEB (transcription factor EB) stimulates the intracellular clearance of pathogenic factors by enhancing autophagy and lysosomal function in multiple kidney diseases [26]. Other than immune genes, KLF10 (Kruppel-like factor 10) represses cell proliferation and inflammation by inducing apoptosis [27] and KLF10 down-regulation contributes to the regeneration of survived renal tubular cells and kidney tissue repair [28]; Epithelial cell kinase EPHA2, a receptor for hepatitis C virus (HCV) in hepatocytes, playing roles in cell-cell repulsion and adhesion, was also shown to play a critical role in tissue repair in kidney injury disease such as renal ischemia-reperfusion injury [29], and therefore EPHA2 down-regulation in IgAN may weaken tissue repair process. Other down-regulated genes include two mitochondrial genes: ACADM (Acyl-CoA dehydrogenase medium chain) catalyzes mitochondrial fatty acid β-oxidation when converting fats to energy and its lower expression was shown to predict poor prognosis for renal cancer [30]; OPA1 (optic atrophy 1) is essential for normal mitochondrial morphology, their down-regulations in IgAN indicate abnormal mitochondria function; As for metabolism, LDLR (low-density lipoprotein receptor) is popularly regarded as receptor and entry point for multiple viruses, however, a study has shown that down-regulation of LDLR changes the intercellular lipids and impairs lymphatic function, which is related to atherosclerosis [31], and another study has shown that LDLR dysfunction induces LDL accumulation and promotes pulmonary fibrosis [32], therefore it is reasonable to hypothesize that LDLR down-regulation in IgAN kidney may change the lipids or induce LDL accumulation to contribute to the injury, this should be an interesting research topic.

On the up-regulation side (Table 5) , there are 14 immunity-related genes: DDX3Y, DBN1, CCR5, CXCL9, CD48, GZMB, PRKCB, CLEC10A, IRX5, TRAF3IP3, PSTPIP1, RASGRP2, NFATC4, ARHGAP22, and the rest includes a Wnt gene VAX2, a DNA repair gene MMS19, two mitochondrial genes PITRM1 and BCKDHA, and two virus infection related genes APOBEC3G and CORO1A. DDX3Y (DEAD-box helicase 3 Y-linked) induces type I interferon responses by enhancing IFNB transcription [33]; DBN1 (Drebrin 1) is an actin-binding protein and is re- quired for chemokine receptor CXCR4 recruitment to immunological synapses [34]; CCR5 (C-C motif chemokine receptor 5) has been studied in a lot of human diseases and in renal diseases [36]. CCL5/CCR5 facilitates inflammatory responses and induces the adhesion and migration of different T cell subsets in immune responses and is involved in various pathological processes [35]; CXCL9 is chemotactic for activated T-cells; CD48 is on the surface of immune cells and participates in their activation and differentiation pathways; GZMB (granzyme B), is a member of serine proteases family expressed in the granules of cytotoxic T-lymphocyte and NK cells; PRKCB (protein kinase c beta) plays a key role in B-cell activation by regulating BCR-induced NFκB activation; CLEC10A (c-type lectin domain containing 10a) is a marker for dendritic cells and enhances their TLR-induced cytokine secretion [37]; IRX5 (iroquois homeobox 5) was shown to promote NFκB

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

signaling via activating the promotor of SPP1 (secreted phosphoprotein 1) in tumor [38], SPP1 is previously called OPN (osteopontin) and was shown to be associated with re- nal failure and might be a marker for renal dysfunction prognosis [39], our follow-on analysis also showed that IRX5 and SPP1 were up-regulated in the IgAN group in both data sets, hence we hypothesis that IRX5 up-regulation in IgAN promoted SPP1 expression and NFκB signaling to enhance type I immunity; TRAF3IP3 (TRAF3 interacting protein 3) participates antiviral innate immune response via RIG-I-MAVS pathway [40, 41]; PSTPIP1 (proline-serine-threonine phosphatase interacting protein 1) promotes the actin polymerization required for synapse induction during T-cell activation, and its mutations lead to PSTPIP1-associated myeloid-related proteinemia inflammatory such as pyogenic arthritis, pyoderma gangrenosum and acne syndromes, which may be involved with heterogeneous nephropathy [42]; RASGRP2 (RAS guanyl releasing pro- tein 2) plays roles on aggregation of platelets and adhesion of T-lymphocytes and neutrophils; NFATC4 (nuclear factor of activated T cell 4) stimulates cytokine IL2 and IL4; ARHGAP22 (Rho GTPase activating protein 22) was shown to be a novel biomarker for tumor-promoting immune infiltration with a lower Th1/Th2 cell ratio, higher DC cell infiltration, higher Treg cell infiltration, and T-cell exhaustion phenotype [43], hence it is interesting to study whether its up-regulation in IgAN implies a similar effect as to immune infiltration. Finally, there are some other notable immunity genes that were not among the top up-regulated 20 genes: PTPRC (protein yyrosine phosphatase receptor type c) is an essential regulator of T- and B-cell antigen receptor signaling; CD4 is a membrane glycoprotein of T lymphocytes; CASP1 (Caspase 1) is the precursors of the inflammatory cytokines IL1B (interleukin-1 beta) and IL18 (interleukin 18); F11L is a ligand for integrin alpha-L/beta-2 involved in memory T-cell and neutrophil transmigration. In addition to the immune genes in IgAN, VAX2 (ventral anterior homeobox 2) regulates the expression of Wnt signaling antagonists and Wnt signalling plays important roles in chronic kidney diseases including IgA nephropathy [44]; MMS19 (MMS19 Homolog, cytosolic iron-sulfur assembly component) participates in DNA repair, replication, and mitosis; PITRM1 (pitrilysin metallopeptidase 1) degrades post-cleavage mitochondrial transit peptides and BCKDHA (branched chain keto acid dehydrogenase E1 subunit alpha) is part of an inner mitochondrial enzyme complex which produces energy, their up-regulations in IgAN imply un- usual mitochondria activities and require further investigation; APOBEC3G (apolipoprotein B mRNA editing enzyme catalytic subunit 3G) was found to be over-expressed in kidney renal clear cell carcinoma with an unfavorable prognosis and correlate to Treg cells as well as myeloid- derived suppressor cells [46], so it might have similar effect on IgAN; CORO1A (coronin 1A) was found to be a key biomarker in renal interstitial fibrosis, a common final pathway in almost all progressive renal diseases including IgAN [47]. In summary, the expression of a large number of immune genes and some other important genes in critical pathways have been altered in IgAN based on GCEI analysis results, a majority in the selected cluster members were down-regulated and a much smaller number of them were up-regulated, future research on gene clusters and pathways may shed light on IgAN pathogenesis.

Table 5: ROC Results of the top 20 down-regulated and 20 up-regulated genes in the decreas- ing order of Pδ in all clusters. AUC: area under the curve; FPR: false positive rate; TPR: true positive rate; Expr.: regulation directions, Up - IgAN group has higher expression than HC+KD, Down - Otherwise; Pδ = TPR - FPR.

GENE	AUC	Cutoff	Expr.	FPR	TPR	P_δ
ARC	0.64	0.13	Down	31.4	60.95	29.55
KLF10	0.63	0.59	Down	45.89	74.29	28.4
UAP1	0.63	0.41	Down	34.78	62.86	28.08
ARIH1	0.65	0.46	Down	35.75	63.81	28.06
LDLR	0.62	0.25	Down	27.54	55.24	27.7
CEBPB	0.61	0.33	Down	34.3	61.9	27.6
TRIM10	0.6	0.26	Down	35.27	62.86	27.59
RHOB	0.64	0.57	Down	42.51	69.52	27.01
EPHA2	0.59	0.54	Down	43.48	70.48	27
ACADM	0.61	0.49	Down	40.1	66.67	26.57
BEX1	0.6	0.25	Down	30.92	57.14	26.22
NFKBIA	0.6	0.49	Down	39.13	64.76	25.63
TM4SF5	0.58	0.6	Down	42.03	67.62	25.59
PPP1R15A	0.62	0.5	Down	42.51	67.62	25.11
TFEB	0.6	0.2	Down	32.85	57.14	24.29

SENP2	0.62	0.44	Down	37.68	61.9	24.22
RIPK2	0.61	0.46	Down	40.58	64.76	24.18
SIRT1	0.63	0.42	Down	33.33	57.14	23.81
LRRC19	0.59	0.43	Down	37.68	60.95	23.27
OPA1	0.61	0.39	Down	37.68	60.95	23.27
DDX3Y	0.51	0.18	Up	53.62	71.43	17.81
DBN1	0.52	0.4	Up	43.96	60	16.04
CCR5	0.53	0.24	Up	55.56	71.43	15.87
VAX2	0.52	0.92	Up	23.19	37.14	13.95
CXCL9	0.51	0.23	Up	56.04	69.52	13.48
PITRM1	0.57	0.56	Up	39.13	52.38	13.25
CD48	0.55	0.39	Up	43.96	57.14	13.18
GZMB	0.55	0.51	Up	41.55	54.29	12.74
APOBEC3G	0.54	0.75	Up	33.33	45.71	12.38
PRKCB	0.51	0.64	Up	37.2	49.52	12.32
CLEC10A	0.55	0.45	Up	41.06	53.33	12.27
IRX5	0.54	0.24	Up	52.66	64.76	12.1
MMS19	0.53	0.78	Up	31.88	43.81	11.93
TRAF3IP3	0.54	0.49	Up	43.96	55.24	11.28
CORO1A	0.54	0.41	Up	47.83	59.05	11.22
ARHGAP22	0.52	0.45	Up	48.79	60	11.21
PSTPIP1	0.54	0.28	Up	52.66	63.81	11.15
BCKDHA	0.51	0.74	Up	34.78	45.71	10.93
RASGRP2	0.52	0.62	Up	35.75	46.67	10.92
NFATC4	0.52	0.62	Up	39.13	49.52	10.39

Conclusion

The binary GCEI is proposed to indicate whether a cluster of genes is normal (0) or abnormal in terms of member gene expression, and the categorical combinatory cGCEI represents the number of abnormal gene clusters. They are highly correlated to different IgAN risks so that they can be used as novel disease molecular sub-typing tools for future IgA nephropathy management and treatments.

Declaration

Conflict of Interest: Aibing Rao is a co-founder of Shenzhen Luwei (Biomanifold) Biotechnology Limited, Shenzhen, China. All of the others declare no conflict of interest.

Ethics Approval: No ethics approval required due to use of public data resources.

References

1. Habas E, Ali E, Farfar K, et al. IgA nephropathy pathogenesis and therapy: Review & updates. Med 101 (2022): e31219.
2. Selvaskandan H, Shi S, Twaij S, et al. Monitoring Immune Responses in IgA Nephropathy: Biomarkers to Guide Management. Front Immunol 11 (2020): 572754.
3. Reich HN, Tritchler D, Cattran DC, et al. A Molecular Signature of Proteinuria in Glomerulonephritis. PLoS ONE 5 (2010): e13451.
4. Grayson PC, Eddy S, Taroni JN, et al. Vasculitis Clinical Research Consortium, the European Renal cDNA Bank cohort, and the Nephrotic Syndrome Study Network. Metabolic path ways and immunometabolism in rare kidney diseases. Ann Rheum Dis 77 (2018): 1226- 1233.
5. Cox SN, Chiurlia S, Divella C, et al. Formalin-fixed paraffin-embedded renal biopsy tis- sues: an underexploited biospecimen resource for gene expression profiling in IgA nephropathy. Sci Rep 10 (2020): 15164.
6. Berthier CC, Bethunaickan R, Gonzalez-Rivera T, et al. Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. J Immunol 189 (2012): 988-1001.
7. Guo F, Zhang W, Su J, et al. Prediction of Drug Positioning for Quan-Du-Zhong Capsules Against Hypertensive Nephropathy Based on the Robustness of Disease Network. Front Pharmacol 109 (2019): 49.
8. Ju W, Nair V, Smith S, et al. Tissue transcriptome-driven identification of epi- dermal growth factor as a chronic kidney disease biomarker. Sci Transl Med 7 (2015): 316ra193.

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

9. Liu P, Lassen E, Nair V, et al. Transcriptomic and Proteomic Profiling Pro- vides Insight into Mesangial Cell Function in IgA Nephropathy. J Am Soc Nephrol 28 (2017): 2961-2972.

10. Rao A. Gene Cluster Expression Index and Potential Indica- tions for Targeted Ther- apy and Immunotherapy for Lung Cancers. Cancer Screen Prev 1 (2024): 24–35.

11. Hajian-Tilaki K et al. Receiver Operating Characteristic (ROC) Curve Analysis for Med- ical Diagnostic Test Evaluation. Caspian J Intern Med Spring 4 (2013): 627-635.

12. Du, Y, Cheng T, Liu C, et al. IgA Nephropathy: Current Understanding and Perspectives on Pathogenesis and Targeted Treatment. Diagnostics 13 (2023): 303.

13. Rollino C, Vischini G, Coppo R. IgA nephropathy and infections. J Nephrol 29 (2016): 463-468.

14. Ufer F, Vargas P, Engler JB, et al. Arc/Arg3.1 governs inflammatory dendritic cell migration from the skin and thereby controls T cell activation. Sci Immunol 1 (2016).

15. Yang S, Jin S, Xian H, et al. Metabolic enzyme UAP1 mediates IRF3 pyrophosphory- lation to facilitate innate immune response. Mol Cell 83 (2023): 298-313.e8.

16. Wang S, Li Z, Chen Y, et al. ARIH1 inhibits influenza A virus replication and facilitates RIG-I dependent immune signaling by interacting with SQSTM1/p62. Virol J 20 (2023): 58.

17. Liu, X, Cen, X, Wu, R, et al. ARIH1 activates STING-mediated T-cell activation and sensitizes tumors to immune checkpoint blockade. Nat Commun 4066 (2023).

18. Kong L, Sui C, Chen T, et al. The ubiquitin E3 ligase TRIM10 pro- motes STING aggregation and activation in the Golgi apparatus. Cell Rep 42 (2023).

19. Guo M, Cao W, Chen S, et al. TRIM10 binds to IFN-α/β receptor 1 to negatively regulate type I IFN signal transduction. Eur J Immunol 51 (2021): 1762-1773.

20. Chung SS, Ahn BY, Kim M, et al. Control of adipogenesis by the SUMO-specific protease SENP2. Mol Cell Biol 30 (2010): 2135-2146.

21. Yang T, Chiang M, Chang C, et al. SENP2 restrains the generation of pathogenic Th17 cells in mouse models of colitis. Commun Biol 629 (2023).

22. Kolesnichenko M, Mikuda N, H¨opken UE, et al. Transcriptional repression of NFKBIA triggers constitutive IKK- and proteasome-independent p65/RelA activation in senescence. EMBO J 40 (2021): e104296.

23. Sun, SC. The non-canonical NF-κB pathway in immunity and inflammation. Nat Rev Immunol 17 (2017): 545–558.

24. Chai L, Dai L, Che Y, et al. LRRC19, a novel member of the leucine-rich repeat protein family, activates NF-κB and induces expression of proinflammatory cytokines. Biochem Biophys Res Commun 388 (2009): 543-548.

25. Ruefli-Brasse AA, Lee WP, Hurst S, et al. Rip2 participates in Bcl10 signaling and T-cell receptor-mediated NF-kappaB activation. J Biol Chem 279 (2004): 1570-1574.

26. Zhang W, Li X, Wang S, et al. Regulation of TFEB activity and its potential as a therapeutic target against kidney diseases. Cell Death Discov 32 (2020).

27. Subramaniam M, Hawse JR, Rajamannan NM, et al. Functional role of KLF10 in mul- tiple disease processes. Biofactors 36 (2010): 8-18.

28. Zhang Y, Bao S, Wang D, et al. Downregulation of KLF10 contributes to the regeneration of survived renal tubular cells in cisplatin-induced acute kidney injury via ZBTB7A-KLF10-PTEN axis. Cell Death Discov 82 (2023).

29. Park JE, Son AI, Zhou R. Roles of EphA2 in Development and Disease. Genes 4 (2013): 334-357.

30. Zhou, L, Yin, M, Guo, F, et al. Low ACADM expression predicts poor prognosis and suppressive tumor microenvironment in clear cell renal cell carcinoma. Sci Rep 9533 (2024).

31. Vachon L, Smaani A, Tessier N, et al. Downregulation of low-density lipoprotein receptor mRNA in lymphatic endothelial cells impairs lymphatic function through changes in intra- cellular lipids. Theranostics 12 (2022):1440-1458.

32. Shi X, Chen Y, Liu Q, et al. LDLR dysfunction induces LDL accumulation and promotes pulmonary fibrosis. Clin Transl Med 12 (2022): e711.

33. Saikruang W, Ang Yan Ping L, Abe H, et al. The RNA helicase DDX3 promotes IFNB transcription via enhancing IRF-3/p300 holocomplex binding to the IFNB promoter. Sci Rep 12 (2022).

34. P´erez-Mart´ınez M, Gord´on-Alonso M, Cabrero JR, et al. F-actin-binding protein drebrin regulates CXCR4 recruitment to the immune synapse. J Cell Sci 123 (2010): 1160- 1170.

35. Zeng Z, Lan T, Wei Y, et al. CCL5/CCR5 axis in human diseases and related treatments. Genes Dis 9 (2022): 12-27.

36. Krensky A, Ahn YT. Mechanisms of Disease: regulation of RANTES (CCL5) in renal disease. Nat Rev Nephrol 3 (2007): 164–170.

37. Heger L, Balk S, Lu¨hr JJ, et al. CLEC10A is a specific

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

marker for human CD1c+ dendritic cells and enhances their toll-like receptor 7/8-induced cytokine secretion. Front Immunol 9 (2018): 744.

38. Huang L, Song F, Sun H, et al. IRX5 promotes NFκB signalling to increase prolifera- tion, migration and invasion via OPN in tongue squamous cell carcinoma. J Cell Mol Med 22 (2018): 3899-3910.

39. Kaleta B. The role of osteopontin in kidney diseases. Inflamm Res 68 (2019): 93-102.

40. Zhu W, Li J, Zhang R, et al. TRAF3IP3 mediates the recruitment of TRAF3 to MAVS for antiviral innate immunity. EMBO J 38 (2019): e102075.

41. Thoresen D, Wang W, Galls D, et al. The molecular mechanism of RIG-I activation and signaling. Immunol Rev 304 (2021): 154-168.

42. Borgia P, Papa R, D'Alessandro M, et al. Kidney involvement in PSTPIP1 associated inflammatory diseases (PAID): a case report and review of the literature. Front Med 8 (2021): 759092.

43. Yang C, Wu S, Mou Z, et al. Transcriptomic analysis identified ARHGAP family as a novel biomarker associated with tumor-promoting immune infiltration and nanomechan- ical characteristics in bladder cancer. Front Cell Dev Biol 9 (2021): 657219.

44. Malik SA, Modarage K, Goggolidou P. The Role of Wnt signalling in chronic kidney dis- ease (CKD). Genes 11 (2020): 496.

45. Du C, Liu WJ, Yang J, et al. The role of branched-chain amino acids and branched-chain α-keto acid dehydrogenase kinase in metabolic disorders. Front Nutr 9 (2022): 932670.

46. Peng T, Liu B, Lin S, et al. APOBEC3G expression correlates with unfavorable prog- nosis and immune infiltration in kidney renal clear cell carcinoma. Heliyon 8 (2022): e12191.

47. Hu Z, Liu Y, Zhu Y, et al. Identification of key biomarkers and immune infiltration in renal interstitial fibrosis. Ann Transl Med 10 (2022): 190.

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

SUPPLEMENTARY FILES

The supplemental materials include ROC curves of the top 9 genes for each of the remaining clusters, in the decreasing order of Pδ.

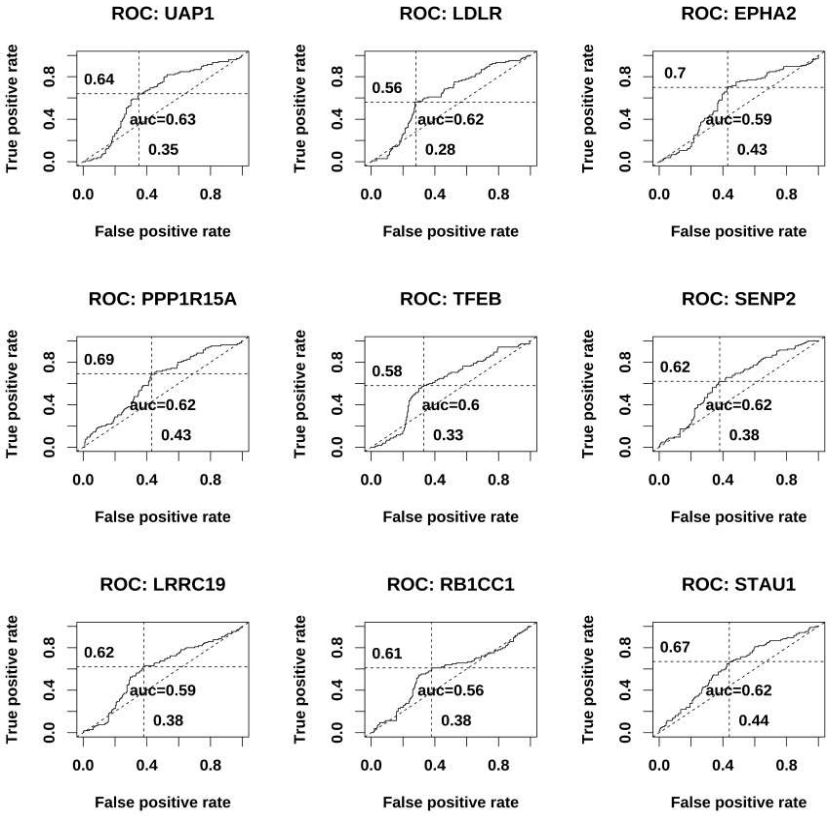

Figure 4: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster INFECTI. Only top 9 genes in the decreasing order of Pδ are displayed.

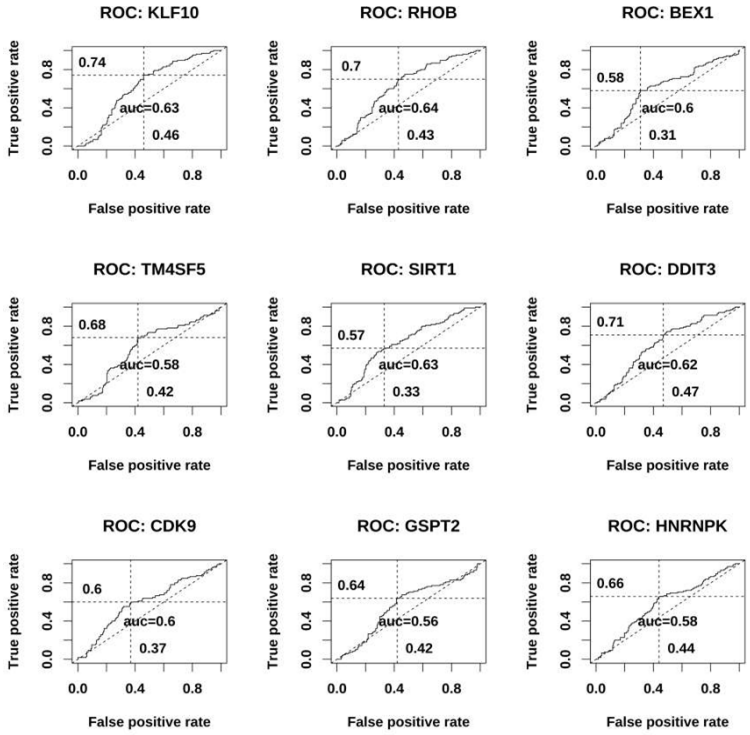

Figure 5: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster CELL CYCLE. Only top 9 genes in the decreasing order of Pδ are displayed.

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

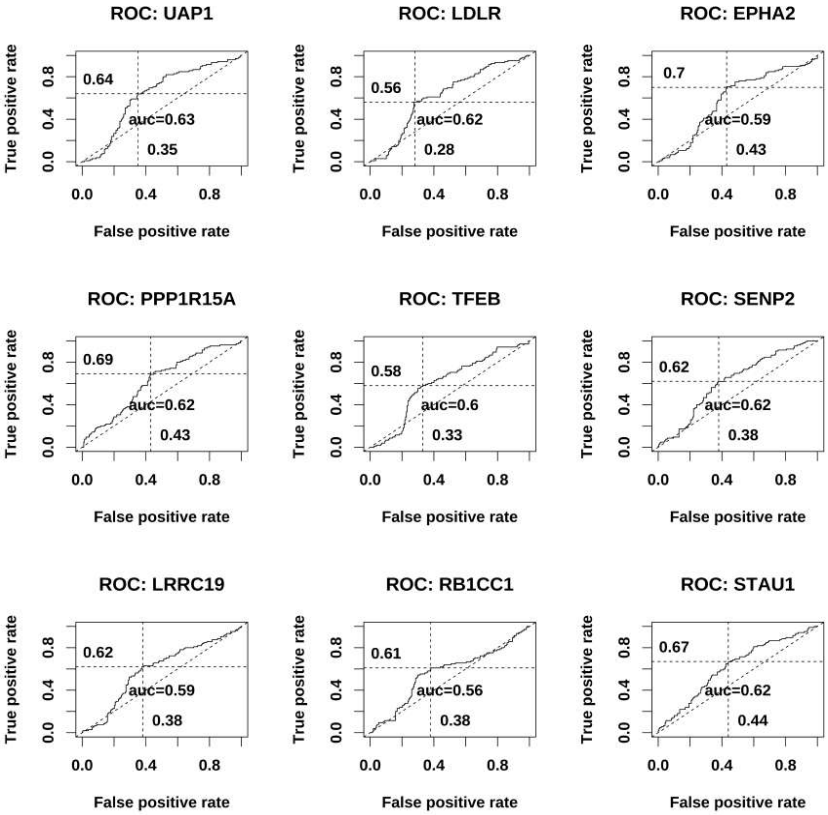

Figure 6: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster CYTOSKELETON. Only top 9 genes in the decreasing order of Pδ are displayed.

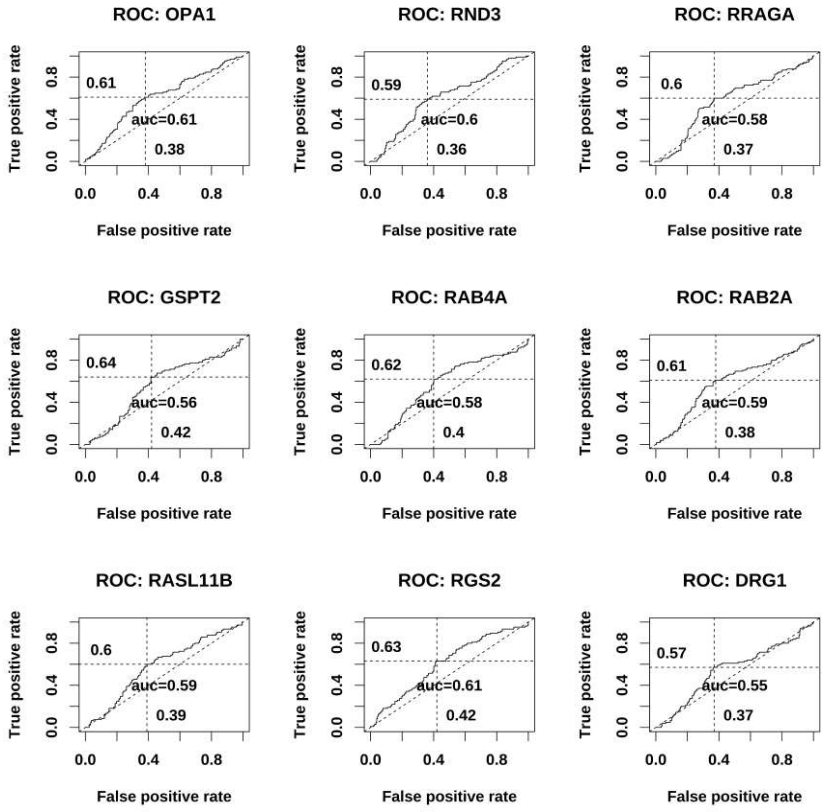

Figure 7: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster DNA REPAIR. Only top 9 genes in the decreasing order of Pδ are displayed.

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

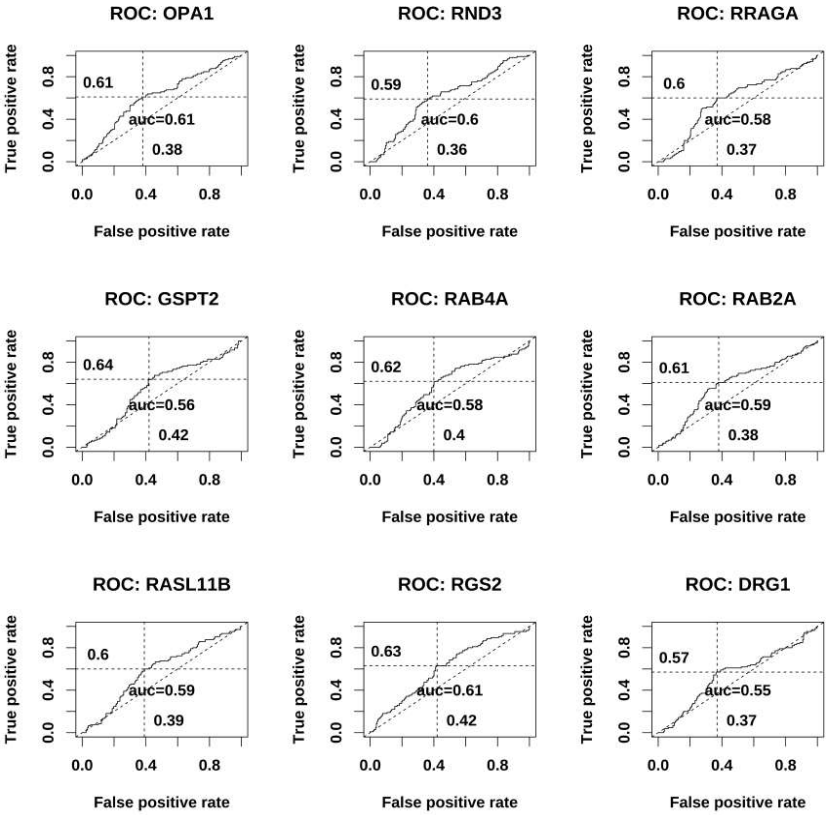

Figure 8: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster GTPASE. Only top 9 genes in the decreasing order of Pδ are displayed.

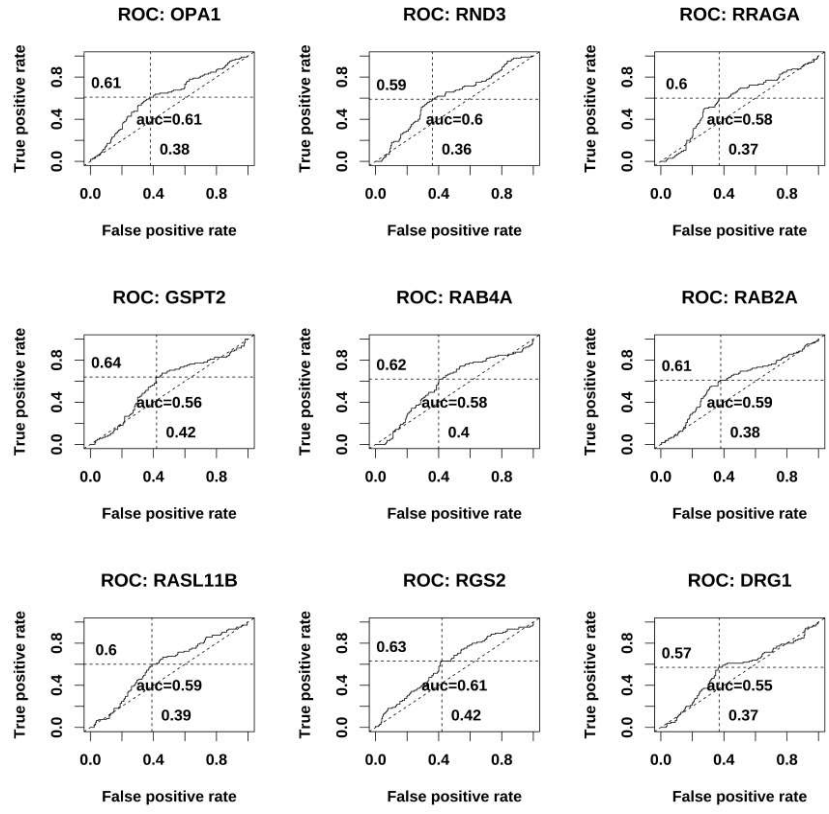

Figure 9: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster MITOCHON. Only top 9 genes in the decreasing order of P_{δ} are displayed.

Rao A, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107091

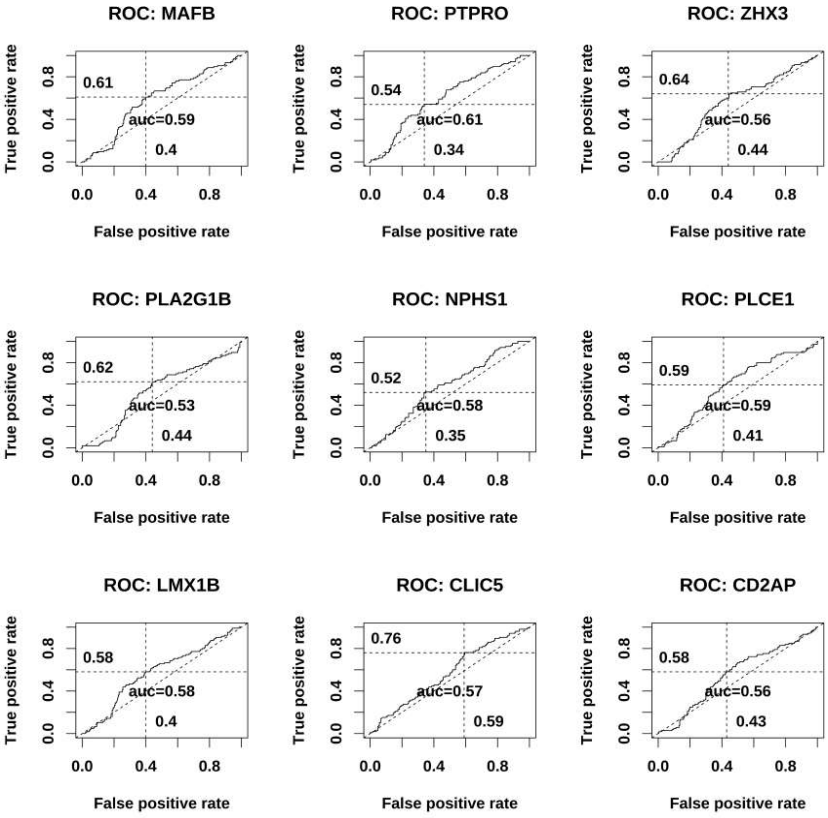

Figure 10: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster PODOCYTE. Only top 9 genes in the decreasing order of Pδ are displayed.

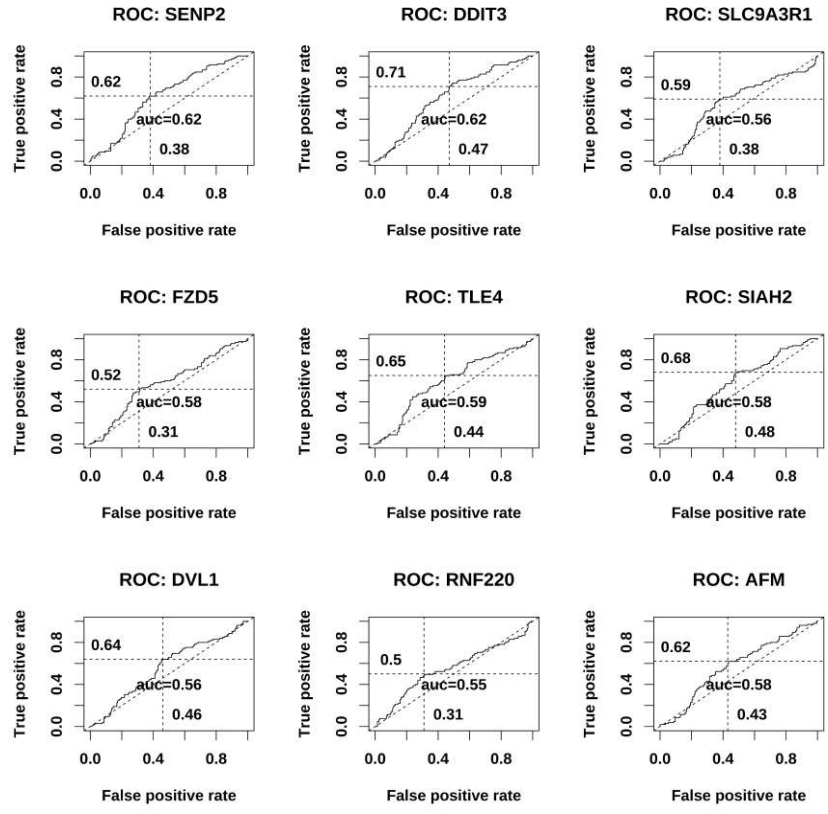

Figure 11: Receiver operating characteristic (ROC) using the gene expression as the predictor of IgAN flags for cluster WNT. Only top 9 genes in the decreasing order of Pδ are displayed.