Construction of Predictive Model of Interstitial Fibrosis and Tubular Atrophy (IFTA) After Kidney Transplantation with Machine Learning Algorithm

Yu Yin  
The First Affiliated Hospital of Nanjing Medical University

Congcong Chen  
The First Affiliated Hospital of Nanjing Medical University

Dong Zhang  
The First Affiliated Hospital of Nanjing Medical University

Qianguang Han  
The First Affiliated Hospital of Nanjing Medical University

Zijie Wang  
The First Affiliated Hospital of Nanjing Medical University

Zhengkai Huang  
The First Affiliated Hospital of Nanjing Medical University

Hao Chen  
The First Affiliated Hospital of Nanjing Medical University

Li Sun  
The First Affiliated Hospital of Nanjing Medical University

Shuang Fei  
The First Affiliated Hospital of Nanjing Medical University

Jun Tao  
The First Affiliated Hospital of Nanjing Medical University

Zhijian Han  
The First Affiliated Hospital of Nanjing Medical University

Ruoyun Tan  
The First Affiliated Hospital of Nanjing Medical University

Min Gu  
The Second Affiliated Hospital of Nanjing Medical University

Xiaobing Ju  
(✉️ doctorjxb73@njmu.edu.cn)  
The First Affiliated Hospital of Nanjing Medical University

Research Article
Abstract

**Background:** Interstitial fibrosis and tubular atrophy (IFTA) are the histopathological manifestations of CKD and one of the causes of long-term renal loss in transplanted kidneys. The purpose of our study is to screen IFTA-related genes with higher importance scores through Random Forest (RF) and further construct IFTA diagnostic model through Artificial Neural Networks (ANNs).

**Methods:** We screened all 162 “kidney transplant” related cohorts in the GEO database and obtained 5 data sets (training sets: GSE98320 validation sets: GSE22459, GSE53605 and GSE76882 survival sets: GSE21374). Differentially expressed genes (DEGs) analysis, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Random Forest (RF), Artificial Neural Network (ANN), Unsupervised Clustering analysis, CIBERSORT analysis were used to analyze the data.

**Results:** A total of 108 common DEGs were identified by taking the intersection of the DEGs of our training sets and validation sets. A total of 15 top IFTA-specific DEGs were screened through the RF, then was used to build ANNs models. The model has good performance in both the training sets [GSE98320 (AUC = 0.9560)] and the validation sets [GSE22459 (AUC = 0.720), GSE53605 (AUC =0.938), GSE76882 (AUC = 0.781)], indicating that we have avoided overfitting while improving the accuracy. Furthermore, samples of survival sets are divided into two clusters using consensus clustering algorithm basing on the expression of 15 top IFTA-specific DEGs. We found significant differences between the two subgroups by survival analysis, and further enrichment analysis and immune cell infiltration analysis were conducted to further explore the causes of survival differences.

**Conclusion:** we identified key biomarkers of IFTA and developed a new IFTA classification model, basing on the combination of RF and ANNs.

1. Introduction

Among 100,800 solid organ transplants worldwide, 69,400 (62.5%) were renal transplants (1), which were the most efficacious therapies for patients with uremia due to a variety of causes (2). The incidence of survival of patients with transplanted kidneys have improved with the development of post-transplantation management, but 40 percent of grafts fail within 10 years of transplantation (3). According to traditional theory, allogeneic immune responses cause irreversible kidney damage, resulting in graft loss (4, 5). The dominating histopathological findings in the end-stage of transplant renal include interstitial fibrosis, tubular atrophy (IFTA) and glomerulosclerosis (6), which is the major cause of graft loss due to progressive loss of function (7). To determine the prognosis of transplanted kidneys, pathogenesis and prediction models of IFTA are essential to understand.

Over the past few years, the technological advancements of gene sequencing and statistical analysis on a large scale have been applied and developed in genetic diagnosis and analysis. These advancements have enabled a deeper studying of rejection after kidney transplantation in greater depth. With the immunogenetic approach to kidney transplantation, Dou et al. have developed a prognostic model, which
is a good predictor of kidney graft survival at 1 and 3 years (8). Zhang et al. discovered 2 novel prognostic genes associated with ischemia-reperfusion injury, using a total of 1,000 specimens were collected from 11 independent cohorts (9). Gui et al. discovered that renal graft fibrosis, which results in chronic renal graft dysfunction, is negatively regulated by ATG16L (10).

In previous study of IFTA diagnosis based on mRNA expression, the quantity of samples and the method of filtering variables are limited, resulting in relatively low sensitivity and specificity (11). With the progress of machine learning algorithms, new approaches can be employed to diagnose diseases. The key benefits of Random Forest (RF) are its accuracy and resistance to overfitting, which makes it a good choices of machine learning algorithms (12). The use of Artificial Neural Networks (ANNs) allows for the creation of nonlinear models, and make it is possible to detect nonlinear relationships and all potential interactions among predictor variables (13). It has been reported that RF and ANN can be used in conjunction to make efficient diagnoses in a wide range of diseases, such as Alzheimer’s disease, heart failure and periodontitis (12, 14, 15).

The purpose of this study is to combine RF and ANN to develop a diagnostic model constructed by gene expression utilizing the expression data of the transplanted kidney in Gene Expression Omnibus (GEO) database, including GSE22459, GSE53605, GSE76882 and GSE98320. Data sets 1 through 3 were used as validation sets, and data sets 4 was used as training sets. Then, we used GSE21374 to further explore the influence of genes used to construct a diagnostic model of IFTA on long-term graft loss.

2. Materials and Methods

Data Retrieval and Organization

We performed a comprehensive search on GEO official website for expression matrix and annotated clinical information of patients who underwent transplantation of the kidney. The screening criteria for 162 "kidney transplant" cohorts in the GEO: 1. All probes in each sample have a value greater than 0; 2. Each patient's sample with publicly available gene expression profile contained biopsy-confirmed IFTA information or survival information about long-term graft loss 3. The total number of cohort samples ≥50. Finally, we selected 5 data sets (GSE22459, GSE53605, GSE76882, GSE98320 and GSE21374) to be included in this study (16-19). As part of the data processing, original data matrix is downloaded, probe annotation is performed, and low-abundance genes in most samples are removed. Each data set contains samples classified as IFTA and non-IFTA, and the detailed description about groups is shown in Table 1.

DEGs Analysis

Through the "limma" package in R (version 4.1.1) (20), we extract the differentially expressed genes (DEGs) in 4 datasets, with adjusted P - values = 0.05 as the filter criteria, then take the intersection of the DEGs of our 4 cohorts for next analysis. DEGs in 4 groups were visualized using the “ggplot2” R packages. For the purpose of exploring DEGs interaction in IFTA of kidney transplantation, the STRING
Functional Enrichment Analysis

In order to research the eligible DEGs of IFTA commonality in transplanted kidney, we extracted the common upregulation and upregulation DEGs of four cohorts and took the intersection of them respectively. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed on common up-regulated and down-regulated DEGs respectively using the “clusterProfiler” R package (21).

Random Forest Screening

For further IFTA-related DEG screening, we used the R package "randomForest". In the first step, error rates were calculated using 1–500 trees. A comprehensive evaluation was conducted in order to determine the optimal tree number for constructing Random Forest (RF) model, with low error rates and high stability. Gini coefficient method was used to calculate the dimensional significance value for screening the important genes as the candidate genes for IFTA diagnosis. The IFTA candidate genes for ANN model development was defined from the top 15 DEGs with significance value greater than 2, which is a acceptable screening index in RF and has been used in similar studies (12, 15).

Artificial Neural Network

As the first step, a normalization of the input data was achieved by converting the 15 candidate DEGs into "Gene Scores", using the min–max method, in which we made a comparison between the median of all sample expression values and the expression value of a single gene in a certain sample. If the expression level of an upregulated gene was higher than the median value, or the expression level of a downregulated gene was lower than the median value, its expression will be valued as 1; otherwise, it will be given 0. The R package "neuralnet" was used to calculate the gene weight and create an ANN model based on the Gene Score (22). The ANN model consists of 3 layers, including hidden, input and output layer. Five hidden layers were set, and two nodes (control/treat) were set in the output layer. Using the "pROC" R package (23), we calculated the area under the receiver operating characteristic (ROC) curve of the training set to evaluate the model's performance. Additionally, the model was validated with other patient cohorts (GSE22459, GSE53605, GSE76882,). Principal component analyses (PCA) were performed and visualized using the “limma” and the “ggplot2” R packages in validation set.

Unsupervised Clustering of Kidney Transplant Patients

In order to identify potential molecular subtypes of transplanted kidneys based on 15 key biomarkers that constructed the ANN model, an unsupervised cluster analysis was performed on the prognosis set using the "ConsensusClusterPlus" R package, with resampling set as 1000. Classification stability was ensured by using the k-means algorithm with 1,000 iterations and an 80% resampling rate. A Kaplan-Meier (K-M)
survival curve was used to determine the survival differences among subtypes of kidney transplant patients.

**CIBERSORT Analysis**

CIBERSORT, a deconvolution algorithm that quantifies cell types based on gene matrix, was used to quantify 22 kinds of immune cell infiltration, and then a map of immune cell infiltration is drawn to determine difference of two subtypes by unsupervised clustering, using the R package "ggpubr".

### 3. Results

**Identification of DEGs in IFTA**

Using the "limma" R package, a total of 108 common DEGs were identified after examining four datasets. There was a significant difference between the IFTA and non-FITA groups in these genes, as shown in Figure 1A-D.

**Functional enrichment**

After setting the filtering conditions and taking the intersection of 4 gene sets, the 108 common DEGs were used to identify a protein interaction network with 82 nodes and 246 relationship pairs (Figure 1E). A total of 21 up-regulated genes and 17 down-regulated genes were extracted from the four gene sets (Figure 2A, B). According to GO analysis, the up-regulated DEGs are mainly enriched in immune-related functions, such as immune system process and immune response (Figure 2C), and down-regulated DEGs mainly regulate oxidation-reduction-related functions, such as oxidation-reduction process, oxidoreductase activity and cellular respiration (Figure 2D). Next, KEGG enrichment analysis was also performed on these DEGs. Up-regulated DEGs enriched in immune-related pathways, such as PD-L1 expression, Th1 and Th2 cell differentiation and Th17 cell differentiation (Figure 2E). Down-regulated DEGs enriched in pathways of cellular REDOX reactions, such as peroxisome and carbon metabolism (Figure 2F).

**Diagnostic biomarker screening**

RF model was built using DEG gene scores to identify reliable diagnostic biomarkers of IFTA. Figure 3A shows the correlation between the number of RF trees and the model error, the best ntree value (ntree = 291) has the lowest error rate. As shown in Figure 3B, the top 30 genes from MeanDecreaseGini are displayed. ZNHIT2 was the most important biomarker, followed by ISOC1, PEX1, SUCLA2 and SLC25A11. Then, a total of 15 top IFTA-specific DEGs with importance greater than 2 were identified for further analysis.

**The ANN Model Construction and Validation**
An ANN analysis was conducted to majorizing the weight of Gene Score transformed by IFTA-specific DEGs. In all, there were 15 input variables, five hidden variables, and two output variables in the ANN diagnostic model (Figure 3C). The detailed description of gene weight can be found in Supplementary Table 1. Using the "pROC" R package, we evaluated the model performance, obtaining an AUC of 0.956 in the training cohort, indicating excellent classification accuracy (Figure 3D). Three independent datasets (GSE22459, GSE53605, GSE76882) were used to test the neural-IFTA model’s ability to predict IFTA occurrence. Figure 4 A-C show that there is a distinction between the IFTA group and the non-IFTA group, as well as differences within the groups. The AUC values of the validation cohorts were 0.720, 0.938, and 0.781. (Figure 4 D-F).

Subgroup Analysis of IFTA

In order to explore the roles of biomarkers in kidney transplantation prognosis, prognosis sets were classified using consensus clustering. The number of iterations was set at 1000 times to ensure stability of classification categories. The results suggested that, when the number of clusters (k) was 2, the samples in the consensus profiles obtained an optimal allocation (Figure 5A, B). The DEGs between the two subgroups, as shown in Figure 5C. According to the K-M survival curve, there were distinct differences in survival between the two subgroups of prognosis sets (Figure 5D). The 22 immune cell scores for GSE21374 samples are shown in Supplementary Table 2 based on CIBERSORT. As shown in Figure 5E, the proportions of plasma cells, CD8 T cell, follicular helper T cell, Tregs, resting Dendritic cells in the cluster A were significantly higher than cluster B, while the proportions of gamma delta T cell and monocyte are opposite. The differentially expressed genes (DEGs) of 2 clusters are extracted using adjusted P-values =0.05 and logFoldChange =1 as filter criteria. According to GO analysis, the DEGs are mainly involved in immune-related functions, such as immune system process, immune response and defense response (Figure 5F). KEGG enrichment analysis shows DEGs enriched in pathways of intracellular signal transduction, such as chemokine signaling pathway and cytokine-cytokine receptor interaction (Figure 5G).

4. Discussion

IFTA is not only commonly occurring histopathological manifestations of CKD, but it also causes long-term renal failure in transplanted kidneys (6, 7). As a result of chronic fibrosis of the transplanted kidney, IFTA develops in the early stages after transplantation, eventually causing renal failure (24). Although the diagnostic model for IFTA can clearly identify it at an early stage and it is helpful to judge the prognosis of transplanted kidneys, the detection method of the process is rare. In this study, we applied the RF and ANN to analyze the target data set in the GEO database, established an IFTA prediction model by analyzing and finding the differential genes in the puncture sample.

From GEO dataset, 108 DEGs were identified between IFTA and non-IFTA samples. According to gene enrichment analysis, the DEGs involved in immune-related pathways were mostly upregulated. There was a significant enrichment in the expression of PD-L1 and differentiation of Th1 and Th2. PD-1
upregulation in CD4 + T cells has been linked to pulmonary fibrosis in previous study (25), however, in renal transplantation fibrosis there has been little accurate study of the PD-1-related pathway. Then the DEGs that were downregulated were mostly involved in oxidation-reduction-related pathways. According to Kang HM and colleagues, a key role in kidney fibrosis development is believed to be defective fatty acid oxidation (26). Thus, the combination of PD-L1 pathway and oxidation-reduction-related pathways may be a novel diagnosis target for IFTA.

A well-known reason for IFTA is ischemia/reperfusion injury during transplantation and subsequent inflammatory response after transplantation. During ischemia/reperfusion injury, generation and production of reactive oxygen species (ROS), dysfunction of mitochondria, and activation of heparinase lead to inflammation and fibrosis, induces epithelial to mesenchymal transition (27). A combination of inflammation interactions resulted in the infiltration of fibroblasts, which promotes the formation of extracellular matrix and irreversible fibrosis and ultimately leads to the loss of renal function (28). With the help of the RF classifier, 108 potential periodontitis markers were screened and 15 key biomarkers were identified, based on MeanDecreaseGini. The 15 key biomarkers are primarily associated with the immune system and the REDOX process, including ZNHIT2, ISOC1, PEX1, SUCLA2, SLC25A11, DLAT, ESRRA, IFITM3, CASP3, KIAA0895(MATCAP2), EMP3, HSPA6, CD33, DHRS11, CD4. Based on the results of RF, only a small number of core genes (CD4, CD33) are associated with the immune process, while most are associated with REDOX. PEX1 and ISOC1 belong to the Peroxisome-related pathway (29, 30). SUCLA2, SLC25A11 and DLAT are involved in tricarboxylic acid cycle of carbon metabolism pathways (31–33). Based on the results of RF and the enrichment analysis of common down-regulated genes, it can be speculated that tissue hypoxia and ischemia are characteristic changes of IFTA, and provide potential targets for future anti-fibrosis therapy. Then, two distinct molecular subgroups of periodontitis were identified through unsupervised clustering base on the 15 key IFTA-related biomarkers. In the survival analysis, there was a visible distinction between the two clusters in terms of K-M survival curve. Therefore, IFTA-related core genes may also be useful for predicting the long-term loss of transplanted kidney. According to DEGs enrichment analysis of two clusters, immunotherapy may also be a potential treatment for fibrosis that reduces graft loss.

One of the reasons we chose RF is that it can automatically explore the interaction among multiple variables and find the variables that matter more when dealing with high-dimensional data. Furthermore, RF can avoid overfitting by generating a large number of decision trees at random. As a result, RF has demonstrated satisfactory accuracy both in modeling and validation groups in previous studies (34). Compared with the traditional linear diagnostic model, although ANN has shortcomings such as overfitting, time-consuming due to large amount of computation, and difficult interpretation of output results, this statistical model based on simulating brain learning does play a unique advantage in complex data patterns such as genes matrix (35, 36). As we use both RF and ANN simultaneously, we hope to improve the model's accuracy without overfitting by combining the advantages of both (37). In our diagnostic model, we firstly combining RF with ANN to enhance the predictive capability of the IFTA model. Furthermore, Gene scoring was used to eliminate batch effects in four data sets, which enabled our prediction model to show good predictive power in both the modeling and 3 validation groups. Our
study shows that our model has a predictability (AUC = 0.956) superior to another IFTA forecasting model integrated COX regression model with LASSO constructed by Yang et al (AUC = 0.8210) (11). Moreover, the AUC of the predictive model achieved 0.938 in the validation set GSE53605, 0.781 in set GSE76882, and 0.720 in set GSE22459, which suggested that our model has a high level of accuracy without overfitting.

Even so, several limitations also existed in this study. In the first place, transplanted kidney biopsy tissue samples were used as model input data, and acquiring tissues is challenging in clinical practice. Additionally, further experimental studies are required to explain the correlation between certain biomarkers and IFTA mechanistically. Third, further exploration of clusters A and B was hampered by a lack of clinical information. Finally, we screened the entire transplant kidney cohorts in the GEO database to build a diagnostic model. However, to build models, machine learning algorithms require a larger number of samples. More independent kidney transplant patient cohorts should be used to evaluate and elevate our model’s performance. In order to verify patient tissue samples, we will collect samples from our hospital.

5. Conclusion

In summary, basing on the combination of RF and ANNs, we developed a new IFTA classification model. Furthermore, we test the model using 3 independent datasets from GEO. Compared to existing biomarkers, our study proves to be a superior diagnostic strategy for clinicians.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Publicly available datasets were analyzed in this study. These data can be found here: https://www.ncbi.nlm.nih.gov/geo/ (training sets: GSE98320 validation sets: GSE22459, GSE53605 and GSE76882 survival sets: GSE21374)

Competing interests

The authors have no competing interests to declare.

Funding
This work was supported by the National Natural Science Foundation of China [grant numbers 82170769, 82070769, 81900684, 81870512, 81770751], Project of Jiangsu Province for Important Medical Talent [grant number ZDRCA2016025], the “333 High Level Talents Project” in Jiangsu Province [grant numbers BRA2017532, BRA2016514], the Standardized Diagnosis and Treatment Research Program of Key Diseases in Jiangsu Province [grant number BE2016791], the Open Project Program of Health Department of Jiangsu Province [grant number JSY-2-2016-099], Jiangsu Province Natural Science Foundation Program [grant number BK20191063].

Authors' contributions

YY, CC, and ZD contributed equally to this work, YY analyzed and interpreted the data and drafted the manuscript. CC and ZD interpreted the data and contributed to the substantial revisions of the manuscript. QH, ZW, ZH, HC, LS, JT, ZH and SF helped to perform the statistical analysis and interpret the data. RT, MG, and XJ made a contribution to the conception and design, analyzed and interpreted the data, supervised the study, provided the

Acknowledgements

We acknowledge the open databases of GEO.

References


### Table

#### Table 1

<table>
<thead>
<tr>
<th>GEO no</th>
<th>Platform</th>
<th>Species</th>
<th>Tissues</th>
<th>No-IFTA</th>
<th>IFTA</th>
<th>Total</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE22459</td>
<td>GPL570</td>
<td>Homo sapiens</td>
<td>kidney biopsy</td>
<td>25</td>
<td>40</td>
<td>65</td>
<td>training set</td>
</tr>
<tr>
<td>GSE53605</td>
<td>GPL571</td>
<td>Homo sapiens</td>
<td>kidney biopsy</td>
<td>45</td>
<td>10</td>
<td>55</td>
<td>validation set</td>
</tr>
<tr>
<td>GSE76882</td>
<td>GPL13158</td>
<td>Homo sapiens</td>
<td>kidney biopsy</td>
<td>139</td>
<td>135</td>
<td>274</td>
<td></td>
</tr>
<tr>
<td>GSE98320</td>
<td>GPL15207</td>
<td>Homo sapiens</td>
<td>kidney biopsy</td>
<td>974</td>
<td>234</td>
<td>1208</td>
<td></td>
</tr>
<tr>
<td>GSE21374</td>
<td>GPL570</td>
<td>Homo sapiens</td>
<td>kidney biopsy</td>
<td>282</td>
<td></td>
<td></td>
<td>prognosis set</td>
</tr>
</tbody>
</table>

### Figures
Figure 1

Differential gene expression and Protein interaction network

Figure 2

Functional Enrichment Analysis of common DEGs

Notes: (A, B): Common up-regulated DEGs (A) and down-regulated DEGs (B) of the four gene sets. (C, D): GO analysis of the common up-regulated DEGs (C) and down-regulated DEGs (D). (E, F): KEGG analysis of the common up-regulated DEGs (E) and down-regulated DEGs (F).
Figure 3

Identification of candidate DEGs by RF and development of an ANN diagnostic model

Notes: (A): Decision trees with different numbers of random trees result in changes in error rates. The error rate is lowest when there are 291 decision trees. (B): Ranking of the top 30 DEGs based on importance scores calculated by the Gini coefficient method. Genes are shown on the y-axis and its importance index on the x-axis. (C): The visualization of the artificial neural network (D): ROC curves of the training sets.
Figure 4

Validation of ANN diagnostic model

Notes: (A, B, C): Principal component analyses (PCA) of GSE22459(A), GSE53605(B), GSE76882(C). (D, E, F): ROC curves of the GSE22459(D), GSE53605(E), GSE76882(F).
Figure 5

Identification of molecular subgroups in IFTA.

Notes: (A): Consensus clustering matrix when $k = 2$. (B): Relative variation of the area under the CDF region at $k = 2 - 9$. (C): Volcanic maps of DEG for 2 subgroups, and the genes that constructed the ANNs model were labeled. (D): Kaplan-Meier curve of 2 subgroups. (E): The differences in infiltrated immune cells and functions. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. (F): GO analysis of the DEGs of 2 subgroups. (G): KEGG analysis of the DEGs of 2 subgroups.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx