BMC Cardiovascular Disorders

RESEARCH **Open Access**

Exploring the predictive values of SERP4 and FRZB in dilated cardiomyopathy based on an integrated analysis

Bin Qi[1], Hai-Yan Wang[1], Xiao Ma[1], Yu-Feng Chi[1] and Chun Gui[1*]

Abstract

Background and objective The aim of this study was to investigate potential hub genes for dilated cardiomyopathy (DCM).

Methods Five DCM-related microarray datasets were downloaded from the Gene Expression Omnibus (GEO). Differentially expressed genes (DEGs) were used for identification. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment, disease ontology, gene ontology annotation and protein-protein interaction (PPI) network analysis were then performed, while a random forest was constructed to explore central genes. Artificial neural networks were used to compare with known genes and to develop new diagnostic models. 240 population blood samples were collected and expression of hub genes was verified in these samples using RT-PCR and demonstrated by Nomogram.

Results After differential analysis, 33 genes were statistically significant (adjusted $P < 0.05$). Functional enrichment of these differential genes resulted in 85 Gene Ontology (GO) functions identified and 6 pathways enriched for the KEGG pathway. PPI networks and molecular complex assays identified 10 hub genes (adjusted $P < 0.05$). Random forest identified *SMOC2* and *SFRP4* as the most important, followed by *FCER1G* and *FRZB*. NeuraHF models (*SMOC2*, *SFRP4*, *FCER1G* and *FRZB*) were selected by artificial neural network model and had better diagnostic efficacy for the onset of DCM, compared with the traditional KG-DCM models (*MYH7*, *ACTC1*, *TTN* and *LMNA*). Finally, *SFRP4* and *FRZB* were expressed higher in DCM verified by RT-PCR and as a factor for DCM identified by Nomogram.

Conclusions We performed an integrated analysis and identified *SFRP4* and *FRZB* as a new factor for DCM. But the exact mechanism still needs further experimental verification.

Keywords GEO analysis, Dilated cardiomyopathy, Artificial neural network, Random forest

*Correspondence:
Chun Gui
guichun@yahoo.com
[1]Department of Cardiology, First Affiliated Hospital, Guangxi Medical University, 6 Shuangyong Road, Nanning, Guangxi 530021, China

Introduction

Dilated cardiomyopathy (DCM) is a specific type of cardiomyopathy that is caused by enlargement of the heart chambers for various reasons and is a common cause of heart failure and death [1].

In adults, dilated cardiomyopathy arises more commonly in men than in women. The prevalence is one in 2500 individuals, with an incidence of seven per 100 000 per year and the familial type might account for 20–48% of all cases [2].

DCM is an impairment of the dilated and contractile function of the left ventricle. Patients with DCM usually have a poor prognosis, with a 5-year survival rate of less than 50% after the initial diagnosis. Patients with DCM often die of congestive heart failure, which is also one of the common causes of heart transplantation [3]. Other clinical manifestations of DCM include arrhythmias, thromboembolism and sudden death [3].

The commonly used clinical diagnostic techniques for DCM have several limitations. The positive detection of AHA in immunomarkers reflects the presence of autoimmune damage in patients and is commonly seen in patients with VMC and its progression to DCM. Brain natriuretic peptide/N-terminal natriuretic peptide levels are often indicative of heart failure, and the etiology of heart failure is not only DCM. Echocardiogram is another kind of commonly used technology that evaluates cardiac function, the individual that it relies on expert more is operated skilled degree and diagnostic experience, make check repeatability is poor. The early diagnosis rate of dilated cardiomyopathy is low and the long-term prognosis is poor. Therefore, there is an urgent need for new diagnostic models to be developed to complement the existing diagnostic inefficiencies.

Technological advances have improved the accuracy of diagnosis, and the precise identification of disease-causing genes by sequencing has provided a solid theoretical foundation and technical support for the development of diagnostic and prognostic models, while the new diagnostic model of DCM can also give sequencing data for in-depth mining [4]. In this study, differentially expressed genes (DEGs) in DCM and normal myocardium were screened from the Comprehensive Gene Expression Database (GEO). We used gene enrichment analysis to interact with protein-protein interaction (PPI) networks to understand the role of these DEGs. We used random forest algorithm to identify key genes expressed in DCM. Then we input these key genes into artificial neural network to construct the genetic diagnosis model of DCM. At the same time, a diagnostic model was constructed using genes that had been proved to be significantly different in the patients of DCM. Finally, we compared the sensitivity and specificity between the two diagnostic models.

Materials and methods

Data selection and preprocessing

In total, five datasets were downloaded for analysis GSE42955 [5], GSE79962 [6], GSE120895 [7], GSE9800 and GSE17800 [8]) (GEO, http://www.ncbi.nlm.nih.gov/geo/). GSE42955 was based on the GPL6244 platform of Affymetrix Human Gene 1.0 ST Array and included 12 DCM patients and 5 controls collected from human myocardial biopsy tissues [5]. GSE79662 was also based on the GPL6244 platform of Affymetrix Human Gene 1.0 ST Array and included 9 DCM patients and 11 controls collected from human myocardial biopsy tissues [6]. GSE42955 and GSE79962 were derived from the same tissue source and based on the same sequencing platform. To obtain a large sample of microarray data set, we combined two sample microarray datasets (GSE42955 and GSE79962) as a training dataset. GSE120895 was based on the GPL570 platform of Affymetrix Human Genome U133 Plus 2.0 Array and included 47 DCM patients and 8 controls collected from human myocardial biopsy tissues [7]. Because GSE120895 contains a larger sample size of DCM patients, we chose it as the validation dataset. These datasets were converted to logarithmic form after standardization, and the R package ComBat was used to remove the batch effects [9]. A training dataset with 37 samples and a validation dataset with 55 samples were obtained using classical and Bayesian correction methods. GES9800 and GSE17800 as the validation dataset, the analysis method is similar to the previous dataset.

Differentially expressed genes (DEGs) screening

After we combined two sample microarray datasets (GSE42955 and GSE79962) as a training dataset, data preprocessing and expression of genes were processed using R software (version 4.1.1). DEG was defined as adjusted P-values needing to be less than 0.05 and ($\log_2$FC| > 1). Data processing methods and visualization of graphs were performed with the help of the "Limma" and "Pheatmap" packages in the R [10].

Function analyzed for DEGs

Gene ontology (GO), disease ontology (DO) enrichment and the Kyoto Encyclopedia of Genes and Genomes (KEGG) were used to illustrate the pathways of the DEGs. We employed DOSE and the clusterProfiler package in R [11, 12]. An adjusted P value (Q-value) of <0.05 was regarded as statistically significant. Please refer to our previous studies for details of our analysis methodology [13].

Protein–protein interaction (PPI) network and potential key gene analyses

The potential interactions between DEGs were studied with the help of the Interacting Genes/Proteins

(STRING) plugin. The PPI network was visualized using Cytoscape software (version 3.8.1) [14]. We used the Molecular Complex Detection (MCODE) app to detect the major and most notable clustering modules. For further subsequent analysis, we set EASE≤0.05 and count≥2 as the cutoff value and MCODE score>8 as the threshold. We have explained the use of this method in detail in a previous manuscript [15].

Random forest (RF) classification

During data processing, DEG needs to be classified, and we use R-package RandomForest for this process [16]. In the first step, we need to find the optimal number of all variables. 4 is the optimal number of variables for the binary tree in the whole node, and 131 is also the optimal number of trees in the random forest, which minimizes the error. Second, the decreasing accuracy method (Gini coefficient method) is used to construct the random forest model and to obtain the dimensional importance values in the model. Based on the importance, the first 4 genes greater than 1 were selected as disease-specific genes for subsequent model construction. Next, the four important genes in the training dataset were selected for unsupervised hierarchical clustering and reclassified, and heat maps were drawn using the Pheatmap package.

Neural network to build disease classification model

The training model of the neural network is selected as the training data set. All the data are normalized and the maximum and minimum values are distinguished from the weight. To construct the artificial neural network model for the variables, we employed the R package for neural networks to perform the analysis. The DCM classification model was constructed by first obtaining the gene weight information, and this process set 5 hidden layers as model parameters. The disease classification score in the model is defined as the sum of the product of the weight scores multiplied by the expression levels of important genes. To ensure accuracy, the model results are then subjected to 5-fold cross-validation, and the results of the 5-fold cross-validation are calculated using a confusion matrix function, which is done using the Caret software package [17]. Validation of the AUC classification performance is calculated using the pROC software package [18].

Additional data verification

The validity of the classification and scoring model of DCM disease and normal samples was verified on the independent dataset GSE120895. The efficiency of the different classifications is calibrated using the area under the ROC curve, and its calculation and visualization of the graphs are done using the pROC software package. The effectiveness of classification was then compared with that of 4 other reported DCM disease biomarkers. At the same time, the optimal threshold of ROC curve and the sensitivity and specificity classification threshold of disease and normal samples under this threshold were calculated.

Study population

In total, 240 patients were recruited from the inpatient department at the First Affiliated Hospital, Guangxi Medical University from 2019-6-1 to 2020-12-31. Ischemic cardiomyopathy due to coronary stenosis was excluded after completion of coronary angiography, and all admitted patients met the diagnostic criteria for dilated cardiomyopathy: (1) Evidence of ventricular enlargement and reduced myocardial contractile function. LVEDd>5.0 cm (women) and LVEDd>5.5 cm (men) (or greater than 117% of the predicted value for age and body surface area, i.e., 2 times the SD of the predicted value+5%); LVEF<45% (Simpsons' method), and LVFS<25%; (2) Excluding cardiac valvular disease, congenital heart disease, or ischemic heart disease [19]. Exclusion criteria included subjects with poor compliance, incomplete clinical data, contrast agent sensitivity and autoimmune diseases. Additionally, subjects with obvious surgical contraindications were excluded. Clinical data collection, biochemical measurements and diagnostic criteria were performed according to internationally standardized methods, following a common protocol. The study adhered to the Declaration of Helsinki of 1975 (http://www.wma.net/en/30publications/10policies/b3/) and its revision in 2008 and the Ethics Committee of First Affiliated Hospital, Guangxi Medical University agreed with the study design (No: Lunshen-2019-KY; Feb. 02, 2019). Informed consent was obtained from all subjects after receiving a full explanation of the study.

RNA isolation, reverse transcription (RT) and quantitative PCR (qPCR)

Fasting blood samples (5 mL) were collected in EDTA and separated by centrifugation at 3000 g for 15 min. The RNA was extracted according to the manufacturer's protocols. Total RNA was eluted in 30 μL of RNase-free water. RNA was reverse transcribed to cDNA with reverse transcriptase kit. The reaction system contained total RNA 2 μg, Enzyme Mix 2 μL, RNase-free water up to 20 μL. The reaction contained 0.2 μL PCR Forward Primer, 0.4 μL PCR Reverse Primer, 3.0 μL cDNA, RNase-free water up to 20 μL. All reactions were run in duplicate. The average of the Ct value was calculated after the PCRs were run in duplicate for each sample. The specific primer design for the validation genes is carried out under the guidance of a professional company, as detailed in the Supplementary Table 1.

Statistical analysis

The statistical packages SPSS 22.0 (SPSS Inc., Chicago, IL, USA) and R software (version 4.1.1) were used for statistical analysis. Quantitative variables were expressed as mean ± standard deviation, and two sample means were compared using the T-test, or if the samples did not conform to a normal distribution, the interquartile spacing was used. The difference in the percentages were compared using the Chi-square test. The predictive accuracy of the risk model was assessed by the discrimination measured by the C statistic and the calibration evaluated by the Hosmer-Lemeshow χ2 statistic. All tests were two-sided, and $P<0.05$ was considered statistically significant.

Results

Differential expression analysis

See Fig. 1 for a flow chart of the analysis of the manuscript. After normalization (Supplementary Table 2), data from 21 DCM patients and 16 control samples were retrospectively analyzed from the conjoint analysis of GSE42955 and GSE79962. After the analysis, 33 DEGs were obtained: 15 genes were significantly upregulated and 18 genes were significantly downregulated ($P<0.05$ and $|\log_2FC| > 1$). The volcanic plot of gene average expression levels was shown in Fig. 2A. The heat map of the screened 33 DEGs in the conjoint dataset was shown in Fig. 2B.

Functional analysis of DEGs

The clusterProfiler software package was used for GO enrichment analysis of the important DEGs obtained in the first step. The results of GO enrichment analysis represent three parts, including biological processes, cellular components, and molecular functions (Fig. 3A). Among these results, relevant biological processes involved in DCM include cell chemotaxis, receptor-mediated endocytosis, leukocyte chemotaxis; relevant cellular components involved in DCM include collagen- containing, extracellular matrix vacuolar lumen, blood microparticle; relevant molecular functions involved in DCM include extracellular matrix structural constituent conferring compression resistance, extracellular matrix structural constituent, growth factor activity. Figure 3B intensively shows part of the GO enriched terms and the significant DEGs involved. KEGG enrichment analysis indicated that the main DEGs enrichment was in the pathway of Phagosome, Complement and coagulation cascades, Ae-RAe signaling pathway in diabetic complications (Fig. 3C). Figure 3D intensively shows part of the KEGG pathways which the significant DEGs involved in.

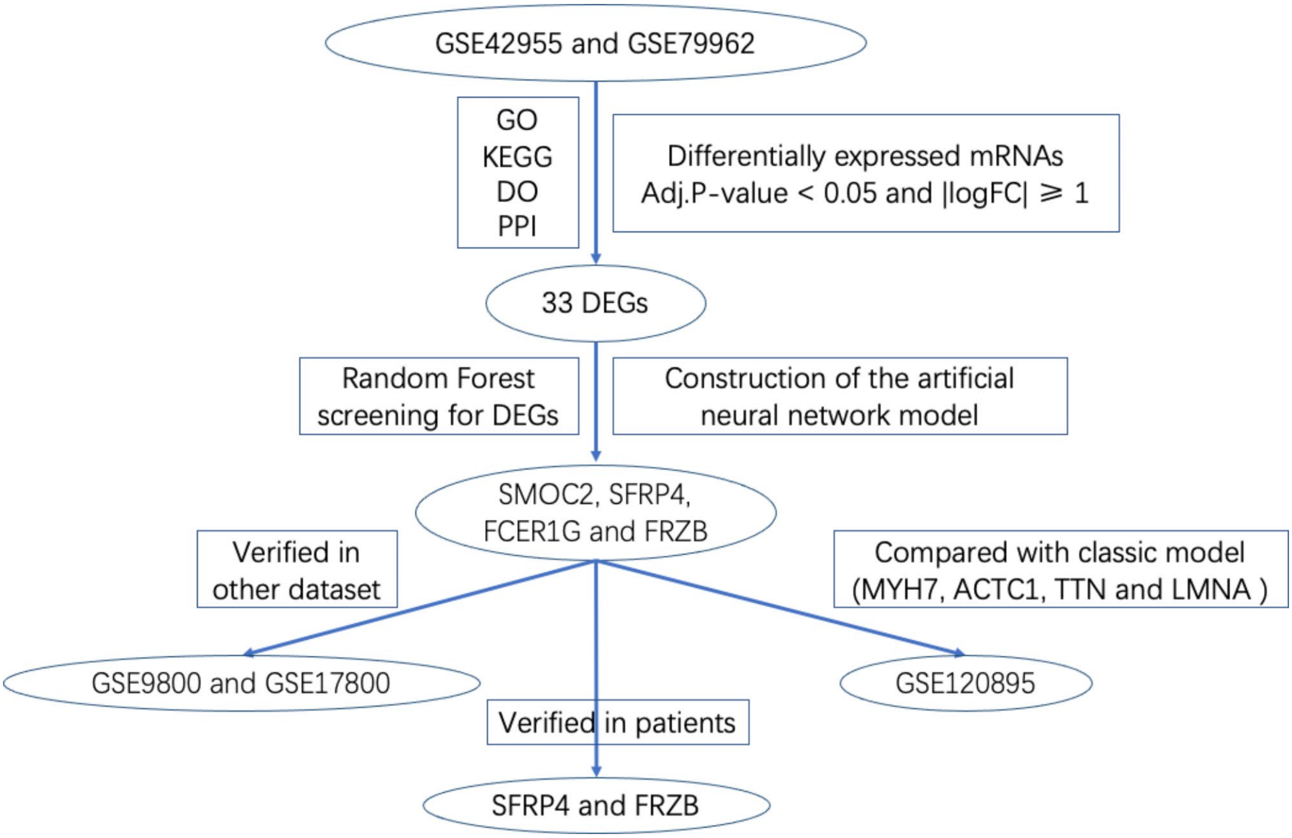
GSE42955 and GSE79962
GO
KEGG
DO
PPI
Differentially expressed mRNAs
Adj.P-value < 0.05 and |logFC| ⩾ 1
33 DEGs
Random Forest screening for DEGs
Construction of the artificial neural network model
SMOC2, SFRP4, FCER1G and FRZB
Verified in other dataset
Compared with classic model (MYH7, ACTC1, TTN and LMNA)
GSE9800 and GSE17800
GSE120895
Verified in patients
SFRP4 and FRZB

Fig. 1 The flowchart for analysis

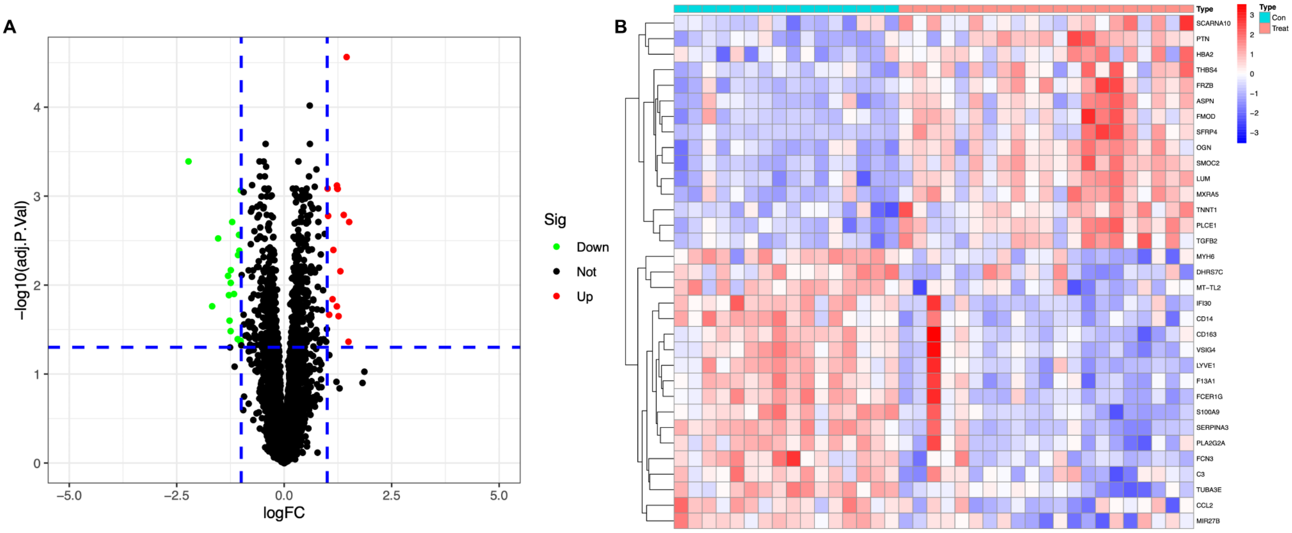

Fig. 2 DEG analysis and Heatmap. (**A**): Volcano plot for DEGs. Items with statistical significance and upregulated expression are marked with red dots, and downregulated expression is marked with green. (**B**): Heatmap for DEGs. The red strip represents high relative expression, and the blue strip represents low relative expression

Detailed data for GO and KEGG were visible in Supplementary Table 3, Supplementary Table 4. Finally, we used the STRING database to develop the PPI network of these DEGs (Fig. 3E). Detected by MCODE methods that 10 genes with the higher degree and MOCDE_score (Supplementary Table 5).

Random forest screening for DEGs

Next, we took the DEGs into the random forest classifier. The lowest error rate occurred when the number of variables was 4; meanwhile, the optimal number of trees in DCM classifier was set to 131 due to the low error rate and stability (Fig. 4A). In the process of building the random forest model, the variable importance of the output results was measured from the Angle of reducing accuracy and mean square error (Gini coefficient method) (Fig. 4B). Four DEGs of greater than 1 importance (Gini coefficient method) were identified as candidate genes for subsequent analysis. Among the four variables, *SMOC2* and *SFRP4* are the most important, followed by *FCER1G* and *FRZB*. In the 37 samples in the combined dataset of GSE42955 and GSE79962, these 4 genes could be used to distinguish between disease and normal samples (Fig. 4C). Among them, the expression level of *FCER1G* was high in normal samples. On the other hand, *SMOC2*, *SFRP4* and *FRZB* were highly expressed in disease samples.

Construction of the artificial neural network model

Based on the combined datasets of GSE42955 and GSE79962, an artificial neural network model based on neural network package was constructed. This dataset has previously been preprocessed to normalize the data. Before starting the calculation, the maximum and minimum values are normalized and the number of hidden layers is set to 5. There are no fixed rules on how many layers and neurons to use when choosing parameters. It is generally believed that the number of neurons should take into account the size of input layer and output layer. Since we selected four key genes to build the model through the random forest tree in the previous step, we set the parameter of the number of neurons as 4. The combined dataset neural network topology of GSE42955 and GSE79962 shows 4 input layers, 5 hidden layers, and 2 output layers in total (Fig. 5A). In order to evaluate the results of the neural network model more effectively, we chose a five-fold cross-validation method (Supplementary Table 6). In order to remove the batch effect of selected key genes in different samples, we scored the genes if the gene was expressed in the sample was greater than the median expression value of the gene in all samples, then the gene was marked with 1 score; otherwise, it was marked with 0 score, and the results of four genes were shown in Supplementary Table 7. The purpose of gene scoring was to calculate its weight in the neural network. Classification efficiency of model scores constructed using gene expression and gene weights. The formula for calculating the classification score of the disease neural network model is as follows: neuraHF = (Gene Expression x Neural Network Weight). The weight of each gene (input layer) reaching the hidden node was detailed in Supplementary Table 8. The weight of each hidden node to the sample attribute (output layer) was detailed in Supplementary Table 9. Correspondingly, we used the same method to construct another neural network model of four identified key genes of DCM (*MYH7*,

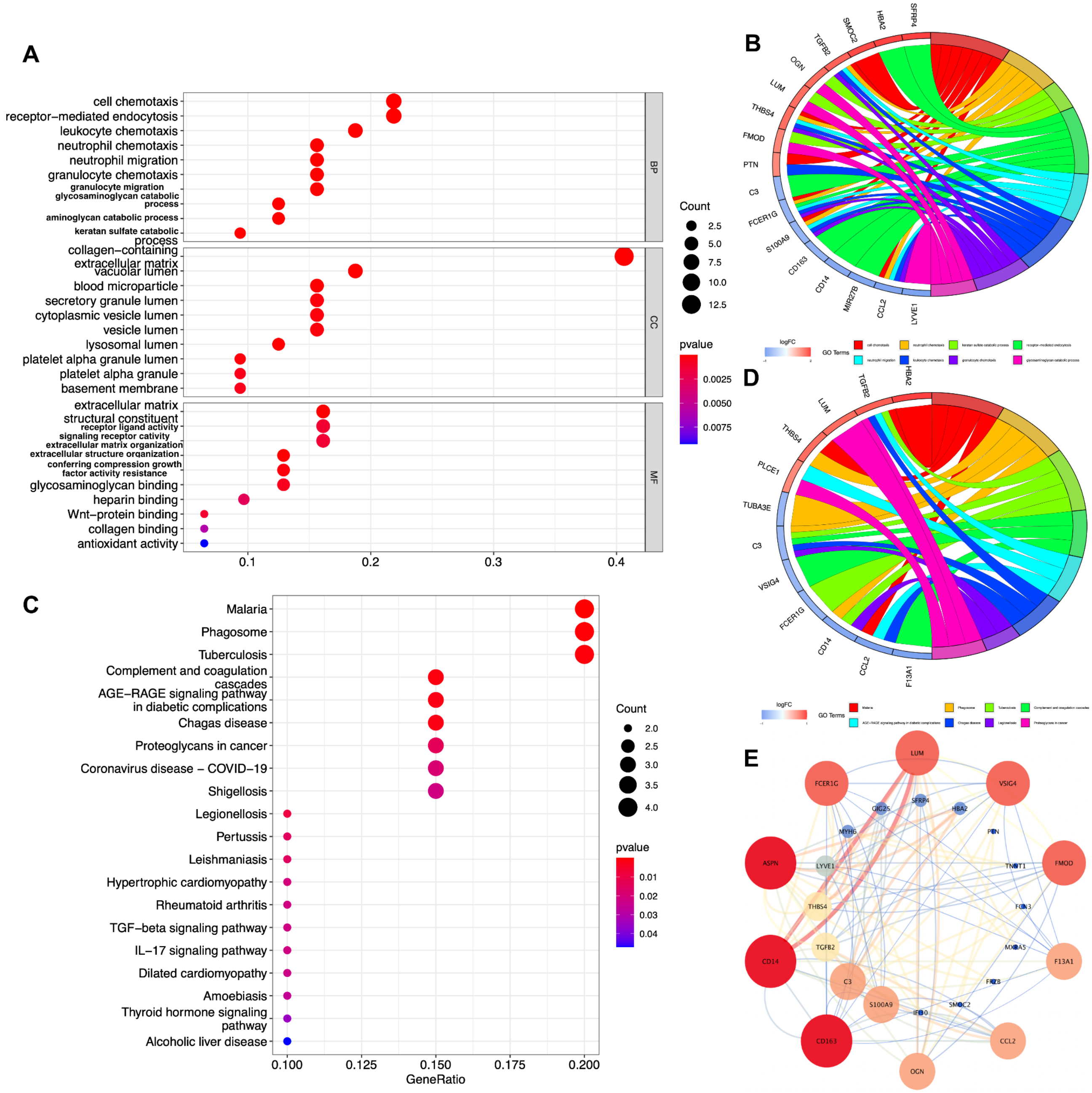

Fig. 3 Functional annotation and PPI network. (**A**): GO analysis for DEGs. (**B**): Ribbons linking the genes with their assigned terms for GO analysis. The logFC is represented by the blue-to-red coding near the marked genes. (**C**): KEGG analysis for DEGs. (**D**): Ribbons linking the genes with their assigned terms for KEGG analysis. The logFC is represented by the blue-to-red coding near the marked genes. (**E**): PPI network of the selected DEGs. Edge stands for the interaction between two genes. The darker is the edge, the darker is the node. A degree was used to describe the importance of protein nodes in the network; darker filling shows a high degree, and white represents a low degree. The significant modules were identified from the PPI network using the molecular complex detection method with a score of > 8.0

ACTC1, *TTN* and *LMNA*) (Fig. 5B), we called this neural network model the KG-DCM.

Validation of artificial neural network model

First, we verified the accuracy of the neural network model(neuralDCM) in the combined datasets of GSE42955 and GSE79962. The accuracy of Control group was 0.938 and the accuracy of treat group was 0.952. The AUC of neuralDCM was 0.975(95%CI 0.921-1.000) (Fig. 6A). The neural network model could distinguish the control group from the treatment group accurately. Then we verified the accuracy of the neural network model (KG-DCM) in the combined datasets of GSE42955 and GSE79962. The accuracy of Control group was 0.562

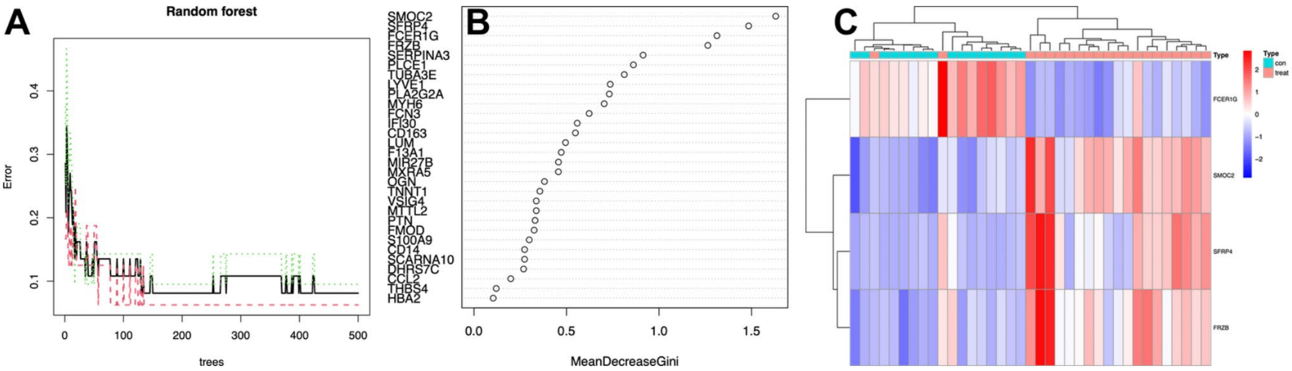

Fig. 4 Random Forest model. (**A**): Relationship between the number of decision trees and the model error. The x-axis represents the number of decision trees, and the y-axis represents the error rate of the constructed model. (**B**): The importance of all variables in the random forest classifier through the Gini coefficient method. The x-axis represents the mean decrease in the Gini index, and the y-axis represents all variables. (**C**): Expression of key genes in disease group and normal group

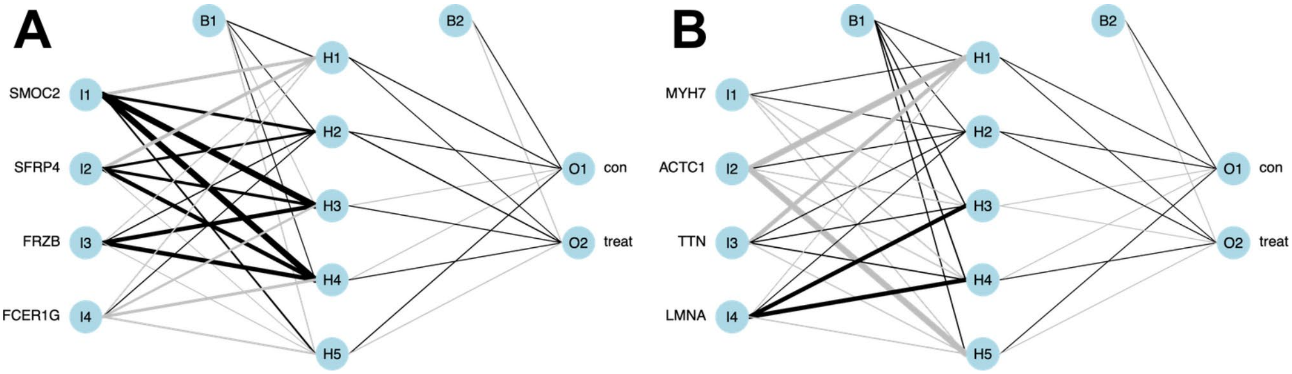

Fig. 5 Artificial neural network model. (**A**): Results of neural network visualization for DEGs. (**B**): Results of neural network visualization for four genes that have been confirmed to be associated with DCM

and the accuracy of treat group was 0.952. The AUC of KG-DCM was 0.789(95%CI 0.616–0.938) (Fig. 6B). An independent validation dataset GSE120895 was used to evaluate the classification efficiency of neuralDCM and KG-DCM and compare their AUC values through maximum and minimum standardized data processing. The accuracy of Control group was 0.875, the accuracy of treat group was 0.660 and the AUC was 0.818(95%CI 0.660–0.932) through the neuralDCM model in the dataset GSE120895 (Fig. 6C). The accuracy of Control group was 0.375, the accuracy of treat group was 0.681 and the AUC was 0.609(95%CI 0.396–0.816) through the KG-DCM model in the dataset GSE120895 (Fig. 6D).

Validation

We selected two microarrays (GSE9800 and GSE17800) with dilated cardiomyopathy to validate the screened hub genes. As shown in Fig. 7 (A-D), four hub genes (*SMOC2*, *SFRP4*, *FCER1G* and *FRZB*) were compared in Normal and DCM in GSE9800, but only *SFRP4* and *FRZB* with significance ($P<0.05$). The same situation could be found in GSE17800 Fig. 7 (E-H). Then, to further verify the function of the two hub genes (*SFRP4* and *FRZB)*, we collected the data of some patients hospitalized, measured the relative gene expression after collecting the peripheral blood and isolation of T cells, and compared the gene expression level between control and DCM. Table 1 shows the general situation of 240 patients with gender and age matching. We considered all the variable data, including the relative expression of two hub genes (*SFRP4* and *FRZB*), age, sex, BMI, blood pressure, serum glucose, lipid profile, renal function, blood uric acid, cardiac enzyme profile, left ventricular diastolic and end-systolic diameters, cardiac ejection fraction, which were the best subset of risk factors to related to DCM risk score and risk model (nomogram)The level of BMI, DBP, LDL-C, Creatinine and LVEDd were higher in DCM. The opposite trend was reflected in the level of SBP, HDL-C, HR, LVEDs and EF where DCM was lower than in the normal group (Fig. 8C). We defined the sores as follows: male=1; female=2. The nomogram had excellent discriminative power with a C-statistic and was well calibrated with the Hosmer-Lemeshow χ^{2} statistic. The predicted probabilities of developing DCM ranged from 0.0002 to 99.6%. After calculation, the relative expression of levels of *SFRP4* and *FRZB*, Sex, Heart Beat, Creatinine,

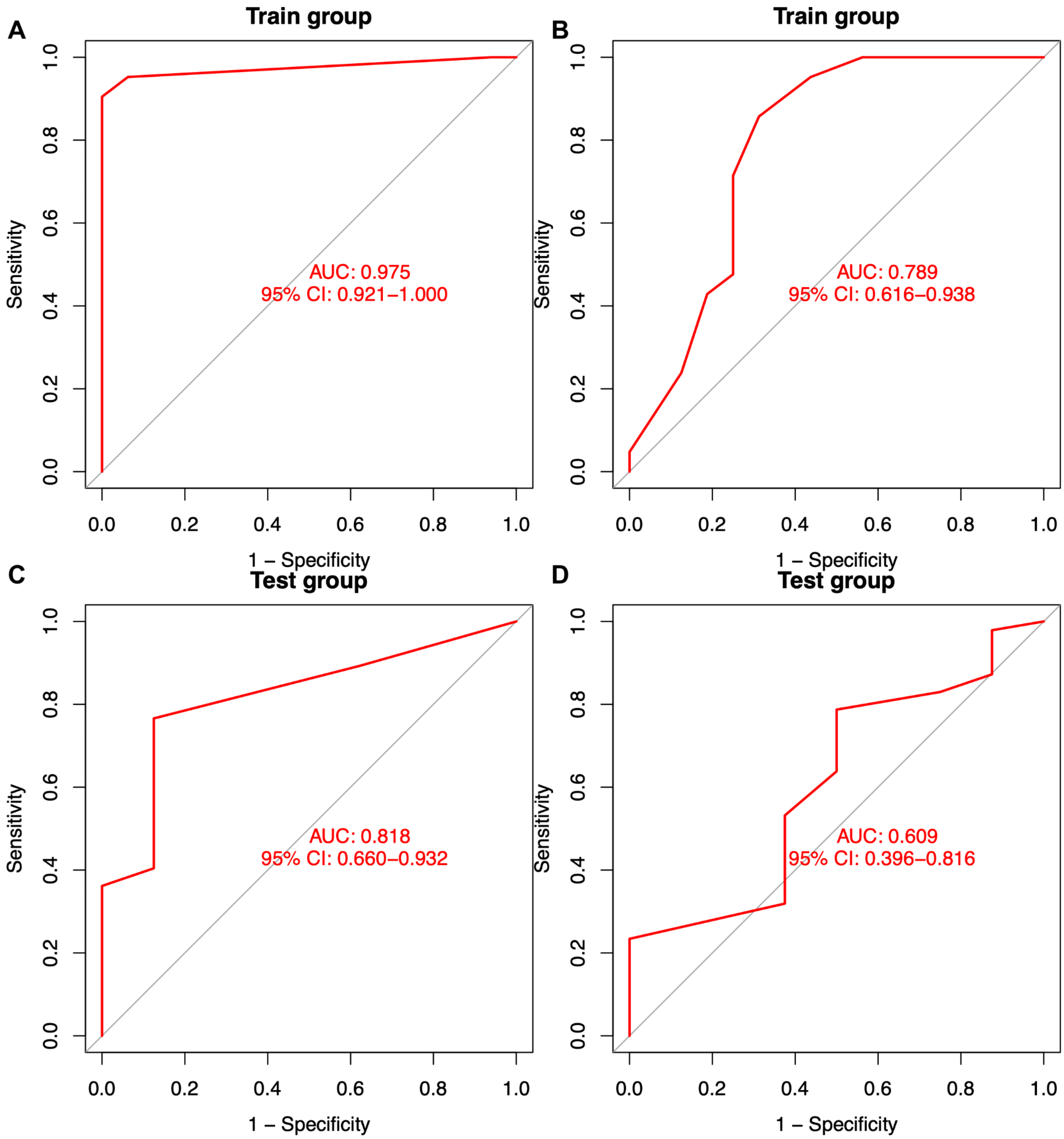

Fig. 6 ROC curve analyses for the diagnostic credibility. (**A** and **C**): Diagnostic model constructed by DEGs. (**B** and **D**): Diagnostic model constructed by four genes that have been confirmed to be associated with DCM

CKMB, LVEDd, LVEDs and EF were significantly related to the risk of DCM, with statistical significance. The relative expression of peripheral blood RT–PCR showed that the expression of *SFRP4* and *FRZB* was statistically significant in the comparison of cases and controls (Fig. 8A and B).

Discussion

In recent years, machine learning algorithms have gradually become a trendy and hot topic, with their powerful computational analysis capabilities to analyze disease diagnosis or prognostic models to inform treatment. In turn, the expression data of genes in public databases can be used as analytical variables to provide a sound analytical basis for biomarkers as diagnostic or prognostic

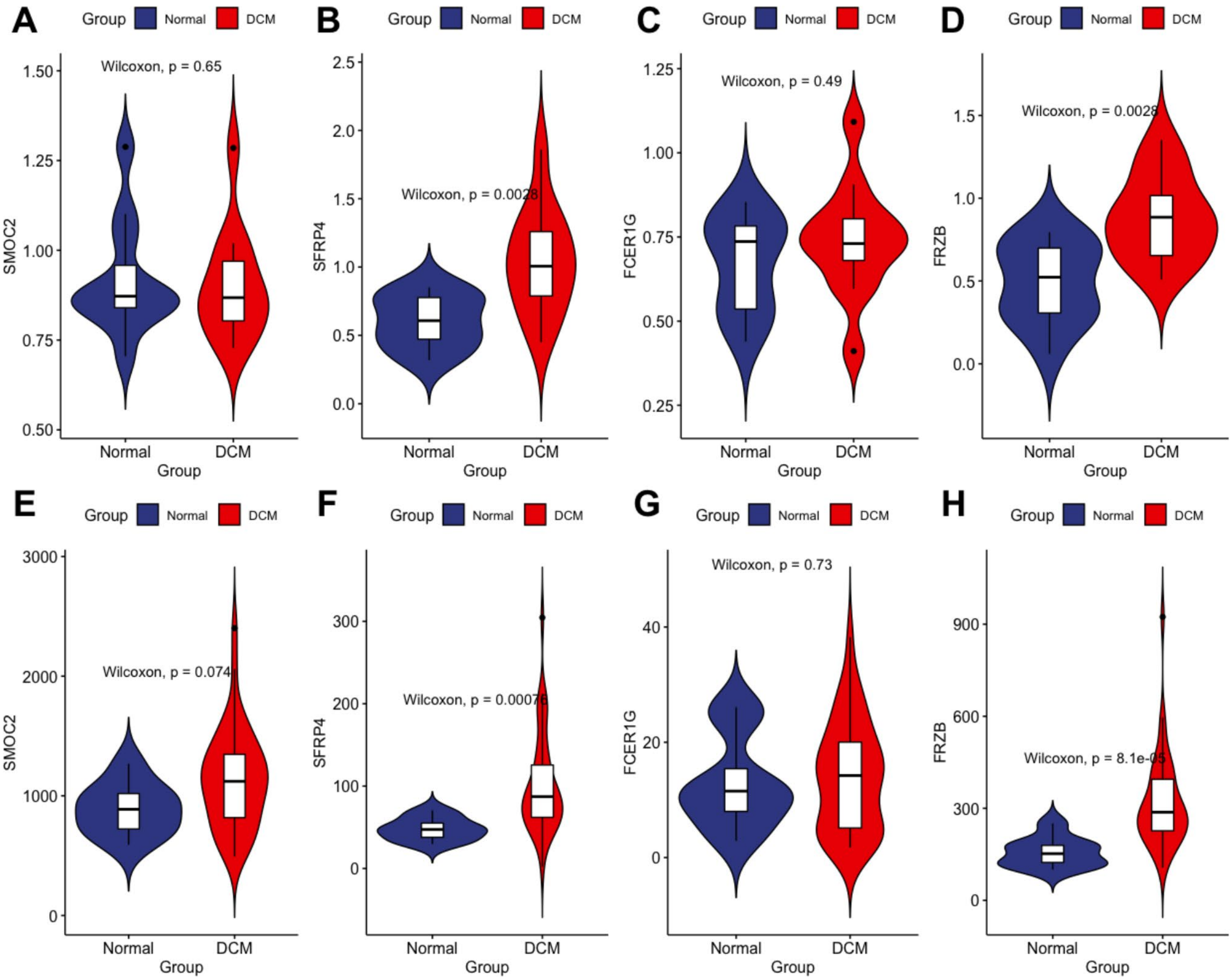

Fig. 7 Expression of four hub genes in different microarrays. (**A** - **D**): Expression in GSE9800. (**E** - **H**): Expression in GSE17800

of diseases [20–23]. As an exclusive diagnosis of DCM, there is often a lack of direct and effective diagnosis. As a result, DCM is often not recognized or treated early, and by the time it is diagnosed, patients often have obvious symptoms of heart failure [24, 25]. Early detection and diagnosis of dilated cardiomyopathy is of great importance to the long-term prognosis of patients. In this study, we calculated the DCM-related DEGs for the first time, and obtain four important candidate DEGs by random forest classifier. Neural network model was used to determine the predictive weight of related genes, and the classification model neuraDCM was constructed. At the same time, we constructed a classification model KG-DCM consisting of four genes that have been identified to be closely related to DCM. We evaluated the classification efficiency of the model neuraHF and the model KG-DCM in combined samples (GSE42955 and GSE79962) and independent samples (GSE120895) respectively. AUC of neuralDCM is highly efficient, and neuralDCM has a better classification efficiency than KG-DCM which composed of DCM key biomarkers. We aimed to develop a diagnostic model based on gene expression data using as many samples as possible from GEO database. We used a new approach [neuraHF = (Gene Expression x Neural Network Weight)] to redefine the importance of genes, and although it was not fully argued, we proved our conclusions in subsequent genetic validation. Ultimately, in other microarrays (GSE9800 and GSE17800) as well as in population sample validation, we identified two hub genes (*SFRP4* and *FRZB)* whose high expression was strongly associated with the development of DCM.

To make the DEGs more accurate, we used two microarrays (GSE42955 and GSE79962) from the same tissue sample source and the same sequencing platform for joint analysis, and at the same time, batch effect was removed to reduce the error. Gene enrichment analysis suggested that the function of DEGs were related to cell chemotaxis, receptor-mediated endocytosis, leukocyte chemotaxis, phagosome and complement. The above enrichment analysis suggested that the differential genes

Table 1 Comparison of demographic, lifestyle characteristics and serum lipid levels among different groups

Parameter	Normal	DCM
Number	88	152
Male/female	26/62	42/110
Age (years)	52.02 ± 8.21	52.38 ± 8.56
Height (cm)	169.71 ± 8.25	168.93 ± 8.03
Weight (kg)	54.20 ± 5.85	53.69 ± 6.79
Body mass index (kg/m^2)	28.66 ± 6.54	29.81 ± 6.39^a
SBP (mmHg)	107.38 ± 13.57	106.55 ± 15.28^a
DBP (mmHg)	77.69 ± 10.54	79.62 ± 9.91^a
TC (mmol/L)	4.97 ± 1.35	5.28 ± 1.23
TG (mmol/L)^1	1.58(0.39)	1.72(0.64)
HDL-C (mmol/L)	1.62 ± 0.55	1.57 ± 0.49^a
LDL-C (mmol/L)	2.97 ± 0.83	3.01 ± 0.74^a
Heart Beat (times/minutes)	77.62 ± 10.72	76.88 ± 11.31^c
Creatinine, (μmol/L)	75.69 ± 12.08	77.35 ± 11.57^a
Troponin T, (μg/L)	0.02 ± 0.01	0.03 ± 0.01
CK, (U/L)	88.62 ± 44.22	86.57 ± 42.21
CKMB, (U/L)	14.68 ± 2.57	13.75 ± 3.87
LVDd(mm)	49.81 ± 19.61	51.67 ± 18.23^c
LVDs(mm)	33.59 ± 11.92	30.12 ± 13.05^c
EF (%)	69.56 ± 14.27	67.32 ± 13.25^c

DCM, dilated cardiomyopathy; *SBP*, systolic blood pressure; *DBP*, diastolic blood pressure; *TC*, total cholesterol; *TG*, triglyceride; *HDL-C*, high-density lipoprotein cholesterol; *LDL-C*, low-density lipoprotein cholesterol; *LVDd*, Left ventricular end diastolic dimension. *LVDs*, left ventricular end-systolic dimension. *EF*, ejection fraction.^{1}Because of not normally distributed, the value of triglyceride was presented as median (interquartile range), the difference between the two groups was determined by the Wilcoxon-Mann-Whitney test. The P value was defined as the comparison of case and control groups. $^a P < 0.05$; $^b P < 0.01$; $^c P < 0.001$

were concentrated in cellular immunity and inflammatory response. Based on MeanDecreaseGini, the first 4 hub genes screened by RF model were included in DEGs classification. We used these four genes to construct an artificial neural network model. Then we used different types of microarrays to calculate the weight of hub genes using artificial neural network. Another diagnostic model KG-DCM was created based on the composition of four known pathogenic genes for DCM. By comparing with KG-DCM model, we found that the neuralDCM model constructed by us was superior to the model constructed by known DCM pathogenic genes in both sensitivity and specificity. The results of AUC scores also showed that our model obtained high AUC scores, indicating that it can separate DCM samples from normal samples with good probability in microarray data. However, as the method has only been validated in our experiments, it is yet to be supported by cohort studies with large samples.

SFRP4 (secreted frizzled related protein 4) is a member of the SFRP family, it contains cysteine-rich domains homologous to the WNT binding site of crimped proteins. SFRPs act as soluble regulators of Wnt signals. Expression of *SFRP4* in ventricular myocytes and expression of apoptosis-related genes [26]. Serum levels of *SFRP4*, an adipocytokine, are significantly elevated in patients with different types of diabetes [27, 28]. Meanwhile circulating *SFRP4* levels were positively correlated with glucose, insulin and hba1c levels [29]. Serum *SFRP4* levels in patients with stable coronary artery disease (CAD) are also positively correlated with body mass index, waist circumference, and triglycerides, all of which are associated with metabolic syndrome [30]. *SFRP4* levels are elevated in human failing hearts due to DCM or CAD [31]. Study results suggest that SFRP4 is a novel biomarker of CAD and might play a role in the development of CAD [32].

FRZB (frizzled related protein) encodes a protein that is secreted and is involved in regulating bone development. Defects in this gene are responsible for susceptibility to female-specific osteoarthritis (OA). Previous studies have demonstrated that *FRZB* is highly evolutionarily conserved in vertebrates and that its function is significantly reduced after knockdown using zebrafish embryos, along with a significant reduction in embryonic vascular integrity. This has also been verified in other experiments, and *FRZB* is considered a key gene in the development and progression of abdominal aortic aneurysms [33]. Some experimental studies have shown that SFRP1 and SFRP2 are beneficial for cardiac remodeling [34, 35], increased left ventricular wall tension may be a potential activator of SFRP3 expression and release in vitro in addition [36]. Secreted crimp-related proteins (SFRPs) bind directly to Wnt ligands and may interfere with both classical and non-classical Wnt pathways [35, 37].

Previous studies have shown that DCM is due to a virus that causes natural killer (NK) cells and macrophages to induce a host immune response, leading to cytokine production and inflammatory cell infiltration. In this process, antigen-specific T lymphocytes and antibody-producing B cells induce an immune-mediated response leading to myocardial necrosis. these cells involved in immune-mediation include killer T cells, helper T cells and natural killer cells. therefore, our choice to validate gene expression using the patient's peripheral blood T cells is highly. We therefore chose to validate gene expression using peripheral blood T cells from patients, which is very convincing [38]. It is noteworthy that all the up-regulated genes (*SFRP4* and *FRZB*) of the 2 genes selected by RF in our neuralDCM model were reported to be involved in the Wnt signaling pathway. Two of them were confirmed to be directly involved in the Wnt signaling pathway [35, 39]. All the up-regulated genes we used to construct the neuralDCM model were involved in the Wnt signaling pathway. It can generally indicate that Wnt signaling pathway is activated in DCM patients, and the model we constructed is accurate and effective in accordance with the real situation. Wingless (Wnt) signaling pathways regulate many important cellular processes

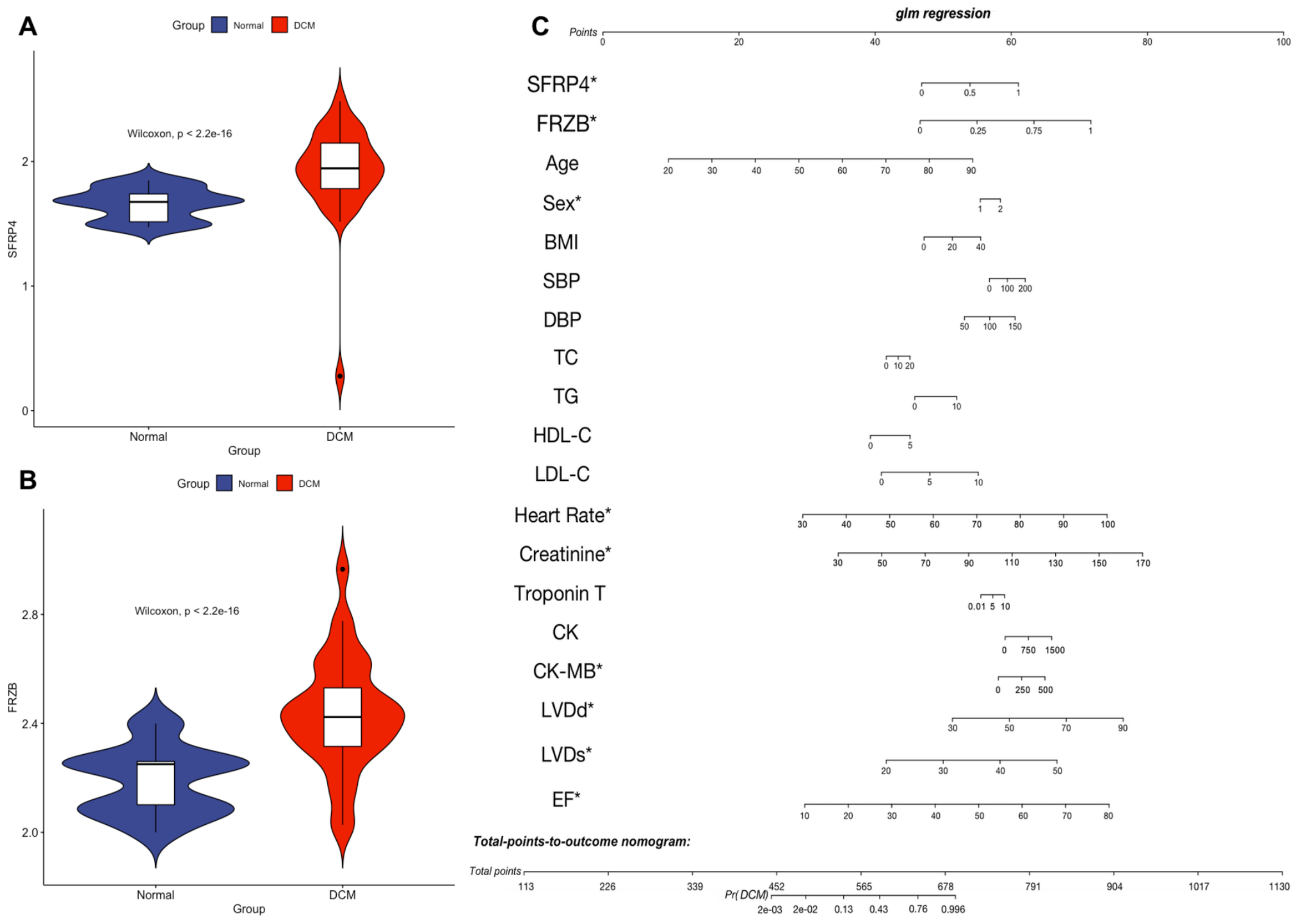

Fig. 8 Expression in patient and Nomogram to estimate individual DCM probability. (**A** - **B**): Expression of two genes in different groups. (**C**): Each predictor variable characteristic has a corresponding point value based on its position on the top point scale and contribution to the model. The probability of DCM for each subject is calculated by summing the points for each variable to obtain a total point value that corresponds to a probability of DCM from the scale presented on the bottom line. *$P < 0.05$

during embryonic development, and proper heart formation requires coordinated Wnt signaling [40, 41]. Wnt activity is generally low in adult organisms and is closely regulated. However, when pathological stress or injury leads to dysregulation, both low and high activity of Wnt signaling is associated with many clinical diseases, including cardiovascular disease [42, 43]. Soluble Wnt modulators are involved in the progression of clinical DCM, and there are potentially complex interactions between different members of the Wnt family. It is conceivable that a better understanding of Wnt signaling in DCM could provide us with new tools in the therapeutic drug apparatus [44].

This manuscript also has some shortcomings, the sample size could be relatively small, which may limit the generalizability of the findings. Additionally, the lack of certain data, such as lifestyle characteristics, raises concerns about potential biases or omissions in the analysis. Third, it fails to explore the specific mechanisms of these two hub genes.

Conclusion

DCM is a specific type of cardiomyopathy that is a common cause of heart failure and death. We performed an integrated analysis that Random Forest and artificial neural networks were constructed for developing a new diagnostic model. Furthermore, RT-PCR and Nomogram were employed to verify the hub genes. *SFRP4* and *FRZB* as a new diagnostic model for DCM. But, the exact mechanism still needs further experimental verification.

Abbreviations

DCM	Dilated Cardiomyopathy
DEG	Differential Expressed Genes
DEIRG	Differentially Expressed Immune-Related Gene
DO	Disease Ontology
GEO	Gene Expression Omnibus
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes And Genomes
MCODE	Molecular Complex Detection
PPI	Protein-Protein Interaction
LVEDd	Left Ventricular Diastolic Diameters
LVEDs	Left Ventricular End-Systolic Diameters
EF	Ejection Fraction

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12872-024-04255-6.

Supplementary Material 1: Table 1: The primer of SFRP4 and FRZB

Supplementary Material 2: Table 2: Raw data pre-processing

Supplementary Material 3: Table 3: GO analysis

Supplementary Material 4: Table 4: KEGG analysis

Supplementary Material 5: Table 5: PPI analysis

Supplementary Material 6: Table 6: Five-fold cross-validation method for neural network

Supplementary Material 7: Table 7: Gene score for neural network

Supplementary Material 8: Table 8: The weight of each gene

Supplementary Material 9: Table 9: The weight of each node

Acknowledgements

Thanks to The National Natural Science Foundation of China (81460063) and Innovative Research Team Project of Guangxi Natural Science Foundation (Grant No. 2018GXNSFGA281006) for the funding of this study.

Author contributions

B.Q. and H.-Y.W. conceived the study, participated in the design, performed the statistical analyses, and drafted the manuscript. C.G. conceived the study, participated in the design and helped to draft the manuscript. X.M. drafted the paper. Y.-F.C. revised the paper. All authors read and approved the final manuscript.

Funding

This study was supported by The National Natural Science Foundation of China (81460063) and Innovative Research Team Project of Guangxi Natural Science Foundation (Grant No. 2018GXNSFGA281006).

Data availability

The datasets generated and/or analyzed during the current study are publicly available.

Declarations

Ethics approval and consent to participate

Prior to their participation in the study, all patients were informed about the experimental procedure and provided written consent. The study was reviewed by an ethics committee. Ethics Committee of First Affiliated Hospital, Guangxi Medical University agreed with the study design (No: Lunshen-2019-KY; Feb. 02, 2019).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 7 May 2023 / Accepted: 14 October 2024

Published online: 18 October 2024

References

1. Merlo M, Pivetta A, Pinamonti B, et al. Long-term prognostic impact of therapeutic strategies in patients with idiopathic dilated cardiomyopathy: changing mortality over the last 30 years. Eur J Heart Fail. 2014;16(3):317–24.
2. Jefferies JL, Towbin JA. Dilated cardiomyopathy. Lancet. 2010;375(9716):752–62.
3. Japp AG, Gulati A, Cook SA, Cowie MR, Prasad SK. The diagnosis and evaluation of dilated cardiomyopathy. J Am Coll Cardiol. 2016;67(25):2996–3010.
4. Zhao J, Lv T, Quan J, et al. Identification of target genes in cardiomyopathy with fibrosis and cardiac remodeling. J Biomed Sci. 2018;25(1):63.
5. Molina-Navarro MM, Roselló-Lletí E, Ortega A, Tarazón E, Otero M, Martínez-Dolz L, et al. Differential gene expression of cardiac ion channels in human dilated cardiomyopathy. PLoS ONE. 2013;8(12). https://doi.org/10.1371/journal.pone.0079792. e79792.
6. Matkovich SJ, Khiami A, Efimov B, Evans IR, Vader S, Jain J, A., et al. Widespread down-regulation of Cardiac mitochondrial and sarcomeric genes in patients with Sepsis. Crit Care Med. 2017;45(3):407–14. https://doi.org/10.1097/CCM.0000000000002207.
7. Witt E, Hammer E, Dörr M, Weitmann K, Beug D, Lehnert K, et al. Correlation of gene expression and clinical parameters identifies a set of genes reflecting LV systolic dysfunction and morphological alterations. Physiol Genomics. 2019;51(8):356–67. https://doi.org/10.1152/physiolgenomics.00111.2018.
8. Ameling S, Herda LR, Hammer E, Steil L, Teumer A, Trimpert C, et al. Myocardial gene expression profiles and cardiodepressant autoantibodies predict response of patients with dilated cardiomyopathy to immunoadsorption therapy. Eur Heart J. 2013;34(9):666–75. https://doi.org/10.1093/eurheartj/ehs330.
9. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
10. Miao L, Yin RX, Zhang QH, et al. A novel circRNA-miRNA-mRNA network identifies circ-YOD1 as a biomarker for coronary artery disease. Sci Rep. 2019;9(1):18314.
11. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
12. Yu G, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015;31(4):608–9.
13. Qi B, Chen JH, Tao L, et al. Integrated Weighted Gene Co-expression Network Analysis Identified that TLR2 and CD40 are related to coronary artery disease. Front Genet. 2020;11:613744.
14. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. J Proteome Res. 2019;18(2):623–32.
15. Wang Y, Miao L, Tao L, et al. Weighted gene coexpression network analysis identifies the key role associated with acute coronary syndrome. Aging. 2020;12(19):19440–54.
16. Deo RC. Machine learning in Medicine. Circulation. 2015;132(20):1920–30.
17. Hengl T, Mendes de Jesus J, Heuvelink GB, et al. SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE. 2017;12(2):e0169748.
18. 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.
19. Weintraub RG, Semsarian C, Macdonald P. Dilated cardiomyopathy. Lancet. 2017;390(10092):400–14.
20. Wang D, Li JR, Zhang YH, Chen L, Huang T, Cai YD. Identification of differentially expressed genes between original breast Cancer and xenograft using machine learning algorithms. Genes (Basel). 2018. 9(3).
21. Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A. A Machine Learning Approach for identifying gene biomarkers guiding the treatment of breast Cancer. Front Genet. 2019;10:256.
22. Tian Y, Yang J, Lan M, Zou T. Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure. Aging. 2020;12(24):26221–35.
23. Xie NN, Wang FF, Zhou J, Liu C, Qu F. Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network. Biomed Res Int. 2020. 2020: 2613091.
24. Dec GW, Fuster V. Idiopathic dilated cardiomyopathy. N Engl J Med. 1994;331(23):1564–75.
25. Alter P, Rupp H, Stoll F, et al. Increased end diastolic wall stress precedes left ventricular hypertrophy in dilative heart failure–use of the volume-based wall stress index. Int J Cardiol. 2012;157(2):233–8.
26. Park JR, Jung JW, Lee YS, Kang KS. The roles of wnt antagonists Dkk1 and sFRP4 during adipogenesis of human adipose tissue-derived mesenchymal stem cells. Cell Prolif. 2008;41(6):859–74.
27. Mahdi T, Hänzelmann S, Salehi A, et al. Secreted frizzled-related protein 4 reduces insulin secretion and is overexpressed in type 2 diabetes. Cell Metab. 2012;16(5):625–33.
28. Brix JM, Krzizek EC, Hoebaus C, Ludvik B, Schernthaner G, Schernthaner GH. Secreted frizzled-related protein 4 (SFRP4) is elevated in patients with diabetes Mellitus. Horm Metab Res. 2016;48(5):345–8.

29. Anand K, Vidyasagar S, Lasrado I, et al. Secreted frizzled-related protein 4 (SFRP4): a novel biomarker of β-Cell dysfunction and Insulin Resistance in individuals with prediabetes and Type 2 diabetes. Diabetes Care. 2016;39(9):e147–8.
30. Hoffmann MM, Werner C, Böhm M, Laufs U, Winkler K. Association of secreted frizzled-related protein 4 (SFRP4) with type 2 diabetes in patients with stable coronary artery disease. Cardiovasc Diabetol. 2014;13:155.
31. Schumann H, Holtz J, Zerkowski HR, Hatzfeld M. Expression of secreted frizzled related proteins 3 and 4 in human ventricular myocardium correlates with apoptosis related gene expression. Cardiovasc Res. 2000;45(3):720–8.
32. Ji Q, Zhang J, Du Y, et al. Human epicardial adipose tissue-derived and circulating secreted frizzled-related protein 4 (SFRP4) levels are increased in patients with coronary artery disease. Cardiovasc Diabetol. 2017;16(1):133.
33. Oh CK, Ko Y, Park JJ et al. FRZB as a key molecule in abdominal aortic aneurysm progression affecting vascular integrity. Biosci Rep. 2021. 41(1).
34. Mirotsou M, Zhang Z, Deb A, et al. Secreted frizzled related protein 2 (Sfrp2) is the key akt-mesenchymal stem cell-released paracrine factor mediating myocardial survival and repair. Proc Natl Acad Sci U S A. 2007;104(5):1643–8.
35. Bovolenta P, Esteve P, Ruiz JM, Cisneros E, Lopez-Rios J. Beyond wnt inhibition: new functions of secreted frizzled-related proteins in development and disease. J Cell Sci. 2008;121(Pt 6):737–46.
36. Askevold ET, Gullestad L, Nymo S, et al. Secreted frizzled related protein 3 in Chronic Heart failure: analysis from the controlled rosuvastatin multinational trial in Heart failure (CORONA). PLoS ONE. 2015;10(8):e0133970.
37. Kawano Y, Kypta R. Secreted antagonists of the wnt signalling pathway. J Cell Sci. 2003;116(Pt 13):2627–34.
38. Lasrado N, Reddy J. An overview of the immune mechanisms of viral myocarditis. Rev Med Virol. 2020;30(6):1–14. https://doi.org/10.1002/rmv.2131.
39. Gay D, Ghinatti G, Guerrero-Juarez CF, et al. Phagocytosis of wnt inhibitor SFRP4 by late wound macrophages drives chronic wnt activity for fibrotic skin healing. Sci Adv. 2020;6(12):eaay3704.
40. van Amerongen R, Nusse R. Towards an integrated view of wnt signaling in development. Development. 2009;136(19):3205–14.
41. Gessert S, Kühl M. The multiple phases and faces of wnt signaling during cardiac differentiation and development. Circ Res. 2010;107(2):186–99.
42. Kim KI, Park KU, Chun EJ, et al. A novel biomarker of coronary atherosclerosis: serum DKK1 concentration correlates with coronary artery calcification and atherosclerotic plaques. J Korean Med Sci. 2011;26(9):1178–84.
43. Clevers H, Nusse R. Wnt/β-catenin signaling and disease. Cell. 2012;149(6):1192–205.
44. He W, Zhang L, Ni A, et al. Exogenously administered secreted frizzled related protein 2 (Sfrp2) reduces fibrosis and improves cardiac function in a rat model of myocardial infarction. Proc Natl Acad Sci U S A. 2010;107(49):21110–5.

Publisher's note