

Original Research

Construction of Prediction Model for Atrial Fibrillation with Valvular Heart Disease Based on Machine Learning

Qiaoqiao Li^{1,2,†}, Shenghong Lei^{1,2,†}, Xueshan Luo^{1,2}, Jintao He^{1,2}, Yuan Fang^{1,2},
 Hui Yang^{1,2}, Yang Liu^{1,2}, Chun-Yu Deng^{1,2}, Shulin Wu^{1,2}, Yu-Mei Xue^{1,2,*}, Fang Rao^{1,2,*}

¹Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, 510080 Guangzhou, Guangdong, China

²Research Center of Medical Sciences, Provincial Key Laboratory of Clinical Pharmacology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, 510080 Guangzhou, Guangdong, China

*Correspondence: xueymeimei@gdph.org.cn (Yu-Mei Xue); raofang@gdph.org.cn (Fang Rao)

†These authors contributed equally.

Academic Editor: Carmela Rita Balistreri

Submitted: 31 March 2022 Revised: 31 May 2022 Accepted: 10 June 2022 Published: 28 June 2022

Abstract

Background: Valvular heart disease (VHD) is a major precipitating factor of atrial fibrillation (AF) that contributes to decreased cardiac function, heart failure, and stroke. Stroke induced by VHD combined with atrial fibrillation (AF-VHD) is a much more serious condition in comparison to VHD alone. The aim of this study was to explore the molecular mechanism governing VHD progression and to provide candidate treatment targets for AF-VHD. **Methods:** Four public mRNA microarray datasets were downloaded and differentially expressed genes (DEGs) screening was performed. Weighted gene correlation network analysis was carried out to detect key modules and explore their relationships and disease status. Candidate hub signature genes were then screened within the key module using machine learning methods. The receiver operating characteristic curve and nomogram model analysis were used to determine the potential clinical significance of the hub genes. Subsequently, target gene protein levels in independent human atrial tissue samples were detected using western blotting. Specific expression analysis of the hub genes in the tissue and cell samples was performed using single-cell sequencing analysis in the Human Protein Atlas tool. **Results:** A total of 819 common DEGs in combined datasets were screened. Fourteen modules were identified using the cut tree dynamic function. The cyan and purple modules were considered the most clinically significant for AF-VHD. Then, 25 hub genes in the cyan and purple modules were selected for further analysis. The pathways related to dilated cardiomyopathy, hypertrophic cardiomyopathy, and heart contraction were concentrated in the purple and cyan modules of the AF-VHD. Genes of importance (*CSRP3*, *MCOLN3*, *SLC25A5*, and *FIBP*) were then identified based on machine learning. Of these, *CSRP3* had a potential clinical significance and was specifically expressed in the heart tissue. **Conclusions:** The identified genes may play critical roles in the pathophysiological process of AF-VHD, providing new insights into VHD development to AF and helping to determine potential biomarkers and therapeutic targets for treating AF-VHD.

Keywords: atrial fibrillation; valvular heart disease; WGCNA; machine learning; specific markers

1. Introduction

Atrial fibrillation (AF) is the most prevalent arrhythmia within the general population [1]. Morbidity and mortality linked to AF represent a significant public health burden worldwide [2]. There are multiple factors contributing to AF, including valvular heart disease (VHD), hypertension, age, obesity, and diabetes [3,4]. VHD is a significant cause of arrhythmia and stroke. Stroke induced by AF-VHD is a more serious condition compared to VHD alone [5,6]. However, the mechanism for the development of VHD into AF-VHD is not yet fully understood. It is therefore essential to investigate the pathogenesis and clarify the precise molecular mechanisms involved in AF-VHD progression.

Lamirault *et al.* [7] identified the gene expression profile associated with human AF-VHD. In their study, eight atrial appendages in 11 patients with AF-VHD and 7 pa-

tients with sinus rhythm with VHD (SR-VHD) undergoing open heart surgery were included in cardiac-specific microarray analysis. The results indicated that 169 genes were differentially expressed between the two groups. Notably, cysteine- and glycine-rich protein 3 (*CSRP3*) was found to be highly expressed in AF-VHD and involved in cardiac contraction. Furthermore, Yan *et al.* [8] and Li *et al.* [9] screened key immune-related genes in AF-VHD. Liu *et al.* [10] also identified feature genes for AF with VHD using integrative transcriptomic, proteomic, and machine learning approaches. In our study, by contrast, we merged related datasets and used weighted gene co-expression network analysis (WGCNA), a statistical method that is similar to cluster analysis but is more biologically meaningful, and machine learning methods to identify specific biomarkers [11,12]. Even though many studies have investigated AF-VHD markers, specific predictive biomarkers are still



lacking to enable early detection. The communicative regulatory mechanisms of AF-VHD also remain poorly understood.

In the present study, co-expression networks were constructed using the dataset GSE115574's expression profile to identify key modules and hub genes related to AF with VHD. The genes were then subjected to gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) ontology enrichment analyses. Subsequently, important genes (*CSRP3*, Transient Receptor Potential Channel Mucolipin 3 (*MCOLN3*), solute carrier family 25 member 5 (*SLC25A5*), and *FGF1* intracellular binding protein (*FIBP*)) were identified using a machine learning approach and their potential clinical significance was determined. Notably, the clinical significance of *CSRP3* and *MCOLN3* was statistically significant. Functional enrichment results showed that *CSRP3* has a strong association with heart development and *MCOLN3* is linked to the calcium channel complex. Additionally, the available literature indicated that *CSRP3* is closely involved in the process of cardiac hypertrophy [13,14]. These observations may link AF with dilated or hypertrophic cardiomyopathy, providing novel evidence for the diagnosis and treatment of AF with VHD in a clinical setting.

2. Methods

2.1 Atrial Fibrillation Datasets Filtration

Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) was used to extract the raw datasets and included if datasets met the following criteria. (1) The expression profiling was acquired by array; (2) The experimental platform belonged to *GPL570* platform; (3) Data sets should include gene expression profiles of human left or right atria tissues; (4) All experiments included controls and the minimum sample size used was three.

2.2 Data Collection

Based on the criteria mentioned above, four publicly microarray datasets (GSE79768, GSE115574, GSE14975 and GSE41177) were filtered and downloaded. In the GSE115574 dataset, the expression matrix of a total of 59 samples was acquired from the human atrial tissue, containing 15 patients diagnosed with AF-VHD, and 15 with SR-VHD. Left atrial appendage and right atrial appendage samples (except 1 patient) were obtained from each patient. In GSE79768, a total of 7 patients with AF-VHD and 6 with SR-VHD provided atrial tissue samples. GSE14975 contained 5 atrial tissue samples from AF-VHD patients and 5 from SR-VHD control. The GSE41177 dataset included 16 patients with persistent AF receiving valvular surgery and 3 patients diagnosed with SR with VHD.

2.3 Data Preprocessing and Differentially Expressed Genes (DEGs) Screening

Furthermore, Probe IDs were mapped to gene symbols in each dataset by extracting them from the respective platform file. The microarray data preprocessing, containing normalization and background correction, was conducted by applying the “Affy” package in R [15]. Then, using the “sva” package of the R software (version 3.6.3) to merge and batch normalization four datasets. Afterwards, the “limma” package was used to identify the DEGs between the AF-VHD and SR-VHD. The statistical cutoff values were an absolute \log_2 FC >0.3 (FC, fold change) and adjust p -value < 0.05 in combined datasets. Volcano plot and Heatmap were generated according to the data above by using R package “ggplot2” and “pheatmap”.

2.4 Functional Enrichment Analysis

In order to explore the biological function of the DEGs and genes in key modules, GO analysis and a KEGG terms enrichment analysis were performed using Metascape tool (<http://metascape.org>) and ClusterProfiler (version 3.6.0) software in R language [16,17]. Enrichment significance thresholds were set at an adjust p -value below 0.05. Furthermore, Gene set enrichment analysis (GSEA) was performed by clusterProfiler (R package) and GSEA plots were visualized by “gseaplot” package [18]. Results with a $|NES| > 1$ and FDR <0.25 were regarded as statistically significant (NES, normalized enrichment score; FDR, false discovery rate).

2.5 Construction of Weighted Gene Co-Expression Network

Based on the median absolute deviation of the genes, we selected the top 5000 genes for WGCNA using the R package “WGCNA” [19,20]. Biological networks were constructed with a value of 9 for the soft thresholding parameter to satisfy the scale-free assumption (linear regression model fitting index $R^2 = 0.91$). We calculated pairwise Pearson's correlation matrix and then transformed it into an adjacency matrix. A Topological Overlap Measure (TOM)-based dissimilarity matrix (DissTOM) was created by transforming the adjacency matrix, and modules were generated by hierarchical average linkage clustering analysis for the gene dendrogram. After acquiring modules, module eigengene (ME), first principal component of the expression matrix of the referred to module, was calculated using the “Module Eigengenes” function. The relationship between clinical parameters and modules were indirectly assessed by looking at the correlation between MEs and clinical traits. Module significance (MS) was calculated by taking the average of the gene scores for all genes within a module. Modules with the highest MS scores were regarded as key modules and selected for subsequent analysis. Additionally, we extracted gene expression profiles of each module genes for further analysis.

2.6 Identification of Candidate Hub Genes in Key Module

The module membership (MM) was also calculated, which was regarded as the degree of association between the ME and the gene expression matrix. Then, the intramodular hub genes were identified based on gene significance (GS) >0.2 and MM >0.8 [21]. Heatmaps were conducted to demonstrate the putative candidate genes' expression patterns with the R package "pheatmap" [22]. We also made Venn diagrams for common DEGs in four public microarray datasets and hub genes in WGCNA.

2.7 TF-miRNA Network of Hub Genes

The TF-miRNA regulatory network was constructed by overlapping TF-hub genes and miRNA-hub genes using the Network-Analyst database (<http://www.networkanalyst.ca>) and then visualized by performing Cytoscape (version 3.7.2) [23–25] (TF, transcription factor).

2.8 Machine Learning: Construction of Lasso and Random Forest Model

We performed least absolute shrinkage and selection operator (LASSO) regression by applying the "glmnet" package in R software to identify the candidate predictive features based on a generalized linear model [26]. Moreover, we constructed Random Forest (RF) Model to identify the important variables by using "Random Forest" package [27]. Finally, we screened real hub gene signatures by intersecting gene signatures from the LASSO and RF.

2.9 Development and Validation of a Prognostic Model

The association analysis between hub genes was performed by using the spearman rank correlation coefficient, illustrated by heat-map. To screen out the potential clinical significance of hub genes, the receiver operating characteristic curve (ROC) was created using "pROC" packages [28]. A multivariate regression formula was built based on the hub genes' expression value and their regression coefficients under the merged datasets. Finally, a nomogram was constructed based on the selected predictive factors identified by using the "rms" package in R to predict the prevalence of AF-VHD. Calibration curves were plotted to evaluate the difference between the predicted probability and the actual probability. In addition, a decision curve analysis (DCA) can be used to measure the net benefit of a predictive model [29].

2.10 Patients

Left atrial appendages were obtained from 6 VHD patients with AF and 5 VHD patients with SR undergoing open heart surgery. Patients with hyperthyroidism, diabetes, hypertension and infectious diseases were excluded from our study. After the surgical operation, liquid nitrogen was immediately applied to the tissue specimens. Human participants in the studies were reviewed and approved by the ethics committee of the Guangdong General

Hospital, Guangdong Academy of Medical Sciences (No. GDREC2017111H). A signed informed consent form was provided to all patients and their legal representatives.

2.11 Western Blot Analysis

The experiment was conducted based on the procedure reported as previously [30]. Antibodies used were mentioned as follows: anti-CSRP3 antibody [1:1000, Abcam Cat# ab172952]; anti-MCOLN3 antibody [1:1000, Thermo Fisher Invitrogen Cat# PA5-109339]; anti-GAPDH antibody [1:5000, Proteintech Cat# 60004-1-Ig].

2.12 Specific Expression Analysis of Hub Genes in Tissue and Cells

We used Human Protein Atlas (HPA) (<https://www.proteinatlas.org/>) tool to validate the mRNA and protein expression levels of the hub genes in all tissues. The Single Cell Type Atlas part in HPA as used to illustrate the expression of hub genes in single specific cell types [31].

2.13 Statistical Analysis

A description of the bioinformatic analyses appeared in corresponding subsections. In all cases, values were expressed as means and standard deviations (SD), and Student's *t*-test was used to determine pairwise statistical significance of the differences between two group means. A *p*-value < 0.05 was defined as statistically significant.

3. Results

3.1 Study Workflow

The flowchart for the study analysis strategy is shown in Fig. 1. Fourteen modules of co-expressed genes were identified via weighted gene correlation network analysis. The cyan and purple modules were identified as the most clinically significant. GO and KEGG analyses were performed on both modules, with similar processes being carried out on the hub genes within the key modules. Crucial candidate genes were identified by intersecting the hub genes in WGCNA and common DEGs in merged datasets. TF-crucial genes and miRNA-crucial gene networks were visualized using Cytoscape. Machine learning methods, including Lasso and random forest, were implemented to select the potential significant genes and to construct a diagnosis prediction model. The ROC curve and DCA were utilized on this prediction model to assess the predictive power. Then, protein expression levels of important genes in AF-VHD were verified using western blot analysis. Specific expression analysis of hub genes in tissue and cell samples was performed to identify specific biomarkers.

3.2 DEGs Screening

Datasets GSE115574, GSE41177, GSE79768, and GSE14975 were included in the analysis. Based on the screening criteria, 819 genes in the merged datasets were screened out as common DEGs, of which 725 genes were

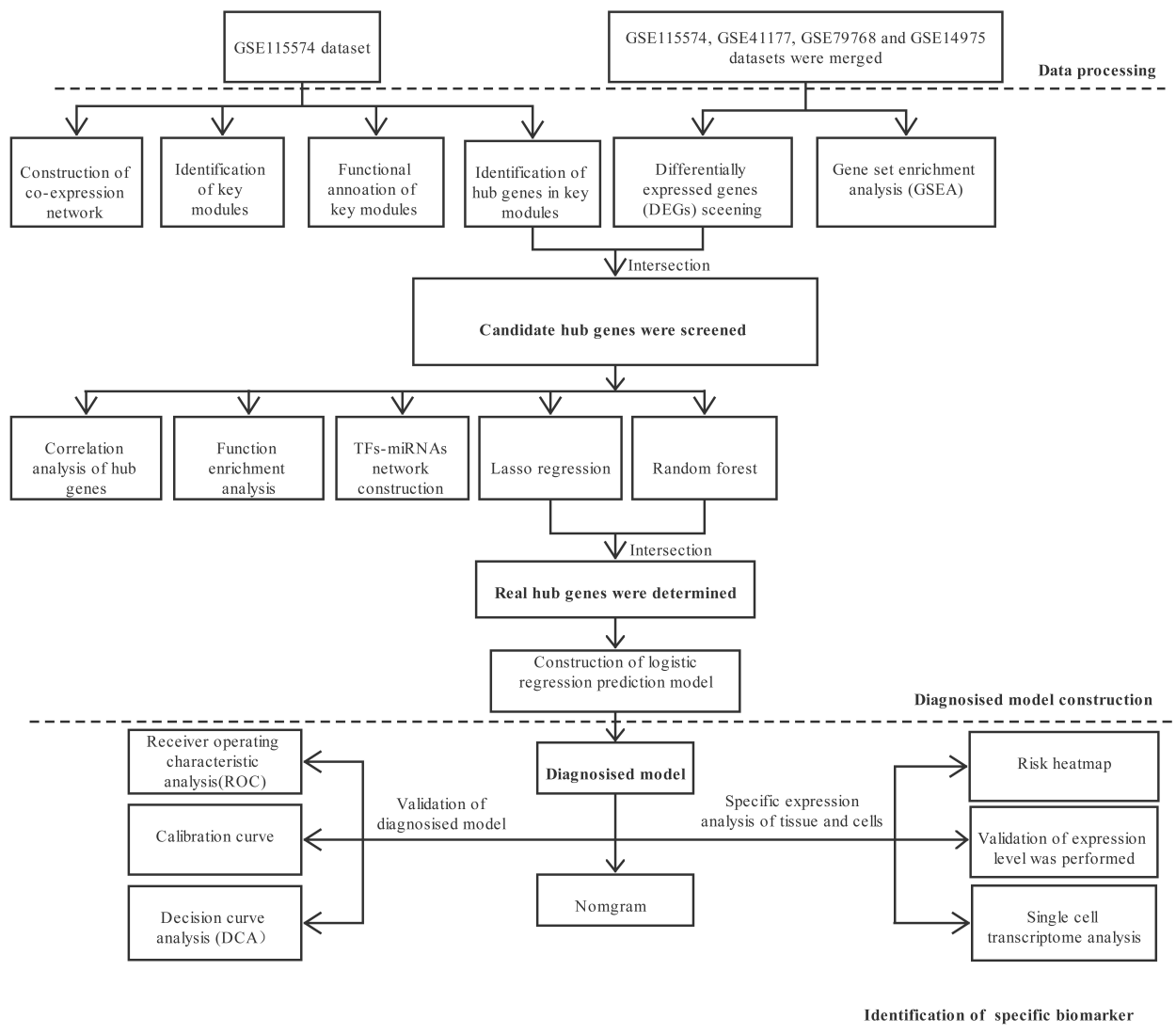


Fig. 1. Flowchart of the analysis strategy.

up-regulated and 94 genes were down-regulated (Fig. 2A). DEGs were ranked according to the fold change expression, and the top 40 were represented using a heatmap (Fig. 2B).

3.3 Functional Enrichment of DEGs

All DEGs were used in functional annotation analyses. The top five significant terms were displayed in bubble plots according to their adjusted p -values (Fig. 3A,B). The GO terms were associated with the molecular functions (MFs), cellular components (CCs), and biological processes (BPs). Those linked with MFs included extracellular matrix structural constituent, extracellular matrix structural conferring tensile strength, IgG binding, electron activity, and heparin binding. CCs included collagen-containing extracellular matrix, mitochondrial inner membrane, endoplasmic reticulum lumen, and collagen trimer. BPs included extracellular matrix organization, collagen fibril organization, neutrophil activation, and neutrophil-mediated

immunity (Fig. 3A). Terms of the enriched KEGG pathway are represented in Fig. 3B, including the phagosome, Fc gamma R-mediated phagocytosis, regulation of actin cytoskeleton, carbon metabolism, and the advanced glycation end products (AGEs) and its synergetic receptor-AGEs-RAGE signaling pathway in diabetic complications. Detailed results are summarized in the **Supplementary Tables 1,2**. The top 6 BP enrichment terms were determined by their adjusted (adj) p -values and BgRatio values. and chord plots were used (Fig. 3C) in order to better understand the molecular functions of DEGs and the potential biological processes in which they could be involved. GSEA was performed based on all genes (Fig. 3D). The AF-VHD groups were enriched in terms of class I major histocompatibility complex (MHC)-mediated antigen presentation, neutrophil degranulation, platelet activation signaling and aggregation, and extracellular matrix organization.

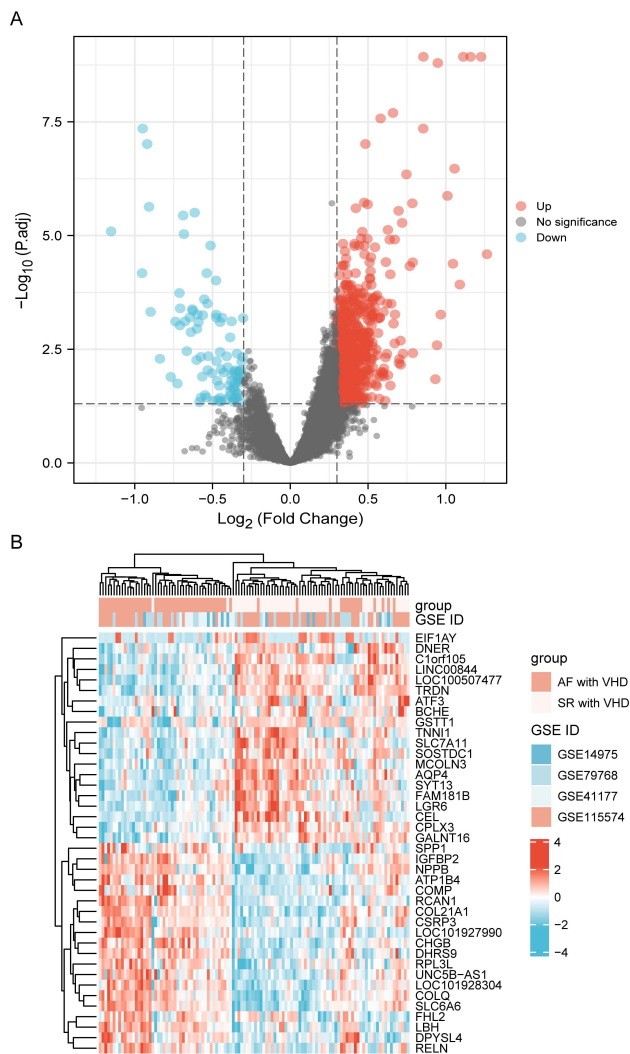


Fig. 2. DEG identification. (A) Volcano plot visualization of differentially expressed genes and their statistical significance. The red dots indicate up-regulated genes, and the blue dots indicate down-regulated genes. (B) Heatmap showing expression profiles of top 40 DEGs. DEGs, differentially expressed genes.

3.4 Construction of Co-Expression Network Using WGCNA

Two outliers (Sample15/GSM3182694; Sample26/GSM3182705) were observed (**Supplementary Fig. 1**) in the atrium samples in GSE115574. A total of 26 AF with VHD samples and 31 SR with VHD samples were included in the analysis after discarding the outliers. To satisfy the scale-free assumption of the constructed biological networks, the soft threshold power $\beta = 9$ was selected for the AF- and SR-VHD samples (**Supplementary Fig. 2A,B**). By calculating the scale-free topology fitting index, the value of R^2 was shown to reach 0.91. The results were represented via a histogram and a linear plot (**Supplementary Fig. 2C,D**). Additionally, an average hierarchical linkage clustering was calculated using TOM-based dissimilarity measurements with a

minimum size of 30 genes. Modules of every gene cluster were identified based on the hierarchical cluster analysis (Fig. 4A). Following the merging of similar modules with a clustering height cut-off of 0.25 (Fig. 4B), a total of 14 modules with a high credibility were obtained, with the initial and merged modules presenting under the clustering tree (Fig. 4C).

3.5 Identification of the Most Significant Modules and Hub Genes

The module-trait relationships illustrate the correlation between the available clinical features (disease status, tissue site) and each module in GSE115574 by calculating the value of MS (Fig. 5A). Notably, the ME of the cyan module ($r = 0.54, p = 1 \times 10^{-5}$) showed the highest linkage with AF-VHD, followed by the purple module ($r = -0.51, p = 4 \times 10^{-5}$). Additionally, the mean GS across all genes in each module was illustrated by the MS values displayed in a bar diagram. The cyan and purple modules showed to be of substantial interest (Fig. 5B). Therefore, the cyan and purple modules were selected as the main focus modules, and scatter plot analysis was conducted to determine the correlation between the GS and MM of the cyan (Fig. 6A) and purple (Fig. 6B) modules. Highly connected genes (hub genes) were defined using module connectivity ($MM > 0.8$) and clinical trait relationship ($GS > 0.2$). Under these criteria, 25 genes were identified as candidates for further analysis. MM and GS values were detailed in **Supplementary Table 3**. A heatmap visualizing gene expression changes for 25 hub genes in merged datasets was shown in Fig. 6C.

3.6 Function Enrichment Analysis of Genes in Key Modules

GO and KEGG analyses were performed to gain a deeper understanding of the biological functions of genes in the cyan and purple modules. The GO results showed that genes in the purple module were mainly clustered in MF, CC, and BP, including “metal ion transmembrane transporter activity”, “voltage-gated ion channel activity”, “sarcooplasm”, “T-tubule”, “muscle system process”, “regulation of blood circulation”, and “regulation of heart contraction” (Fig. 5C). In addition, KEGG analysis of genes in the purple modules showed that they were mainly enriched in the following terms: adrenergic signaling in cardiomyocytes, HIF-1 signaling pathway, hypertrophic cardiomyopathy, and dilated cardiomyopathy (Fig. 5D). The top five terms from the Metascape analysis of the cyan module included carbon metabolism, striated muscle cell development, cardiac muscle cell development, nucleoid, and regulation of membrane permeability. The top 20 cluster-enriched sets are shown in Fig. 5E.

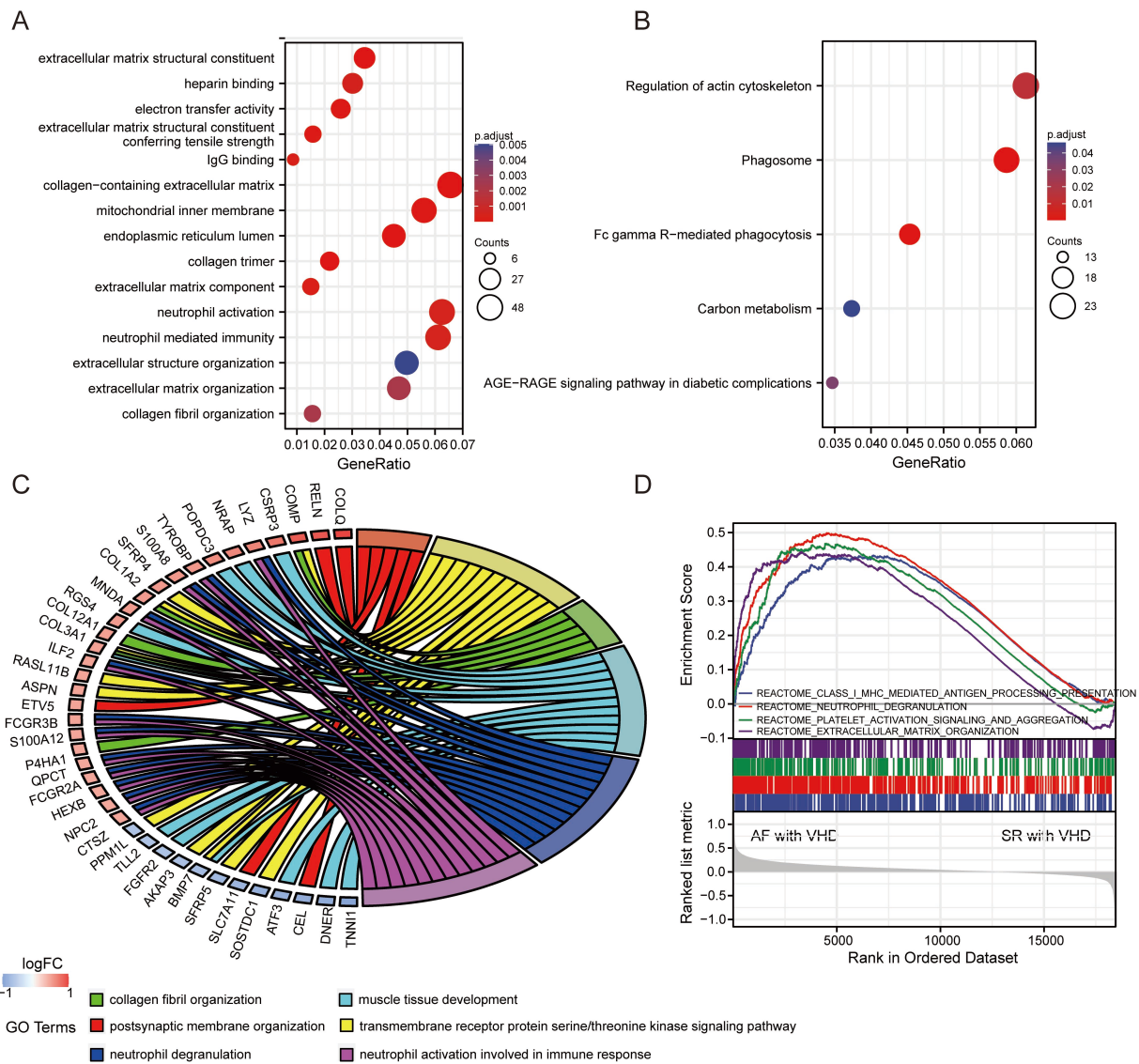


Fig. 3. Functional enrichment analysis of DEGs and GSEA results. (A) GO term enrichment with top five significant adjusted p -values for DEGs illustrated in relation to biological process, molecular function, and cellular component. (B) Top five enriched KEGG pathways for DEGs. (C) Chord plots for gene enrichment analysis. (D) GSEA plots.

3.7 Screening Candidate Hub Genes

In order to screen the candidate hub genes for further analysis, DEGs of the combined datasets (GSE115574, GSE41177, GSE79768, and GSE41177) were overlapped with the hub genes in the key modules using a Venn diagram. A total of 15 candidate hub genes were identified (Fig. 7A). They were mainly enriched in the Mitogen-Activated Protein Kinase (MAPK) signaling pathway (gene ratio 4/10), arrhythmogenic right ventricular cardiomyopathy (gene ratio 3/10), dilated cardiomyopathy, cardiac muscle contraction (gene ratio 2/10), structural constituent of muscle, calcium channel activity (gene ratio 4/14), and calcium ion transmembrane transporter activity (gene ratio 3/14; Fig. 7B). To determine which TFs and miRNAs may be responsible for the altered candidate hub gene expres-

sion, transcription factor and miRNA analyses were performed using NetworkAnalyst. A total of 61 TFs and 96 miRNAs were identified (Fig. 7C). The correlation between 11 hub genes was determined using Spearman's rank test. A positive correlation was observed between *CSRP3* and *SLC25A5* (Fig. 7D).

3.8 Identification of Real Hub Genes in AF-VHD

Lasso and RF analyses were performed to screen signatures within 15 candidate hub genes in AF-VHD. First, the Lasso algorithm identified ten signatures—*CSRP3*, *MCOLN3*, *SLC25A5*, *FIBP*, *ABCF1*, *ACTN2*, *ASTN2*, *CACNA2D2*, *OTOGL*, and *DUSP3*—under the condition of the best penalty parameter (λ) (Fig. 8A,B). Using RF, the top five most important variables were screened, which

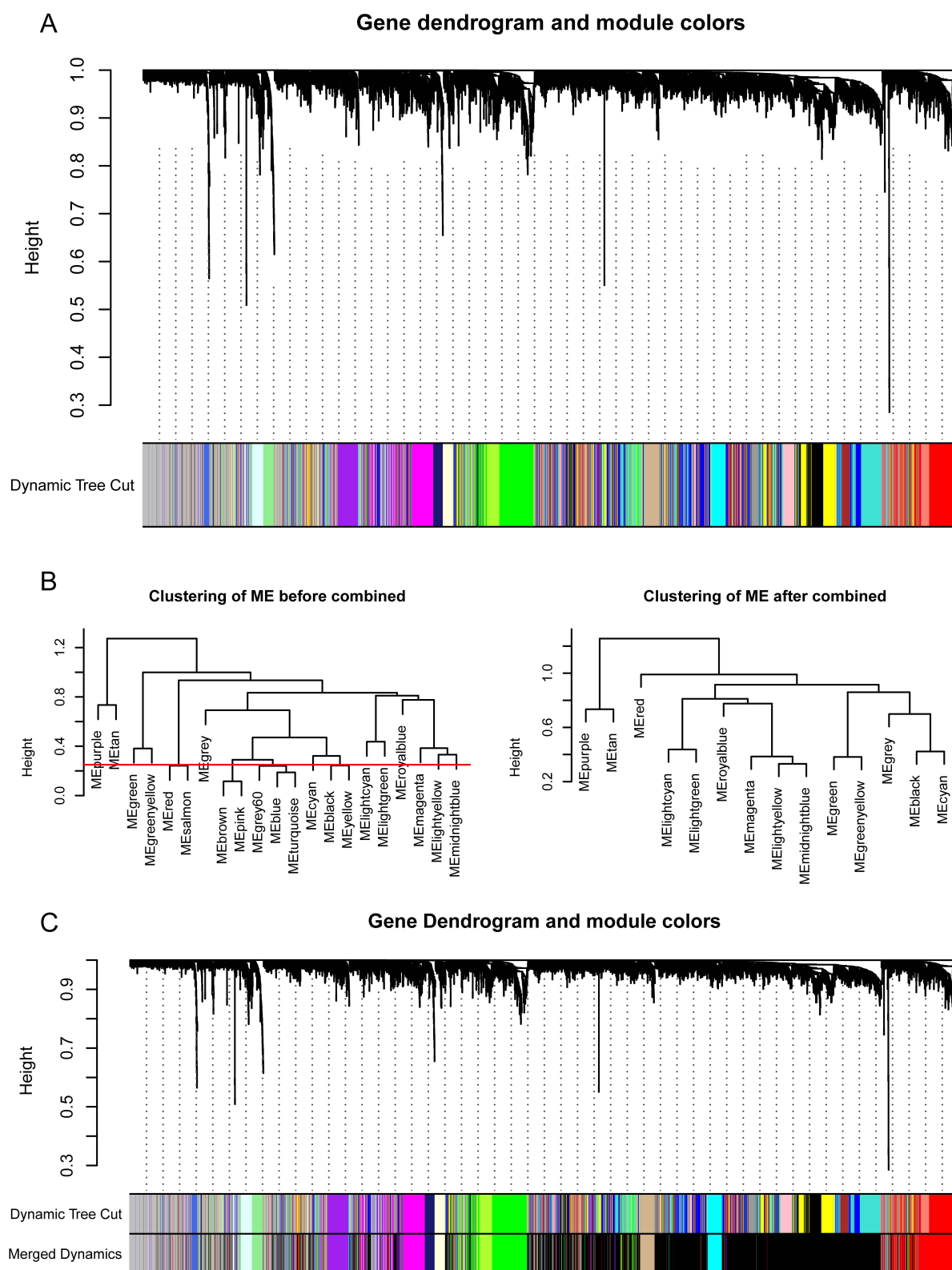


Fig. 4. Construction of weighted gene co-expression network of AF-VHD and SR-VHD samples. (A) Gene dendrograms and modules were acquired using average linkage hierarchical clustering with dissimilarity according to topological overlap. Color beneath each row is a reflection of module assignment; there are different colors for different modules. (B) Dendrogram clustering was performed with 0.25 as height to identify similar modules. (C) After dynamic tree cutting and merging, 14 gene modules were obtained.

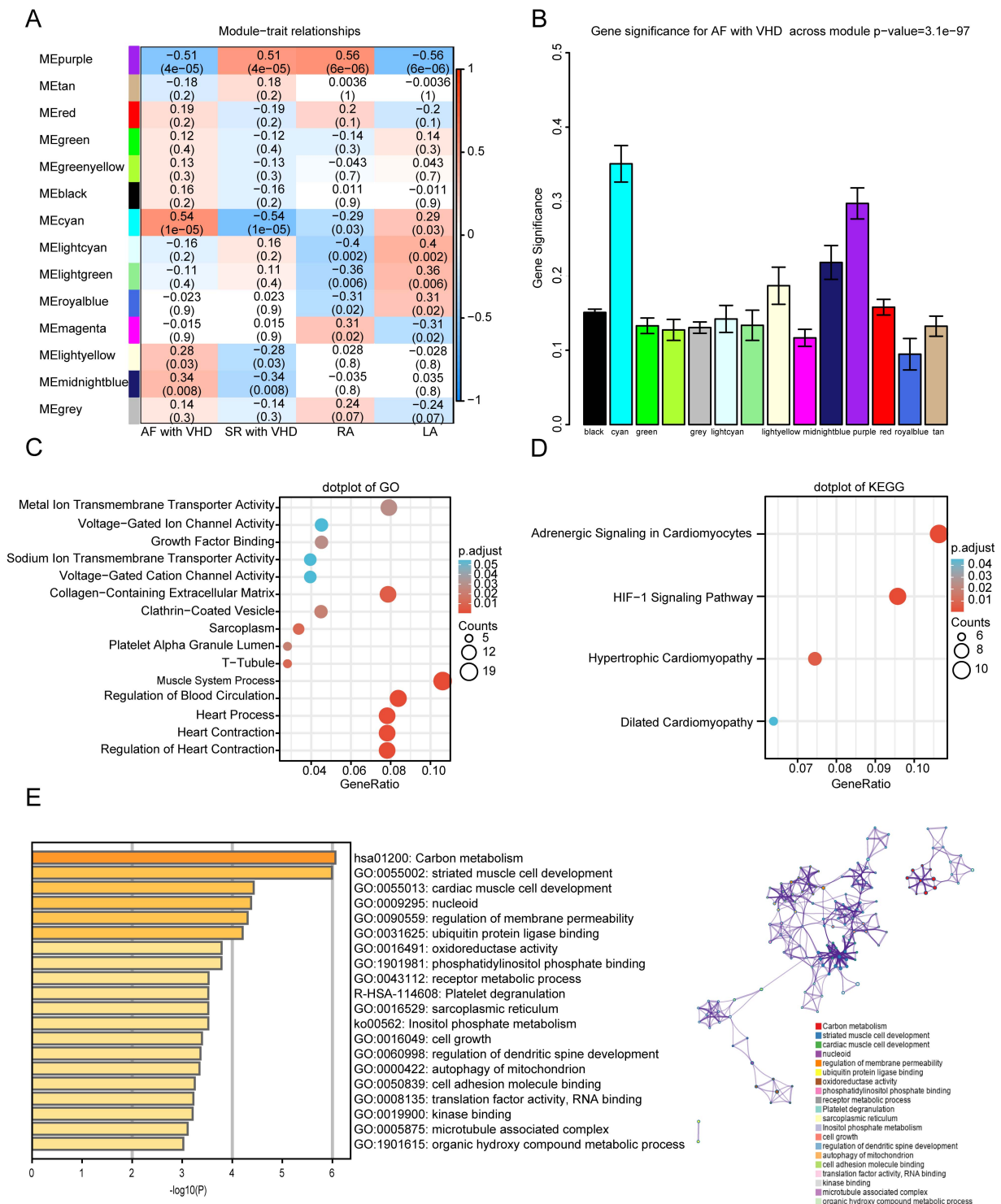


Fig. 5. Functional enrichment of the key module. (A) Module-trait relationship plot. There is one row for each module eigengene and one column for each trait. Correlation coefficients and p -values are displayed in each cell. Red represents a positive correlation, and blue represents a negative correlation. RA, right atrium; LA, left atrium. (B) Bar plot of module significance (MS) defined as the average absolute value of gene significance (GS) for all genes in a module. The cyan and purple modules are the most promising. (C,D) GO enrichment and KEGG analyses for purple module. (C) Top five GO category terms of biological process (BP), cellular component (CC), and molecular function (MF) were identified. (D) Top four terms of KEGG analysis. A p -value (adjusted) < 0.05 was considered significant. (E) Top 20 clusters from Metascape GO enrichment analysis of cyan module-associated genes.

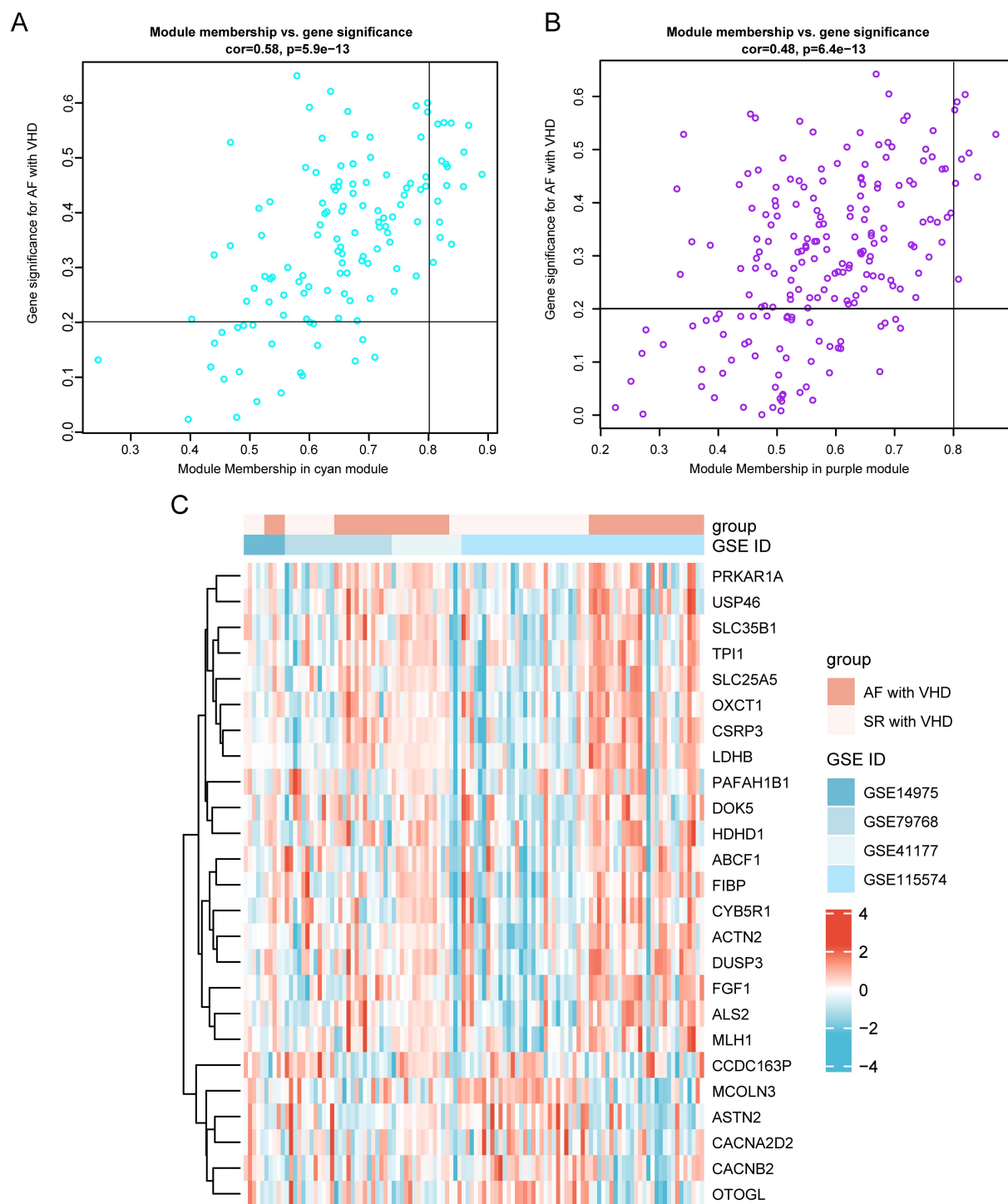


Fig. 6. Identification of hub genes in the key module. (A) Scatterplot of disease status for gene significance (GS) vs. disease status for module membership (MM) in cyan module. (B) Scatterplot of GS vs. MM for purple module. Hub genes were screened based on the following criteria: GS > 0.2 and MM > 0.8. (C) Although the two groups of samples are not completely separated, heatmaps can show the expression level and trend of change patterns of hub genes in key modules.

were *CSRP3*, *OXCT1*, *SLC25A5*, *FIBP*, and *MCOLN3* (Fig. 8C,D). After intersecting the gene signatures selected by Lasso and RF, *CSRP3*, *MCOLN3*, *SLC25A5*, and *FIBP* were determined to be the real hub genes in AF-VHD

(Fig. 8E). The mRNA expression levels of hub genes in SR- and AF-VHD were investigated (Fig. 8F).

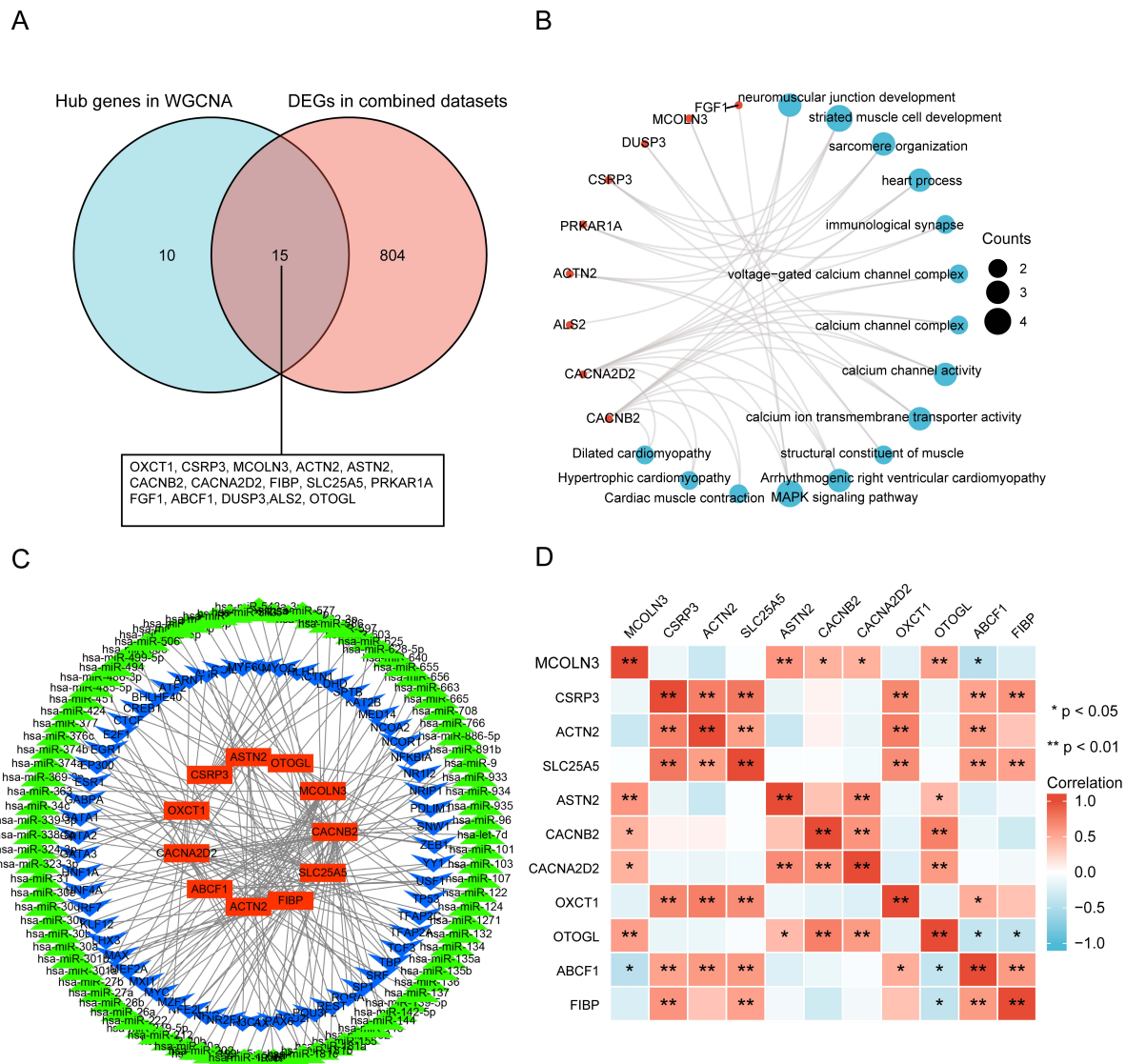


Fig. 7. Overlap of genes in key modules with DEGs and construction of TF-miRNA networks. (A) Venn diagram represents unique and shared genes between common DEGs in merged datasets and hub genes in WGCNA. A total of 15 candidate hub genes were identified. (B) Functional enrichment analysis of candidate hub genes. (C) TF-miRNA regulatory network. Red rectangle represents hub genes, blue V represents TFs, and green triangle represents miRNAs. (D) Heatmap with Spearman's correlations among 11 hub genes.

3.9 Construction and Validation of Diagnosis Model

The ROC analysis was conducted to further validate the diagnostic value of the hub genes in merged datasets. The results demonstrated that *CSRP3* (Area Under Curve, AUC 0.843), *MCOLN3* (AUC 0.771), *SLC25A5* (AUC 0.795), and *FIBP* (AUC 0.735) had a general ability to discriminate between AF with VHD and SR with VHD (Fig. 9A). Multiple biomarkers were combined as a sensitive screening index for AF-VHD in order to improve the diagnosis sensitivity. A combined diagnosis model of four vital genes was used to show that the AUC value of AF-VHD reached 0.866 (95% confidence interval (CI): 0.798–0.935; Fig. 9B). Calibration curves revealed that the diagnosis model had a good performance when predicting AF-VHD incidence (Fig. 9C). The blue line in the DCA curve

remained above the gray and black lines between 0 and 0.8, implying that decisions based on the diagnosis model may be beneficial to AF-VHD patients (Fig. 9D). A nomogram was established using the RMS package for the diagnosis of AF-VHD based on the four crucial genes (*CSRP3*, *MCOLN3*, *SLC25A5* and *FIBP*) (Fig. 9E).

3.10 Key Genes Specifically Expressed in Heart Tissue

Four genes were selected to calculate the risk score according to their coefficients, where risk score = $-20.0615 + 1.838 \times CSRP3 + -0.9575 \times MCOLN3 + 0.1821 \times SLC25A5 + 0.1969 \times FIBP$. Based on the median risk score, patients were categorized into high-risk and low-risk groups. As indicated in Fig. 10A, the prevalence of AF-VHD was significantly increased with an increased risk

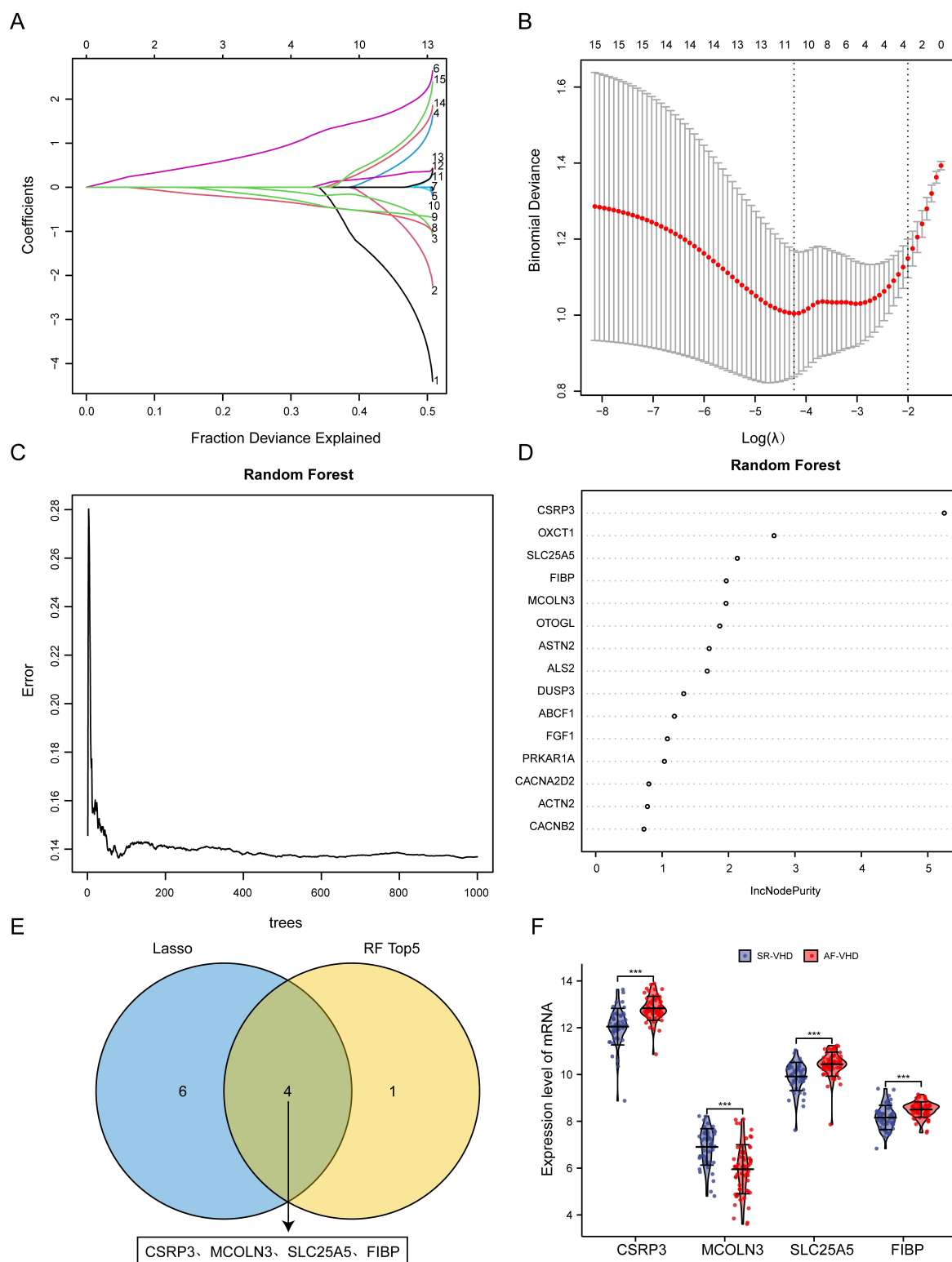


Fig. 8. Identification of real hub gene signatures using machine learning. (A) Lasso coefficient profiles of the key genes in AF-VHD. Coefficients are illustrated via corresponding fraction deviance. (B) Selection of tuning parameter in Lasso regression models. This is a plot of the binomial deviance metrics (the y-axis) against $\log(\lambda)$ (the bottom x-axis). Red dot indicates average deviance values for each model with a given λ . (C) Error rate of the random forest model (1000 trees). (D) Variable importance measure plot with horizontal axis as IncNodePurity. (E) Venn diagram showing Lasso and RF sharing four common genes (*CSR3P3*, *MCOLN3*, *SLC25A5*, and *FIBP*). (F) mRNA expression levels of *CSR3P3*, *MCOLN3*, *SLC25A5*, and *FIBP* in AF-VHD and SR-VHD sample groups based on merged datasets.

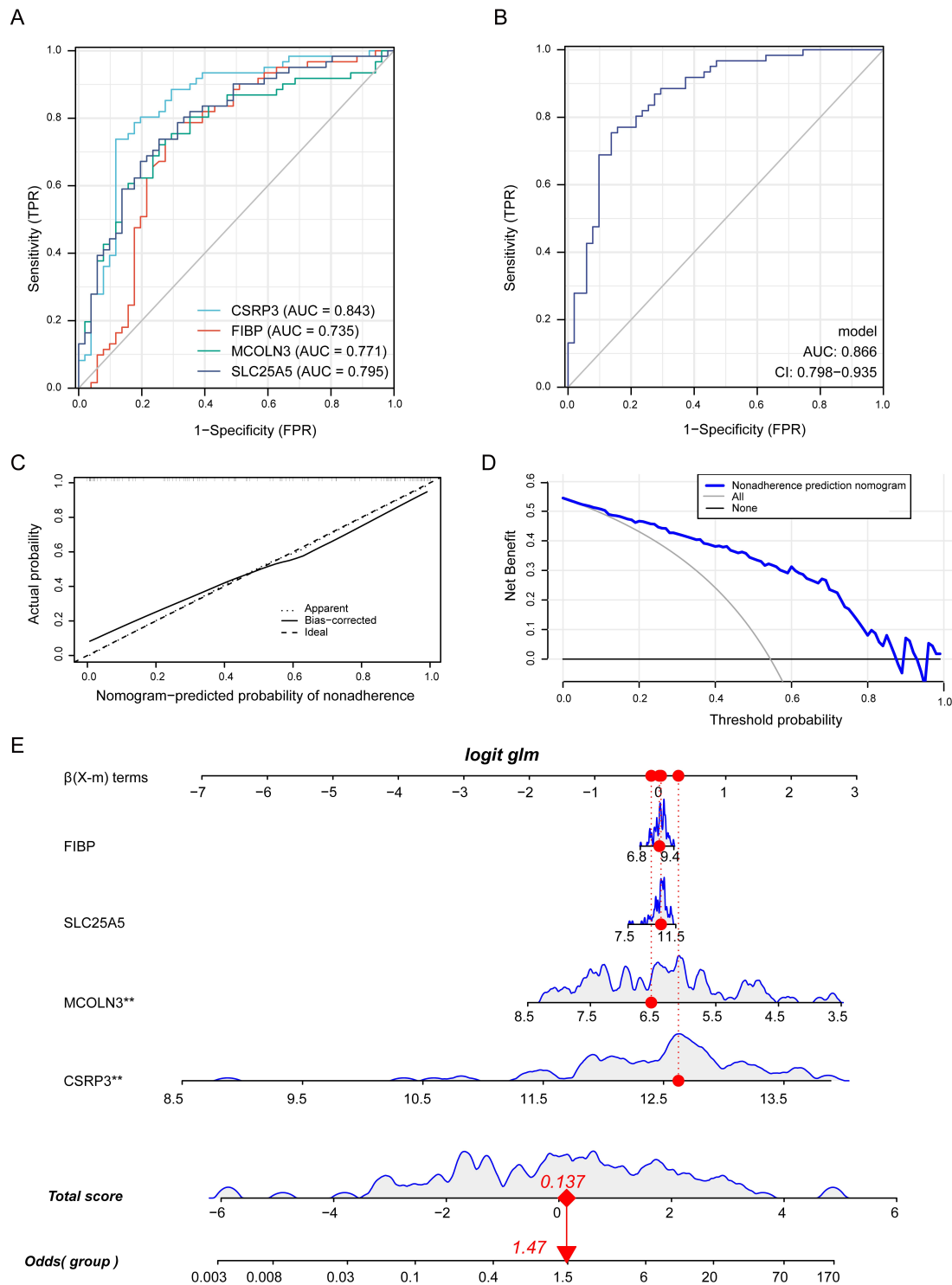


Fig. 9. Identification of crucial genes and evaluation of their clinical significance. (A) ROC analysis for individual *MCOLN3*, *CSRP3*, *SLC25A5*, and *FIBP* genes in AF-VHD vs. SR-VHD in merged datasets. (B) Estimation of clinical diagnostic efficacy of four crucial gene signature (AUC = 0.866, 95% CI = 0.798–0.935) (*Logistic regression model* = $-20.0615 + 1.838 \times CSRP3 + -0.9575 \times MCOLN3 + 0.1821 \times SLC25A5 + 0.1969 \times FIBP$) (ROC, receiver-operating characteristic; AUC, area under curve). (C) Predictive ability of the diagnosis model represented by a calibration curve. (D) Decisions based on the diagnosis model had a better net benefit and broader threshold probability. (E) Construction of a nomogram model based on the four crucial genes. The potential clinic significance of *CSRP3* and *MCOLN3* were statistically significant.

score in combined GSE datasets. The AF-VHD group had higher risk scores compared to the SR-VHD group (Fig. 10B). The *CSRP3* and *MCOLN3* protein expression levels in left atrial appendages from SR and AF patients with VHD were determined. Results indicated that the protein levels of *CSRP3* were higher in the AF-VHD samples than in the SR-VHD controls, whereas the *MCOLN3* expression did not show a significant difference in the two groups (Fig. 10C,D). The HPA tool showed that *CSRP3* was specifically expressed in the heart and skeletal muscle (Fig. 10E). The Single Cell Type Atlas in HPA indicated the expression of *CSRP3* in different cell types. All data were derived from the available published single-cell RNA sequencing analysis (Fig. 10F).

4. Discussion

AF is one of the most common tachyarrhythmias observed in the clinic. It increases patient morbidity and mortality, imposes an economic burden on patients, and seriously affects their quality of life [32]. VHD dramatically increases the risk of AF [33,34]. Nevertheless, the mechanism for development of VHD into AF-VHD is still not completely understood. Therefore, it is essential to investigate the progression of AF-VHD and to identify specific biomarkers and potential therapeutic targets.

Based on previous studies, it has been shown that immune infiltration and atrial fibrosis are involved in the pathophysiological process of AF-VHD [8–10]. However, despite much work in this field, accurate and specific diagnostic biomarkers for AF-VHD are lacking. Our study identified four such biomarkers: *CSRP3*, *MCOLN3*, *SLC25A5*, and *FIBP*. *CSRP3* and *MCOLN3* in particular have important biological and clinical implications.

Microtubule-associated protein *CSRP3*, affiliated with the cysteine-rich protein (*CSRP/CRP*) family, is expressed in both cardiac and muscle tissue. *CSRP3* plays a pivotal role in the development and maintenance of cardiac cytoarchitectural organization [35,36]. It was found to be differentially expressed in AF-VHD and potentially related to myocardial contractility [7]. Evidence has shown that *CSRP3* mutations can result in both hypertrophic cardiomyopathy (HCM) and dilated cardiomyopathy (DCM) in patients [34,35]. HCM and DCM have been shown to be related to the occurrence and development of AF, which indirectly revealed that *CSRP3* might be connected to the development of AF-VHD. Li *et al.* [36] also demonstrated that cardiomyocytes derived from human embryonic stem cells with *CSRP3* deficiency mimic heart failure (HF) 30 days after differentiation, increasing reactive oxygen species generation and exhibiting mitochondrial damage and impaired Ca^{2+} handling. By restoring Ca^{2+} homeostasis, verapamil can trigger an inhibitory effect on HCM and HF, indicating that elevated intracellular Ca^{2+} concentration plays a critical role in the pathogenesis of *CSRP3* deficiency [36]. Our experimental results implied that *CSRP3* is highly expressed in AF-VHD, and the higher the *CSRP3* expression,

the greater the odds of belonging to the AF-VHD group. Additionally, the enrichment analysis revealed that *CSRP3* is related to heart development. Single-cell analysis also showed that *CSRP3* is specifically expressed in myocardial cells. In conclusion, *CSRP3* was shown to be specifically highly expressed in AF-VHD, with a potential clinical significance in the diagnosis of AF-VHD, which may indicate that *CSRP3* is a potential biomarker, as well as a promising therapeutic target, for AF-VHD.

SLC25A5, also known as ANC3 or ANT2, is a member of the mitochondrial carrier subfamily of solute carrier protein genes. *SLC25A5* is highly expressed in high energetic demand organs, such as the heart, kidney, liver, and spleen, contributing to mitochondrial energy metabolism regulation and apoptosis prevention [37,38]. However, the role of *SLC25A5* in AF is unknown. In our study, we found that *SLC25A5* is positively correlated with *CSRP3*, indicating that *SLC25A5* might also participate in the process of myocardial hypertrophy.

MCOLN3, also known as *TRPML3*, is a gene with the highest correlation with AF-VHD among the crucial genes. *MCOLN3/TRPML3* is a cation channel permeable to Ca^{2+} expressed in multiple subcellular compartments with dynamic localization. The findings by Kim *et al.* [39] has demonstrated that Ca^{2+} is released with a robust response when *MCOLN3/TRPML3* is activated intracellularly. It has been previously shown that *MCOLN3* regulates autophagy by specifically interacting with mammalian GABA(A) Receptor-Associated Protein Like 1 (ATG8) homologue GATE16 [40]. Nevertheless, the presence of *MCOLN3* in the heart has been rarely reported. Recently, Düzen *et al.* [41] found that the expression of *MCOLN3* is up-regulated in patients with non-valvular AF (NVAf). It was the first study to reveal the expression pattern of a leukocyte TRP channel gene in NVAf. Our study indicated from the reverse side that the *MCOLN3/TRPML3* expression was reduced in valvular AF (VAF), which proved the expression of *MCOLN3* was associated with AF from another perspective. However, the results showed no difference in *MCOLN3* protein level between the AF-VHD and control groups. This might be due to an insufficient sample size in the study. In summary, *MCOLN3* plays an essential role in the regulation of Ca^{2+} trafficking, which may mediate the development of AF-VHD. However, the detailed action mechanism for *MCOLN3* in AF-VHD remains poorly understood, and further research is needed.

FIBP interacts directly with the fibroblast growth factor 1 (*FGF1*) [42]. *FIBP* is known to be involved in the FGF receptor signaling pathway and platelet aggregation [42,43]. Although few studies have reported on the effect of *FIBP* on AF or VHD, Lu *et al.* [44] demonstrated that *FGF1* might be involved in AF via modification of oxidative stress and sodium/calcium homeostasis, suggesting that *FIBP* may genetically interact with *FGF1* to regulate the development of AF.

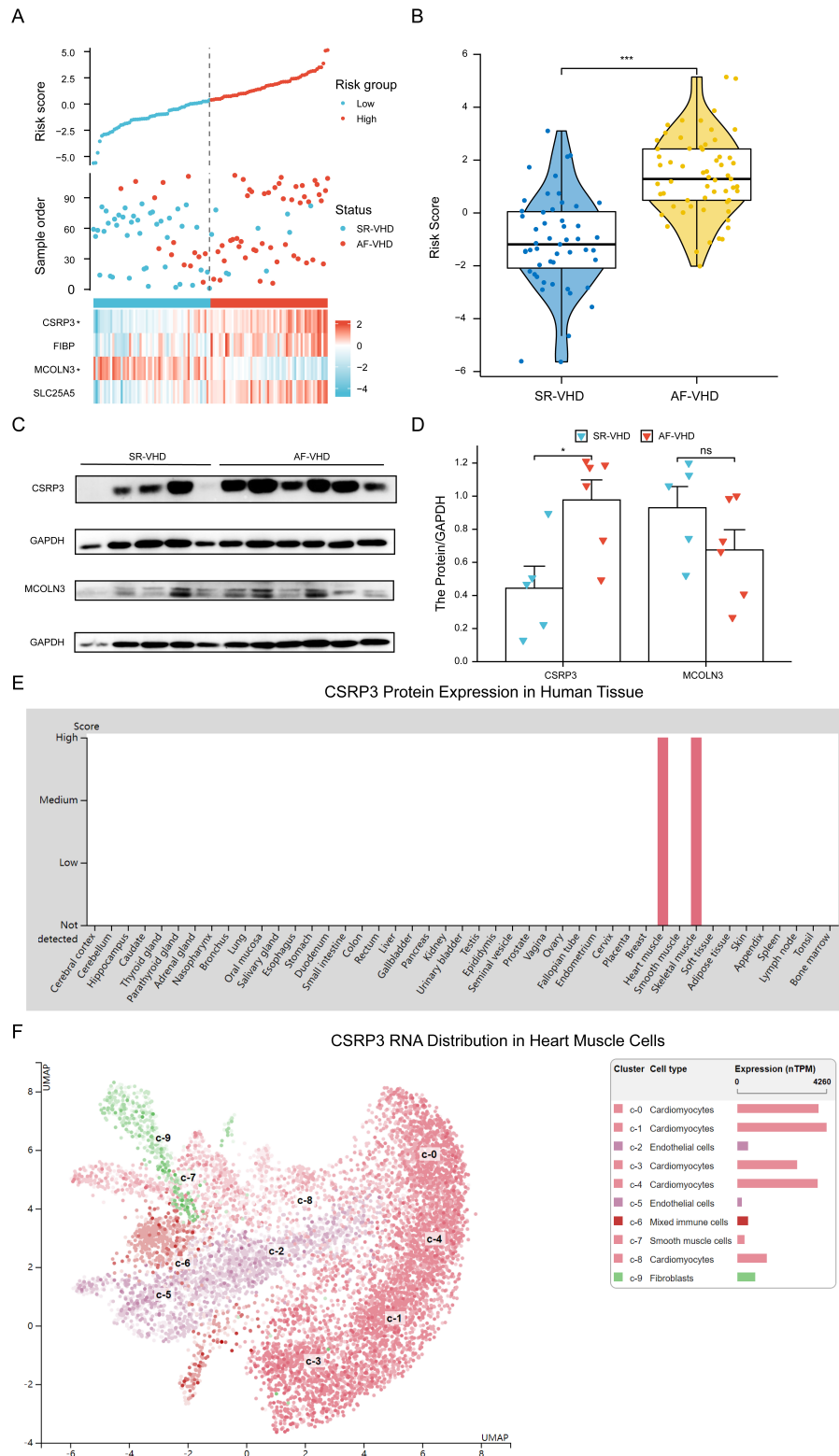


Fig. 10. *CSRP3* is specifically expressed in the heart. (A) Risk score distribution, disease status, and gene expression values of final predictors in the combined datasets. Dotted black line represents the median risk score cutoffs for classifying patients as high- or low-risk. Blue dots represent SR-VHD, and red dots represent AF-VHD. Heatmap showing gene expression values for corresponding sample. (B) Risk score differences between SR-VHD and AF-VHD. (C,D) *CSRP3* and *MCOLN3* protein expression levels were analyzed and quantified in atrial tissue from AF-VHD and SR-VHD patients. (E) In human tissue, *CSRP3* expression was predominantly found in heart and skeletal muscle. (F) The mRNA levels of *CSRP3* expression in different cell types showed that *CSRP3* was mainly expressed in cardiomyocytes.

Our study has some limitations. First, due to the limited sample size derived from the public database, further research with a larger sample size should be conducted to strengthen the conclusion. Second, the biological and molecular functions of these molecules will need to be determined through further experimental studies.

5. Conclusions

In ysummary, four crucial genes (*CSRP3*, *MCOLN3*, *SLC25A5*, and *FIBP*) associated with development of AF-VHD were identified using comprehensive bioinformatics yanalysis. Based on their biological function and clinical value, these genes may be associated with the pathophysiological process of AF-VHD. These findings can facilitate the diagnosis and development of novel therapeutic targets for clinical disorders involving AF-VHD.

Data Availability

The raw datasets were available from the GEO database (<http://www.ncbi.nlm.nih.gov/geo/>; GSE115574, GSE41177, GSE79768 and GSE14975).

Author Contributions

Conception and design—QL, SL, YX and FR; Collection and assembly of data—QL, SL, XL, JH and YF; Data analysis and interpretation—QL, YL and HY; Manuscript writing—QL, SL, XL, JH, YF, HY, YL, CD, SW, YX and FR. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

The authors are accountable for aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The studies involving human participants were reviewed and approved by the ethics committee of Guangdong General Hospital, Guangdong Academy of Medical Sciences (No. GDREC2017111H). The patients/participants provided their written informed consent to participate in this study.

Acknowledgment

This study used data or information from the GEO database. We thanked the GEO database for the provided data.

Funding

This work was supported by grants from the National Natural Science Foundation of China (Nos. 81870254 and 81670314), and the High-level Hospital Construction Plan (No. DFJH201808 and DFJH201925).

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at <https://doi.org/10.31083/j.rcm2307247>.

References

- [1] Esposito G, Piras P, Evangelista A, Nuzzi V, Nardinocchi P, Pannarale G, *et al.* Improving performance of 3D speckle tracking in arterial hypertension and paroxysmal atrial fibrillation by using novel strain parameters. *Scientific Reports*. 2019; 9: 7382.
- [2] Thorolfssdottir RB, Sveinbjornsson G, Sulem P, Helgadottir A, Gretarsdottir S, Benonisdottir S, *et al.* A Missense Variant in PLEC Increases Risk of Atrial Fibrillation. *Journal of the American College of Cardiology*. 2017; 70: 2157–2168.
- [3] Lee SH, Ryu S, Lee J, Seo DC, Kim BJ, Sung K. Association between self-reported physical activity and incident atrial fibrillation in a young Korean population. *Scientific Reports*. 2019; 9: 4222.
- [4] Benjamin EJ, Rice KM, Arking DE, Pfeufer A, van Noord C, Smith AV, *et al.* Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry. *Nature Genetics*. 2009; 41: 879–881.
- [5] Han D, Zhang QY, Zhang YL, Han X, Guo SB, Teng F, *et al.* Gallic Acid Ameliorates Angiotensin II-Induced Atrial Fibrillation by Inhibiting Immunoproteasome-Mediated PTEN Degradation in Mice. *Frontiers in Cell and Developmental Biology*. 2020; 8: 594683.
- [6] d'Arcy JL, Coffey S, Loudon MA, Kennedy A, Pearson-Stuttard J, Birks J, *et al.* Large-scale community echocardiographic screening reveals a major burden of undiagnosed valvular heart disease in older people: the OxVALVE Population Cohort Study. *European Heart Journal*. 2016; 37: 3515–3522.
- [7] Lamirault G, Gaborit N, Le Meur N, Chevalier C, Lande G, Demolombe S, *et al.* Gene expression profile associated with chronic atrial fibrillation and underlying valvular heart disease in man. *Journal of Molecular and Cellular Cardiology*. 2006; 40: 173–184.
- [8] Yan T, Zhu S, Zhu M, Wang C, Guo C. Integrative Identification of Hub Genes Associated With Immune Cells in Atrial Fibrillation Using Weighted Gene Correlation Network Analysis. *Frontiers in Cardiovascular Medicine*. 2021; 7: 631775.
- [9] Li S, Jiang Z, Chao X, Jiang C, Zhong G. Identification of key immune-related genes and immune infiltration in atrial fibrillation disease based on bioinformatics analysis. *Journal of Thoracic Disease*. 2021; 13: 1785–1798.
- [10] Liu Y, Bai F, Tang Z, Liu N, Liu Q. Integrative transcriptomic, proteomic, and machine learning approach to identifying feature genes of atrial fibrillation using atrial samples from patients with valvular heart disease. *BMC Cardiovascular Disorders*. 2021; 21: 52.
- [11] Altenbuchinger M, Schwarzfischer P, Rehberg T, Reinders J, Kohler CW, Gronwald W, *et al.* Molecular signatures that can be transferred across different omics platforms. *Bioinformatics*. 2017; 33: I333–I340.
- [12] Ai D, Wang Y, Li X, Pan H. Colorectal Cancer Prediction Based on Weighted Gene Co-Expression Network Analysis and Variational Auto-Encoder. *Biomolecules*. 2020; 10: 1207.
- [13] Burke MA, Cook SA, Seidman JG, Seidman CE. Clinical and Mechanistic Insights into the Genetics of Cardiomyopathy. *Journal of the American College of Cardiology*. 2016; 68: 2871–2886.
- [14] Pirruccello JP, Bick A, Wang M, Chaffin M, Friedman S, Yao J, *et al.* Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy.

Nature Communications. 2020; 11: 2254.

- [15] Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004; 20: 307–315.
- [16] Yu G, Wang L, Han Y, He Q. ClusterProfiler: an R Package for Comparing Biological Themes among Gene Clusters. *OMICS: A Journal of Integrative Biology*. 2012; 16: 284–287.
- [17] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, *et al*. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*. 2019; 10: 1523.
- [18] Innis SE, Reinaltt K, Civelek M, Anderson WD. GSEAplot: a Package for Customizing Gene Set Enrichment Analysis in R. *Journal of Computational Biology*. 2021; 28: 629–631.
- [19] Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*. 2005; 4: Article17.
- [20] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008; 9: 559.
- [21] Hu X, Bao M, Huang J, Zhou L, Zheng S. Identification and Validation of Novel Biomarkers for Diagnosis and Prognosis of Hepatocellular Carcinoma. *Frontiers in Oncology*. 2020; 10: 541479.
- [22] Lai S, Hua X, Gao R, Zeng L, Song J, Liu J, *et al*. Combinational Biomarkers for Atrial Fibrillation Derived from Atrial Appendage and Plasma Metabolomics Analysis. *Scientific Reports*. 2018; 8: 16930.
- [23] Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research*. 2019; 47: W234–W241.
- [24] Yang W, Niu L, Zhao X, Duan L, Li Y, Wang X, *et al*. Development and validation of a survival model based on autophagy-associated genes for predicting prognosis of hepatocellular carcinoma. *American Journal of Translational Research*. 2020; 12: 6705–6722.
- [25] Yang J, Chen C, Jin X, Liu L, Lin J, Kang X, *et al*. Wfs1 and Related Molecules as Key Candidate Genes in the Hippocampus of Depression. *Frontiers in Genetics*. 2021; 11: 589370.
- [26] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2010; 33: 1–22.
- [27] Lawrence RL, Wood SD, Sheley RL. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*. 2006; 100: 356–362.
- [28] Kumar RD, Swamidass SJ, Bose R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. *Nature Genetics*. 2016; 48: 1288–1294.
- [29] Iasonos A, Schrag D, Raj GV, Panageas KS. How to Build and Interpret a Nomogram for Cancer Prognosis. *Journal of Clinical Oncology*. 2008; 26: 1364–1370.
- [30] Li X, Deng C, Xue Y, Yang H, Wei W, Liu F, *et al*. High hydrostatic pressure induces atrial electrical remodeling through angiotensin upregulation mediating FAK/Src pathway activation. *Journal of Molecular and Cellular Cardiology*. 2020; 140: 10–21.
- [31] Li C, Liu T, Liu Y, Zhang J, Zuo D. Prognostic value of tumour microenvironment-related genes by TCGA database in rectal cancer. *Journal of Cellular and Molecular Medicine*. 2021; 25: 5811–5822.
- [32] Piccini JP, Fauchier L. Rhythm control in atrial fibrillation. *The Lancet*. 2016; 388: 829–840.
- [33] Thomas KL, Jackson LR, Shrader P, Ansell J, Fonarow GC, Gersh B, *et al*. Prevalence, Characteristics, and Outcomes of Valvular Heart Disease in Patients with Atrial Fibrillation: Insights from the ORBIT-AF (Outcomes Registry for Better Informed Treatment for Atrial Fibrillation). *Journal of the American Heart Association*. 2017; 6: e006475.
- [34] Chen YT, Wang J, Wee ASY, Yong QW, Tay ELW, Woo CC, *et al*. Differential MicroRNA Expression Profile in Myxomatous Mitral Valve Prolapse and Fibroelastic Deficiency Valves. *International Journal of Molecular Sciences*. 2016; 17: 753.
- [35] Rashid MM, Runci A, Polletta L, Carnevale I, Morgante E, Foglio E, *et al*. Muscle LIM protein/CSRP3: a mechanosensor with a role in autophagy. *Cell Death Discovery*. 2015; 1: 15014.
- [36] Li X, Lu W, Li Y, Wu F, Bai R, Ma S, *et al*. MLP-deficient human pluripotent stem cell derived cardiomyocytes develop hypertrophic cardiomyopathy and heart failure phenotypes due to abnormal calcium handling. *Cell Death & Disease*. 2019; 10: 610.
- [37] Stepien G, Torroni A, Chung AB, Hodge JA, Wallace DC. Differential expression of adenine nucleotide translocator isoforms in mammalian tissues and during muscle cell differentiation. *Journal of Biological Chemistry*. 1992; 267: 14592–14597.
- [38] Cl  men  on B, Babot M, Tr  z  guet V. The mitochondrial ADP/ATP carrier (SLC25 family): Pathological implications of its dysfunction. *Molecular Aspects of Medicine*. 2013; 34: 485–493.
- [39] Kim SW, Kim DH, Park KS, Kim MK, Park YM, Muallem S, *et al*. Palmitoylation controls trafficking of the intracellular Ca²⁺ channel MCOLN3/TRPML3 to regulate autophagy. *Autophagy*. 2019; 15: 327–340.
- [40] Choi S, Kim HJ. The Ca²⁺ channel TRPML3 specifically interacts with the mammalian ATG8 homologue GATE16 to regulate autophagy. *Biochemical and Biophysical Research Communications*. 2014; 443: 56–61.
- [41] D  zen IV, Yavuz F, Vuruskan E, Saracoglu E, Poyraz F, G  ks  l  k H, *et al*. Leukocyte TRP channel gene expressions in patients with non-valvular atrial fibrillation. *Scientific Reports*. 2017; 7: 9272.
- [42] Kolpakova E, Wiedlocha A, Stenmark H, Klingenberg O, Falnes PO, Olsnes S. Cloning of an intracellular protein that binds selectively to mitogenic acidic fibroblast growth factor. *Biochemical Journal*. 1998; 336: 213–222.
- [43] Fr  bel J, Cadeddu R, Hartwig S, Bruns I, Wilk CM, K  ndgen A, *et al*. Platelet Proteome Analysis Reveals Integrin-dependent Aggregation Defects in Patients with Myelodysplastic Syndromes. *Molecular & Cellular Proteomics*. 2013; 12: 1272–1280.
- [44] Lu YY, Cheng CC, Huang SY, Chen YC, Kao YH, Lin YK, *et al*. Fibroblast Growth Factor 1 Reduces Pulmonary Vein and Atrium Arrhythmogenesis via Modification of Oxidative Stress and Sodium/Calcium Homeostasis. *Frontiers in Cardiovascular Medicine*. 2022; 8: 813589.