

Original Research

Identification of Potential Biomarkers Associated with Dilated Cardiomyopathy by Weighted Gene Coexpression Network Analysis

Qixin Guo^{1,*},†, Qiang Qu^{1,†}, Luyang Wang^{1,†}, Shengen Liao¹, Xu Zhu¹, Anning Du¹,
Qingqing Zhu¹, Iokfai Cheang¹, Rongrong Gao¹, Xinli Li^{1*}¹Department of Cardiology, The First Affiliated Hospital of Nanjing Medical University, 210029 Nanjing, Jiangsu, China*Correspondence: Guoqixin@stu.njmu.edu.cn (Qixin Guo); xinli3267@njmu.edu.cn (Xinli Li)

†These authors contributed equally.

Academic Editor: Rajesh Katare

Submitted: 30 June 2022 Revised: 13 July 2022 Accepted: 18 July 2022 Published: 17 August 2022

Abstract

Background: Dilated cardiomyopathy (DCM) is one of the main causes of systolic heart failure and frequently has a genetic component. The molecular mechanisms underlying the onset and progression of DCM remain unclear. This study aimed to identify novel diagnostic biomarkers to aid in the treatment and diagnosis of DCM. **Method:** The Gene Expression Omnibus (GEO) database was explored to extract two microarray datasets, GSE120895 and GSE17800, which were subsequently merged into a single cohort. Differentially expressed genes were analyzed in the DCM and control groups, followed by weighted gene coexpression network analysis to determine the core modules. Core nodes were identified by gene significance (GS) and module membership (MM) values, and four hub genes were predicted by the Lasso regression model. The expression levels and diagnostic values of the four hub genes were further validated in the datasets GSE19303. Finally, potential therapeutic drugs and upstream molecules regulating genes were identified. **Results:** The turquoise module is the core module of DCM. Four hub genes were identified: GYPC (glycophorin C), MLF2 (myeloid leukemia factor 2), COPS7A (COP9 signalosome subunit 7A) and ARL2 (ADP ribosylation factor like GTPase 2). Subsequently, Hub genes showed significant differences in expression in both the dataset and the validation model by real-time quantitative PCR (qPCR). Four potential modulators and seven chemicals were also identified. Finally, molecular docking simulations of the gene-encoded proteins with small-molecule drugs were successfully performed. **Conclusions:** The results suggested that ARL2, MLF2, GYPC and COPS7A could be potential gene biomarkers for DCM.

Keywords: weighted gene coexpression network analysis; dilated cardiomyopathy

1. Introduction

Dilated cardiomyopathy (DCM) is a clinical phenotype that manifests as heart failure due to a combination of genomic [1,2], epigenetic [3] and external factors. The prevalence of DCM varies from 1/250 to 1/2500 [4], occurring in greater proportions than ischemic cardiomyopathy [5], and is the leading cause of heart failure. Many new drugs and devices are being used to improve the long-term prognosis of patients, such as angiotensin receptor neprilysin inhibitor (ARNI) [6], sodium-glucose cotransporter 2 (SLGT2) inhibitors [7], and left heart assist devices. However, clinical decision-making [8] in DCM is mainly based on heart failure without taking into account the heterogeneity [9] of DCM.

Previous studies have shown that mutations in genes encoding cytoskeletal, myosin, mitochondrial, bridging granule, nuclear membrane, and RNA-binding proteins are associated with DCM [5]. However, there is considerable heterogeneity in the genetic testing panel, especially with the development of whole-genome sequencing, which has seen many variations, making it difficult to distinguish pathogenic and nonpathogenic variants [10]. Abnormal genetic variants defined by individual studies may be normal

in other populations due to sample size, ethnicity, and other factors. Currently, genetic diagnosis is used to rule out genes carried by other family members of the proband, but some carriers pass the genes on without any cardiac events. The fundamental mechanisms of DCM remain poorly understood. Finding efficient and low-cost diagnostic methods to identify hub genes has been a major challenge [11].

Thanks to the development of high-throughput technologies, transcriptional analysis based on multiple datasets has been used to determine the pathological mechanisms of diseases. Disease pathogenesis and progression are not caused by a single gene but by synergistic effects in a complex network [12]. Complementary to this is weighted gene coexpression network analysis (WGCNA), a widely used systems bioinformatics technique to assess associations between genomic and external sample features by constructing scale-free gene coexpression networks.

Molecular targeted therapies have been widely used in oncological diseases to assist physicians' decisions based on the expression of star molecules, but have not yet been introduced to the clinic for use in cardiovascular diseases. This paper is dedicated to investigate the possibility of molecular targets for the diagnosis and treatment of clinical



diseases, to find a new marker that can be used for general screening of DCM or to improve the prognosis of patients after intervention. Therefore, this study collected DCM-related genes and used multiple databases to further search for hub genes and targets for potential accurate treatment or diagnosis of DCM.

2. Methods

2.1 Data Processing

The raw data of two eligible microarray datasets (GSE120895 and GSE17800) based on platform GPL570 were downloaded from the GEO database (87 patients with DCM and 16 controls after data merging). Moreover, GSE19303 (73 patients with DCM and 8 controls) was used as the validation set. The Limma package was used to screen DEGs between DCM and controls, followed by data normalization [13]. Bias and variability of the datasets were removed using the Combat function in the sva package. $p < 0.05$ adjusted by the false discovery rate (FDR) was considered significant. The data processing procedure of our research is illustrated in the workflow (Fig. 1).

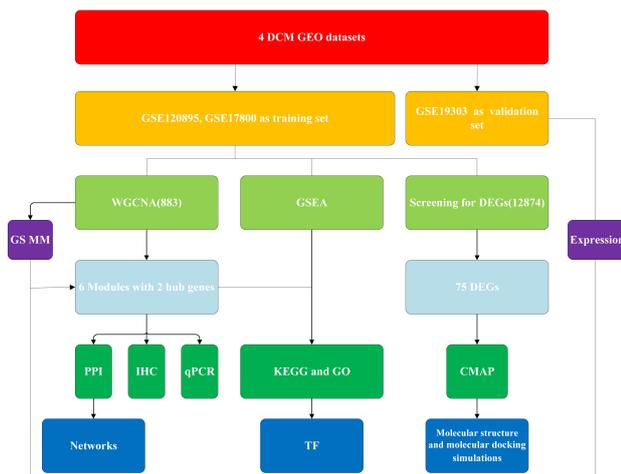


Fig. 1. The workflow of our research. DCM, dilated cardiomyopathy; GEO, Gene Expression Omnibus; GO, Gene Ontology; PPI, protein-protein interaction; CMAP, Connectivity Map; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, gene set enrichment analysis; TF, transcription factor; DEG, differentially expressed gene; WGCNA, weighted gene coexpression network analysis; IHC, immunohistochemical; qPCR, quantitative polymerase chain reaction.

2.2 Weighted Gene Correlation Network Analysis (WGCNA)

A coexpression network of 884 genes with commonly upregulated or downregulated expression in DCM/control was constructed using WGCNA, which is a widely used systems biology approach that helps identify the relation-

ship between genes and disease phenotypes. The main processes were as follows: (1) the hclust function was used for hierarchical clustering analysis; (2) a power of $\beta = 7$ was selected according to the scale-free topology criterion; (3) genes with coexpression relationships were grouped through gene network connectivity, the minimum module size was set to 30, and each module was assigned a unique color label; (4) the relationship between modules and phenotypes was calculated to identify biologically meaningful modules with Pearson correlation analysis; (5) GO annotation and KEGG pathway enrichment analysis were performed for these functional modules; and (6) a PPI network was constructed.

2.3 Hub Gene Identification and Validation

In WGCNA, module membership (MM) is defined as the correlation of the module eigengene and the gene expression profile, while gene significance (GS) is defined as the correlation between the gene and the trait. Among the modules of interest in our study, $GS > 0.3$ and $MM > 0.8$ were defined as hub genes in the candidate gene modules.

The least absolute contraction and selection operator (LASSO) method [14], which is suitable for high-dimensional data restoration, was used to select the optimal risk factor prediction characteristics from gene datasets. The A radiomics score (Rad-score) of each gene was calculated by selecting a linear combination of features, which are weighted by their respective coefficients. Then, the variables selected by the LASSO method were used to obtain the final factors for establishing the model. Harrell's C index and the area under the curve (AUC) were measured to quantify the discrimination performance of the model based on the experimental set and validation set.

2.4 Gene Set Enrichment Analysis (GSEA)

GSEA is a computational method used to assess whether a predefined set of genes displays statistical significance and consistency differences between two biological states [15]. The c5.all.v7.4.entrez and c2.cp.kegg.v7.4.entrez datasets in the MsigDB database were used as reference gene sets, and the clusterprofile package was utilized to perform GSEA with integrated gene expression data. $p < 0.05$ was considered significant.

2.5 Potential Transcription Factors (TFs) and Drugs

The promoter sequences of target genes were searched using the University of California Santa Cruz (UCSC) and National Center for Biotechnology Information databases (NCBI). Potential binding transcription factors were searched using the JASPAR database and further screened according to transcription direction and related levels. Gene expression profiling interactive analysis (GEPIA) was used to verify the correlations and identify potential transcription factors. Finally, the structural information combined with the interval key structural domain was presented using UniProt.

The CMap database establishes links between genes, compounds and diseases based on similar and opposite gene expression profiles. In this study, the DEGs of the DCM and control groups were grouped according to expression differences. Then, the DEGs were loaded to the “Query” page. In this study, drugs with connectivity scores >90 or <-90 and p values less than 0.05 were selected as drug candidates. The 2D and 3D structures of the candidate compounds were obtained from PubChem.

2.6 Quantitative Reverse Transcription Polymerase Chain Reaction (qRT-PCR) Verification

The study was approved by the Medical Ethics Committee of Nanjing Medical University, and a total of 16 patients were recruited to provide blood samples, and informed consent was obtained from patients before participation. Expression levels of 4 hub genes were verified in 8 DCM blood samples and 8 normal blood samples (**Supplementary Table 1**). Total RNA was extracted using TRIzol reagent according to the manufacturer’s instructions. Total RNA was reverse transcribed into cDNA by PrimeScript RT Master Mix (TaKaRa, Japan) after measuring the corresponding concentration. Then qRT-PCR was performed using Power SYBR Green PCR Master Mix (No. A25742; Thermo Fisher Scientific, Waltham, MA, USA). Finally, the relative expression levels of miRNAs and target genes were calculated according to the $2^{-\Delta\Delta Ct}$ method. GAPDH levels were used to normalize mRNA expression levels. The procedure for PCR is shown below: In brief, samples were incubated at 95 °C for 5 min, followed by 40 cycles at 95 °C for 10 s and finally at 60 °C for 20 s. The sequences of the primers are shown below: 5'-CAACGAATTTGGCTACAGCA-3' and 5'-AGGGGAGATTCAGTGTGGTG-3' for GAPDH, 5'-GGACATCGACACCATCTCCC-3' and 5'-TAGTTCCGCCAGTAGGACCG-3' for ARL2; 5'-TCCCTTTGCTATTCACCGTCA-3' and 5'-CACCCGACATTCCCAGCATC-3' for MLF2; 5'-GCCGGATGGCAGAATGGAG-3' and 5'-GGAGGGAGACTAGGACGATGG-3' for GYPC; 5'-ATGAGTGCGGAAGTGAAGGTG-3' and 5'-GCTCTTAACATTGGGCATGTC-3' for COPS7A.

2.7 Protein Expression Analysis in DCM

The Human Protein Atlas was used to validate the immunohistochemistry of potential target genes. This database facilitates the systematic study of transcriptome and gene pathology expression of coding genes in different tissue types. Staining of core gene proteins in human myocardial tissue based on immunohistochemical techniques. In addition, tissue types and staining levels were retrieved from the database to analyze the quality of the data to interpret the results [16,17].

2.8 Active Components-Targets Docking

We retrieved the 3D structure of the receptor from the Uniprot and RCSB protein databases, and the corresponding simplified molecular-input line-entry system (SMILES) for the receptor ligand was obtained from the PubChem database. After the ligand energy was adjusted to a minimum by using Chem 3D software 18.0 (PerkinElmer, Waltham, MA, USA), the crystals were imported into Py-mol 2.4.0 software (Schrödinger, L. & DeLano, W., CA, USA) for dehydration, hydrogenation and ligand separation, followed by AutoDockTools-1.5.7 [18] to construct docking grid boxes for each target. Docking is done with Autodock Vina 1.1.2 (The Scripps Research Institute, San Diego, CA, USA) [19].

3. Results

3.1 Identification of Differentially Expressed Genes (DEGs)

To identify DEGs associated with DCM, we screened the integrated normalized data and obtained 12,748 genes after adjusting for batch effects. As shown in the volcano map (Fig. 2), compared to controls, there were 75 DEGs in the DCM sample, of which 39 were upregulated and 36 were downregulated. The 20 most upregulated and 20 downregulated genes in DCM are visualized in the heatmap.

3.2 Construction of a Coexpression Network of DCM-Related Genes

It is reasonable to consider that certain genes with similar expression patterns may perform their functions as a global network and participate in similar biological processes. The 883 DEGs with absolute deviation were used for WGCNA. A total of six gene modules were obtained, which are represented by branches of the clustering tree, and different colors are shown in Fig. 3. In addition, modular enrichment analysis showed that primary enrichment in biological process (BP) corresponding to each module included mitotic DNA damage checkpoint, neutrophil degranulation, protein-containing complex disassembly and muscle system process. Primary enrichments in cell component (CC) involved integral components of endoplasmic reticulum membrane, tertiary granule and mitochondrial inner membrane. Primary enrichments in molecule function (MF) consisted of E-box binding and structural constituent of ribosome. Primary enrichments in the KEGG pathway were chemical carcinogenesis - reactive oxygen species and diabetic cardiomyopathy (Fig. 4). The PPI network of DEGs was constructed with STRING and visualized with Cytoscape (3.8.2). As shown in the **Supplementary Fig. 1**, the PPI network contained nodes and edges in a circular distribution. The nine most important modules were identified by the Cytoscape plugin MCODE.

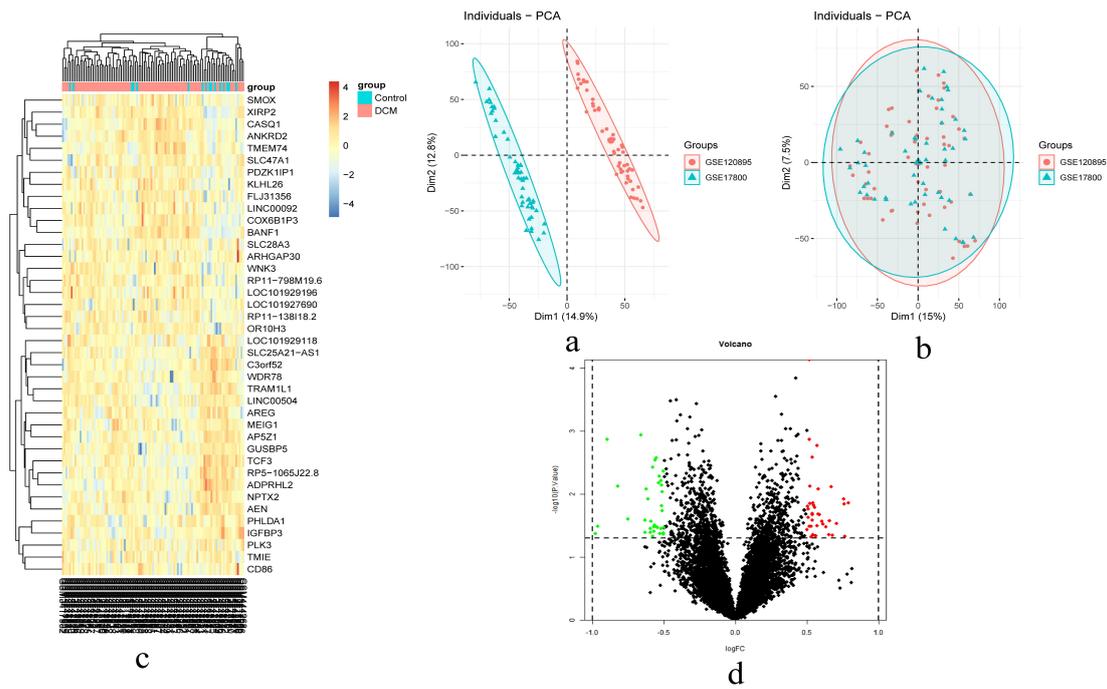


Fig. 2. The result of DEGs identification. (a) Principal component analysis (PCA) before batch effects removal. (b) PCA after batch effects removal. (c) A heatmap of 20 most up-regulated and 20 most down-regulated genes. (d) Volcano plot of the DCM-Control. Abbreviations: DCM, dilated cardiomyopathy; logFC, log₂ fold-change.

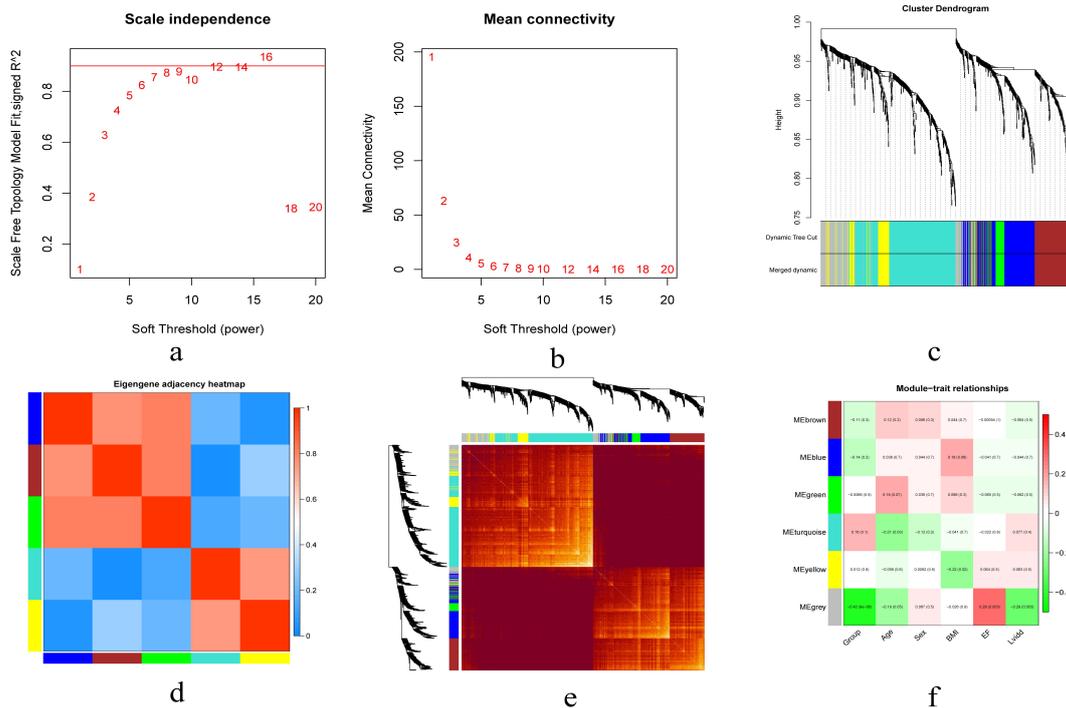


Fig. 3. Construction of gene co-expression modules. (a,b) Analysis of network topology for various soft-thresholding powers. (c) Hierarchical cluster dendrogram of DCM-related genes based on one dissimilarity measure. (d) Hierarchical cluster heatmap of the adjacencies in the eigengene network. (e) Topology overlay heat map: Both rows and columns indicate individual genes, and dark yellow and red indicate a high degree of topological overlap. (f) Module-phenotype associations. Each row corresponds to a module eigengene, and each column corresponds to a clinical feature. Abbreviations: DCM, dilated cardiomyopathy; BMI, body mass index; EF, ejection fraction; LVIDd, left ventricular internal diameter at end-diastole.

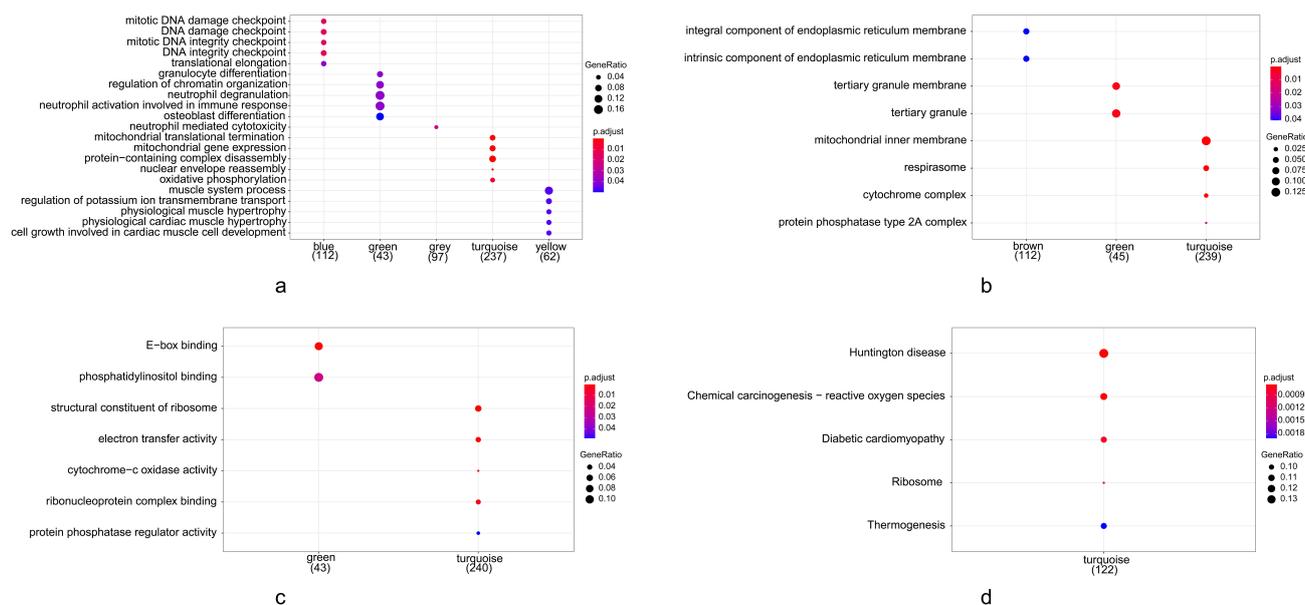


Fig. 4. GO annotation and KEGG pathway enrichment analysis. (a) Biological process of module. (b) Cell component of module. (c) Molecule function of module. (d) KEGG pathway of module. Abbreviations: GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Gene and Genomes.

3.3 Correlation of Coexpressed Genes and Modules with Phenotype

Correlations between the modules of the DCM and control groups and the phenotypes of the six clinics (group, age, sex, body mass index, ejection fraction, left ventricular internal diameter at end-diastole) were calculated, and their corresponding p values were analyzed for magnitude. Except for the gray module, the blue module ($r = -0.14$) had the strongest negative correlation, while the turquoise module ($r = 0.16$) had the strongest positive correlation (Fig. 3).

3.4 Screening and Validation of Hub Genes

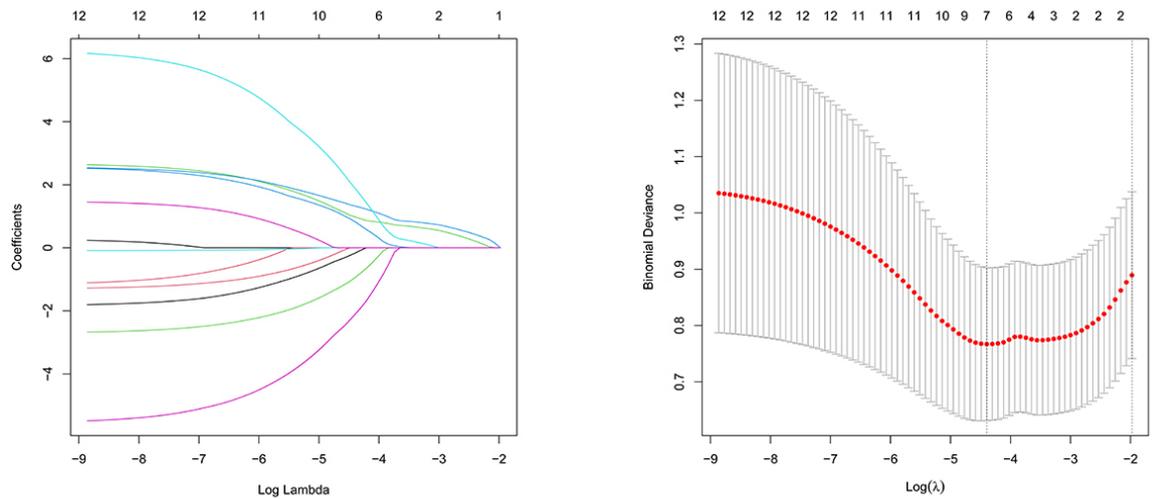
Based on GS and MM values (Supplementary Fig. 2), 12 genes in the yellow module were identified as hub genes RPS15, OXA1L, MLF2, GYPC, GNB2L1, EMC4, COPS7A, BANF1, ATP5G2, ARL2, AP2S1 and AK2. Through LASSO regression, four variables, GYPC (glycophorin C), MLF2 (myeloid leukemia factor 2), COPS7A (COP9 signalosome subunit 7A) and ARL2 (ADP ribosylation factor like GTPase 2), were ultimately identified and used to construct the model. ROC curve analysis was performed on the regression model to predict DCM in the training set with an area under the curve (AUC) of 0.853. To further test the diagnostic efficacy, we used the GSE19303 datasets as the validation set, and the AUC was 0.89, indicating that the model gene had a high diagnostic value. The C-index for the prediction model of the cohort was 0.90, and it was 0.85 through bootstrapping validation, demonstrating the model's good discrimination. In addition, we found that there were significantly upregulated in the expression levels of GYPC ($p < 0.0001$), COPS7A ($p = 0.011$), ARL2 (p

< 0.0001) and MLF2 ($p < 0.0001$) in validation set. These findings suggest that these four genes are highly associated with the occurrence of DCM and could be used as further detectable biomarkers. Moreover, We compared the mRNA expression levels of the 4 hub genes in DCM tissues and paired normal tissues. As shown in Fig. 5, compared with their expression in paired normal samples, the expression of ARL2 ($p = 0.012$), MLF2 ($p < 0.0001$), GYPC ($p < 0.0001$) and COPS7A ($p = 0.026$) was significantly increased in DCM samples. In addition, we observed hub gene protein expression changes in myocardial tissue by immunohistochemistry of tissue sections from the Human Protein Atlas database (Fig. 6).

3.5 Transcription Factors and Drug-Molecules-Targets Docking Analysis

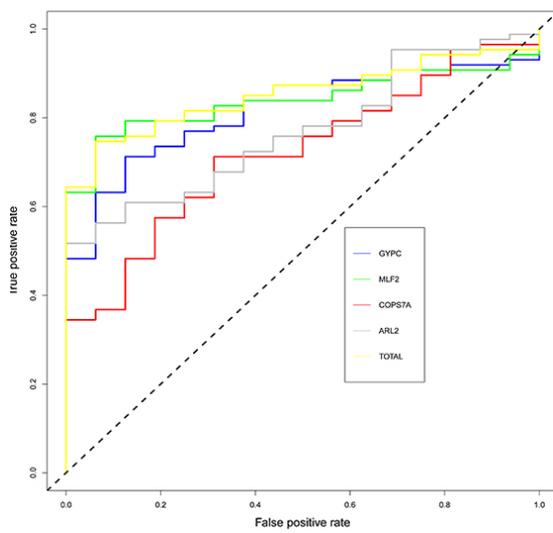
After validating the JASPAR preselected transcription factors by gene expression correlation ($p < 0.05$ and $R > 0.5$), the following results were obtained that the transcription factor upstream of ARL2 was CCCTC-binding factor (CTCF), GYPC was zinc finger protein 384 (ZNF384), COPS7A was zinc finger and BTB domain containing 7A (ZBTB7A) and MLF2 was Sp2 transcription factor (SP2) (Fig. 7).

To find drugs for DCM, we searched the Cmap database and subsequently obtained seven drug candidates, namely, trihexyphenidyl, meclufenamic acid, daunorubicin, simvastatin, doxorubicin, dirithromycin and spiperone (Table 1). The 2D (Supplementary Fig. 3) and 3D structures of these drugs were provided by PubChem, and their corresponding active human validated targets were

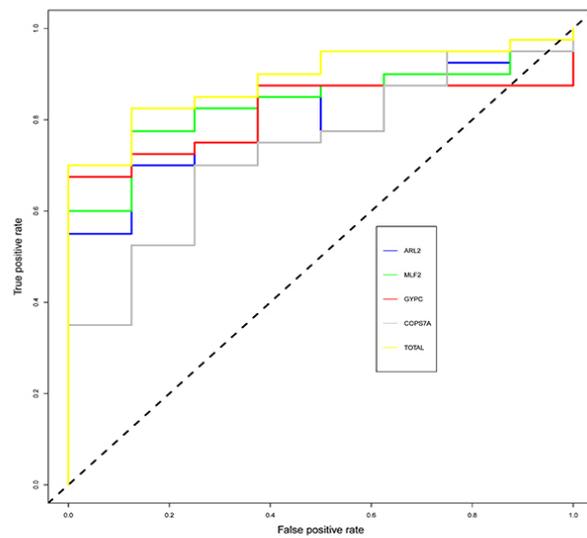


a

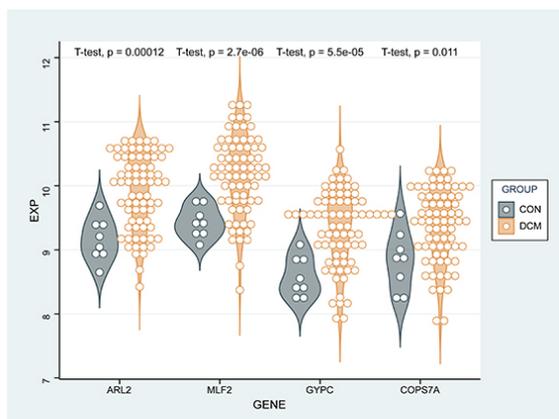
b



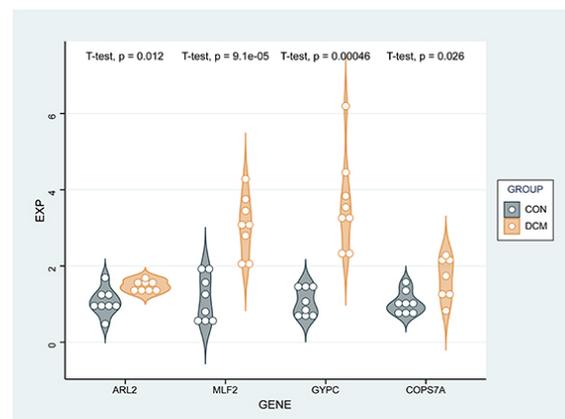
c



d



e



f

Fig. 5. Screening and verification of key genes. (a,b) LASSO model. (c,d) ROC curves for training set and validation set. (e,f) Expression of hub genes in the validation set and qPCR. Abbreviations: LASSO, least absolute shrinkage and selection operator; ROC, receiver operator characteristic; qPCR, quantitative polymerase chain reaction.

Table 1. Seven most significant small molecule chemicals.

Cmap name	Mean	Enrichment	Specificity	Percent non-null	Canonical SMILES	Targetname
Trihexyphenidyl	0.666	0.946	0.0056	100	<chem>C1CCC(CC1)C(CCN2CCCCC2)(C3=CC=CC=C3)O</chem>	NULL
Meclofenamic acid	0.543	0.796	0	100	<chem>CC1=C(C(=C(C=C1)Cl)NC2=CC=CC=C2C(=O)O)C1</chem>	PTGS1, TTR, AKR1C3, PTGS2, AKR1C1, AKR1C2, ABCB11
Daunorubicin	-0.623	-0.846	0.0175	100	<chem>CC1C(C(CC(O1)OC2CC(C3=C2C(=C4C(=C3O)C(=O)C5=C(C4=O)C(=CC=C5)OC)O)(C(=O)C)O)N)O</chem>	BLM, CBF, GNAI1, HDAC1, HDAC10, HDAC11, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HTT, MAPT, POLK, RGS12, RUNX1, TOP1, TOP2A, USP1, YAP1
Simvastatin	-0.581	-0.822	0	100	<chem>CCC(C)(C)C(=O)OC1CC(C=C2C1C(C(C=C2)C)CCC3CC(CC(=O)O3)O)C</chem>	BSLCO1B1, NR2E3, MDM4, MDM2, IDH1, ICMT, HMGCR, CYP3A4, CYP2D6, CYP2C9, CYP2C8, CYP2C19, CYP1A2, ABCB11
Doxorubicin	-0.65	-0.845	0.1353	100	<chem>CC1C(C(CC(O1)OC2CC(CC3=C2C(=C4C(=C3O)C(=O)C5=C(C4=O)C(=CC=C5)OC)O)(C(=O)CO)O)N)O</chem>	ABCC3, ABCC4, BCL2, BCL2L1, C1SD1, CYP1A2, CYP2C19, CYP2C9, CYP2D6, DHCR7, EBP, EPAS1, HIF1A, POLK, PPM1D, SLCO1B3, TOP1, TOP2A, TOP2B, TP53
Dirithromycin	-0.559	-0.826	0	100	<chem>CCC1C(C2C(C(C(C(C(C(C(=O)O1)C)OC3CC(C(C(O3)C)O)(C)OC)C)OC4C(C(CC(O4)C)N(C)C)O)(C)O)C)NC(O2)COCCOC)C(C)O</chem>	PTGER2, NLRP1
Spiiperone	-0.529	-0.873	0.0614	100	<chem>C1CN(CCC12C(=O)NCN2C3=CC=CC=C3)CCCC(=O)C4=CC=C(C=C4)F</chem>	ACBC11, ADRA1A, AVPR1A, CHRM1, CHRM4, CHRM5, CRHBP, CRHR2, DRD1, DRD2, DRD3, DRD4, GNA15, HCRTR1, HTR1A, HTR1D, HTR2A, HTR2B, HTR6, HTR7, MDM2, MDM4, OXTR, TAAR1, TRHR

Canonical SMILE, Internationally recognized atomic structure; Targetname, Validated targets with activity.

identified. The spatial structure of the gene and the spatial structure of the drug can be used for further molecular docking simulations to seek possible mechanisms of occurrence.

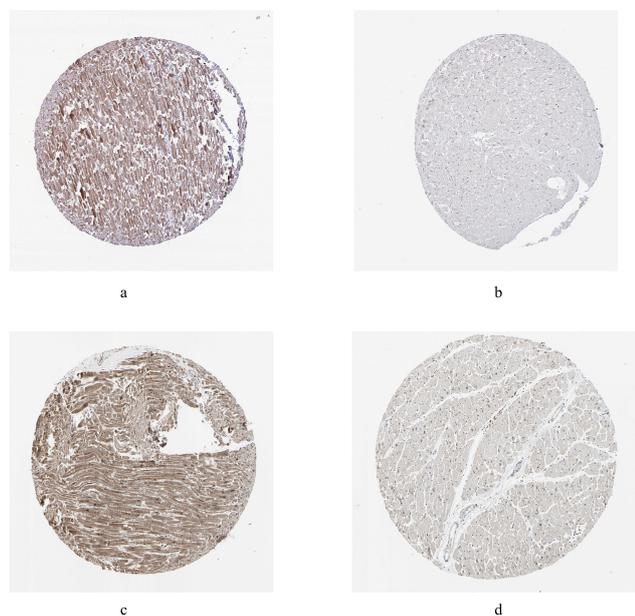


Fig. 6. Immunohistochemistry of the hub genes based on the Human Protein Atlas database (HPAD). (a) Protein levels of ARL2 in myocardial tissue. (b) Protein levels of COPS7A in myocardial tissue. (c) Protein levels of MLF2 in myocardial tissue. (d) Protein levels of GYPC in myocardial tissue.

Molecular docking simulations were used to delve into the possible therapeutic mechanisms of these drugs. The binding energy between two counterparts was calculated to predict their affinity. Binding energies below 0 indicated that the two molecules bind spontaneously, with smaller binding energies leading to a more stable conformation. The 3D structure of MLF2 was not available in the Protein Data Bank and ligands corresponding to GYPC and COPS7A were not suitable for molecular docking simulation. The binding energies of ARL2 with Trihexyphenidyl, Meclofenamic acid, Daunorubicin, Simvastatin, Doxorubicin, Dirithromycin and Spiperone were -6.0 , -6.7 , -7.2 , -7.6 , -7.4 , -6.7 and -5.0 kcal/mol, respectively. Fig. 8 detailed the local structure of molecular docking. It has been illustrated that Doxorubicin acts as an induction target for DCM as positive controls [20]. From the point of view of binding energy this suggested that Simvastatin may be a target for the treatment of DCM (the binding energy of the protein is lower than that of the positive control).

3.6 Gene Set Enrichment Analysis

Gene set enrichment analysis showed that, compared to control samples, the DCM group was significantly enriched in cell component (CC) such as pseudopodium, luminal side of membrane, respiratory chain complex IV,

MHC protein complex and high-density lipoprotein particle; in molecule function (MF) such as chemokine receptor binding, chemokine activity, glutathione binding, oligopeptide binding, MHC class II protein complex binding and MHC protein complex binding; in signaling pathways such as Chemical carcinogenesis - DNA adducts, Drug metabolism - cytochrome P450, Glyoxylate and dicarboxylate metabolism and Asthma; and in reactome such as Chondroitin sulfate biosynthesis, CS/DS degradation, Defective B3GAT3 causes JDSSDHD and Defective B3GALT6 causes EDSP2 and SEMDJL1 (Fig. 9 and Table 2).

4. Discussion

DCM is one of the leading causes of cardiac insufficiency. Despite the increasing number of improved treatments, due to the aging population structure, the proportion of disease occurrence and the number of illnesses is increasing, especially in China. DCM has a serious impact on the quality of life of patients and causes a heavy social and economic burden. The comprehensive molecular mechanisms involved in DCM are still unclear, and advances in high-throughput technologies and bioinformatics distribution can provide a more comprehensive and in-depth understanding of the disease process.

In this study, a systematic collection of DCM-associated genes, followed by WGCNA using multiple clinical features, resulted in the identification of six modules, of which the turquoise coexpression module was the most significantly associated with the occurrence of DCM. We selected the turquoise module as the key module for DCM occurrence and used it for the subsequent hub gene search. In addition, these coexpression modules may interact with each other during DCM. The expression profiles of the hub genes were extracted to construct the LASSO model, and ROC curve analysis showed that the LASSO model had high AUC values for both the training and test sets and could be used as a biomarker for DCM.

TFs regulate gene expression by binding to the promoter regions of target genes; therefore, regulating the biology and binding properties of TFs can be used for targeted therapy [21]. SP2 acts as a cofactor to recruit and increase interactions between the SP2 - Pbx1:Prep1 - Nf-y complex, which in turn promotes genomic binding [22]. The complex activates downstream targeted proteins to control cell proliferation and apoptosis. CTCF is a transcriptional regulatory protein that encodes 11 highly conserved zinc finger structural domains that allow different combinations of structural domains to bind different DNA target sequences and proteins [23]. For example, these domains can bind to a complex containing histone acetyltransferase (HAT) and act as a transcriptional activator or to a complex containing histone deacetylase (HDAC) and act as a transcriptional repressor; if a domain binds to a transcriptional insulator element, it can block communication between the enhancer

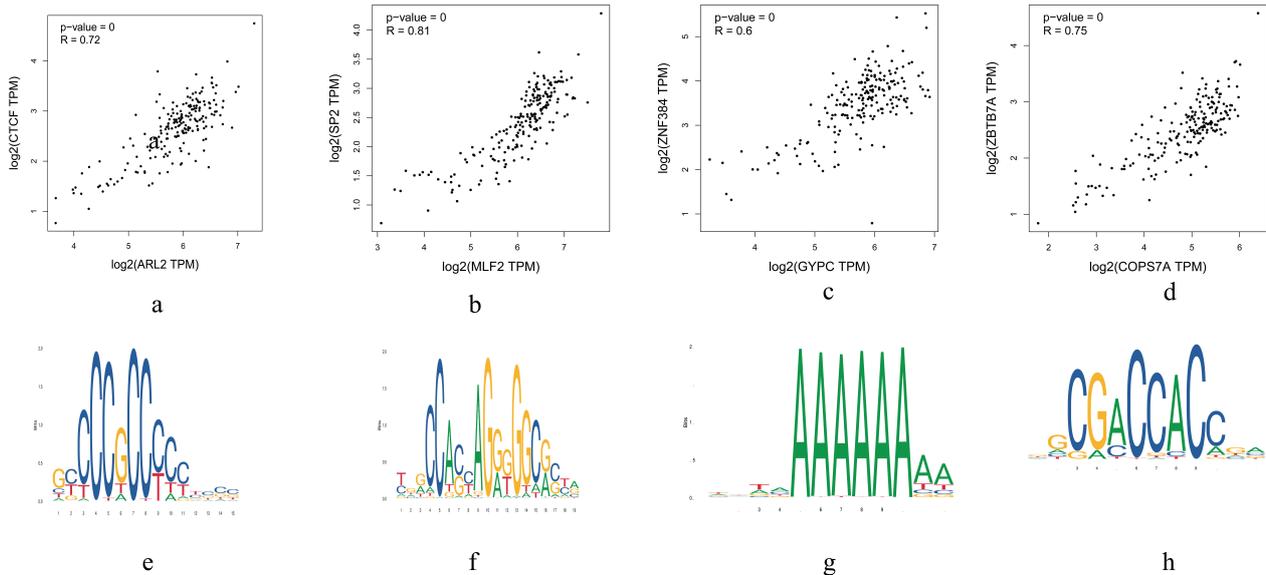


Fig. 7. Transcription factor validation and binding threshold. (a–d) Scatter plots of gene expression correlations corresponding to the most likely transcription factors bound by the hub genes, respectively. (e–h) Binding sites for transcription factors.

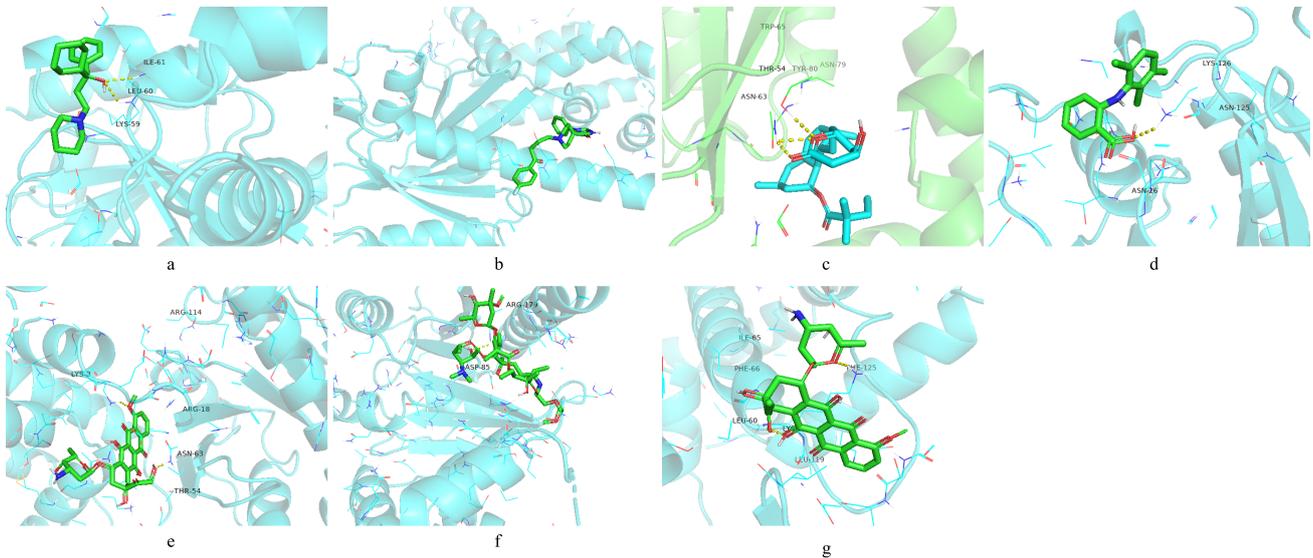


Fig. 8. The results of the molecular docking simulations. (a) ARL2 with Trihexyphenidyl. (b) ARL2 with Meclofenamic acid. (c) ARL2 with Daunorubicin. (d) ARL2 with Simvastatin. (e) ARL2 with Doxorubicin. (f) ARL2 with Dirithromycin. (g) ARL2 with Sipiperone.

and the upstream promoter, thereby regulating blot expression [24,25]. Subsequent *in vitro* and *in vivo* experiments are needed for verification of the modes of action of the TFs screened in this study.

Our results indicated that ARL2 and MLF2 were the central genes obtained after modular analysis and expression validation in DCM. ARL2 is a small G-protein that belongs to the Arf-like small G-protein subfamily, which promotes mitosis by acting on cytoskeletal tissue [26]. Previous studies [27] have shown that ARL2 overexpression increases polymerizable soluble heterodimers, while ARL2 depletion increases microtubule dynamic instability. Thus,

ARL2 depletion significantly reduced the percentage of cells in G2/M phase and mitotic cells. In addition, ARL2 is involved in regulating mitochondrial functions [28], including the maintenance of mitochondrial morphology, motility and ATP levels. ARL2 consumption reduces mitochondrial membrane potential and regulates downstream protein factors to promote mitochondrial fusion and activity [29]. MLF2 and MLF1 are members of the myeloid leukemia factor (MLF) family and are important paraphyletic homologs, sharing nearly 40% identity [30]. The MLF family regulates apoptosis and transcription processes by blocking the association between HS1-associated protein X-1 and HtrA

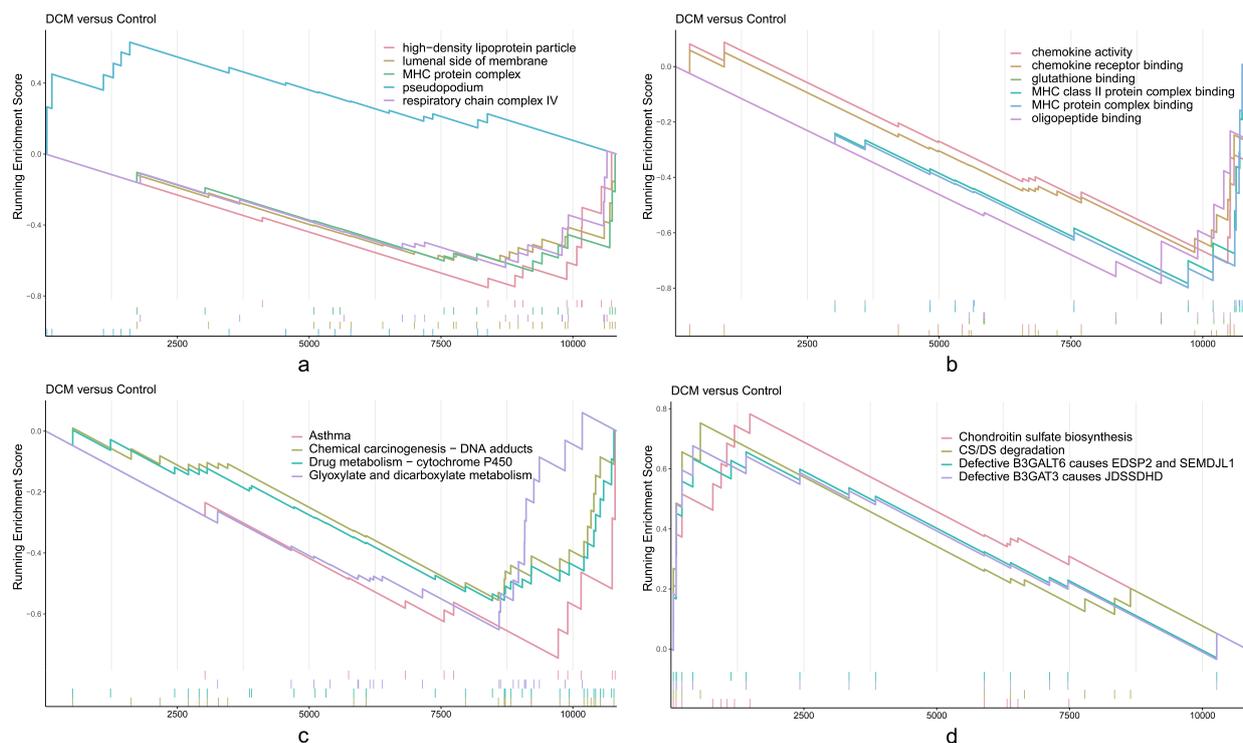


Fig. 9. GSEA. (a) Cell component of gene set. (b) Molecule function of gene set. (c) KEGG pathway of gene set. (d) Reactome of gene set. Abbreviations: GSEA, gene set enrichment analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Table 2. GO, KEGG pathway and Reactome enrichment analysis of DEGs in the DCM and Control samples.

Term	Description	set size	enrichment score	NES	<i>p</i> value	q values
GO:0031143	pseudopodium	16	0.630406438	1.765318418	0.005649718	0.094103654
GO:0098576	luminal side of membrane	23	-0.604675899	-1.906359382	0.002123142	0.059312504
GO:0045277	respiratory chain complex IV	16	-0.639024812	-1.813987412	0.006369427	0.097489377
GO:0042611	MHC protein complex	15	-0.660610096	-1.860053648	0.004158004	0.078901538
GO:0034364	high-density lipoprotein particle	10	-0.753329992	-1.876851266	0.004048583	0.078901538
GO:0042379	chemokine receptor binding	22	-0.670673966	-2.092976246	0.00209205	0.144831662
GO:0008009	chemokine activity	14	-0.709970379	-1.957303241	0.001976285	0.144831662
GO:0043295	glutathione binding	11	-0.782922238	-1.997545034	0.002057613	0.144831662
GO:1900750	oligopeptide binding	11	-0.782922238	-1.997545034	0.002057613	0.144831662
GO:0023026	MHC class II protein complex binding	12	-0.783391223	-2.040328703	0.002053388	0.144831662
GO:0023023	MHC protein complex binding	14	-0.798539963	-2.201478969	0.001976285	0.144831662
hsa05204	Chemical carcinogenesis - DNA adducts	27	-0.555294552	-1.802179492	0.002053388	0.097124772
hsa00982	Drug metabolism - cytochrome P450	32	-0.556436183	-1.909942264	0.003960396	0.137320359
hsa00630	Glyoxylate and dicarboxylate metabolism	21	-0.652250567	-2.059525422	0.001934236	0.097124772
hsa05310	Asthma	10	-0.745038804	-1.854621025	0.006048387	0.153475695
R-HSA-2022870	Chondroitin sulfate biosynthesis	14	0.782742829	2.134302347	0.001941748	0.216045732
R-HSA-2024101	CS/DS degradation	10	0.752675863	1.845047523	0.004048583	0.222934147
R-HSA-3560801	Defective B3GAT3 causes JDSSDHD	14	0.676721314	1.845213822	0.001941748	0.216045732
R-HSA-4420332	Defective B3GALT6 causes EDSP2 and SEMDJL1	15	0.65615414	1.832055703	0.003868472	0.222934147

Abbreviations: GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Gene and Genomes; DCM, dilated cardiomyopathy; DEG, differentially expressed gene.

serine peptidase 2 to inhibit the maturation of HtrA2, thus maintaining normal mitochondrial function [31]. These regulatory patterns and action sites are similar to those of

the previous DCM gene families. Considering the findings our study, ARL2 and MLF2 may serve as therapeutic targets.

We used enrichment analysis of modules and GSEA to explore the key mechanisms of DCM. Excitingly, in contrast to previous results [32], although different datasets were used, the key to the analysis was all about functional changes in the cellular matrix with abnormal mitochondrial function. We also found that the regulated factor MLF2 may act in both leukemia and DCM, similar to previous studies [33] in which the MLLT family was involved in the pathology of DCM, suggesting that modulation of this core target may benefit both difficult diseases. Unlike previous studies [33], we did not only perform integration of multiple datasets and validation at the dataset level, we also performed expression validation through molecular biology experiments, which made our experimental results more reliable. Moreover, we further explored subsequent therapeutic strategies from a pharmacological perspective.

We must acknowledge that there are still limitations to this research. First, the verification of hub genes and their functions has only been tested in human venous blood, but not in other animals model or even human clinical trials. Second, inhibition and overexpression experiments of hub genes have not been completed and need to be further supplemented.

5. Conclusions

Overall, our study demonstrated that ARL2, MLF2, GYPC and COPS7A could be potential gene biomarkers for DCM. However, the identification of the potential key pathways and genes was based on bioinformatic tools and will require further validation by molecular experiments. The extent to which the upregulation and downregulation of ARL2 and MLF2 contribute to DCM development and the specific modes of action of TFs in DCM patients remain to be tested.

Availability of Data and Materials

The deidentified participant data will be shared on a request basis. Please directly contact the corresponding author to request data sharing.

Author Contributions

QG—Conceptualization, Methodology, Software, Investigation, Formal Analysis, Writing - Original Draft; QQ—Data Curation, Visualization, Writing - Original Draft; LW—Visualization, Data Curation, Investigation; SL—Resources, Supervision; XZ—Software, Validation; AD—Visualization, Writing - Review & Editing; QZ—Validation, Investigation; IC—Data Curation, Writing - Review & Editing; RG—Software, Methodology; XL—Conceptualization, Resources, Supervision, Writing - Review & Editing.

Ethics Approval and Consent to Participate

The research protocol was approved by the Ethics Committee of Nanjing Medical University (license number: 2017-SR-086).

Acknowledgment

The authors thanked the patients and investigators who participated in GEO for providing the data and the authors appreciated Jiajin Chen, Department of Biostatistics, School of Public Health, Nanjing Medical University for the statistical guidance, appreciated Mengli Chen and Mengsha Shi for experimental instruction.

Funding

This research received no external funding.

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.fbl2708246>.

References

- [1] Owens AT, Day SM. Reappraising Genes for Dilated Cardiomyopathy: Stepping Back to Move Forward. *Circulation*. 2021; 144: 20–22.
- [2] Rosenbaum AN, Agre KE, Pereira NL. Genetics of dilated cardiomyopathy: practical implications for heart failure management. *Nature Reviews Cardiology*. 2020; 17: 286–297.
- [3] Kim SY, Zhang X, Schiattarella GG, Altamirano F, Ramos TAR, French KM, *et al*. Epigenetic Reader BRD4 (Bromodomain-Containing Protein 4) Governs Nucleus-Encoded Mitochondrial Transcriptome to Regulate Cardiac Function. *Circulation*. 2020; 142: 2356–2370.
- [4] McKenna WJ, Judge DP. Epidemiology of the inherited cardiomyopathies. *Nature Reviews Cardiology*. 2021; 18: 22–36.
- [5] McNally EM, Mestroni L. Dilated Cardiomyopathy. *Circulation Research*. 2017; 121: 731–748.
- [6] Ambrosy AP, Mentz RJ, Fuizat M, Cleland JGF, Greene SJ, O'Connor CM, *et al*. The role of angiotensin receptor–neprilysin inhibitors in cardiovascular disease—existing evidence, knowledge gaps, and future directions. *European Journal of Heart Failure*. 2018; 20: 963–972.
- [7] Verma S, McGuire DK, Kosiborod MN. Two Tales: one Story. *Circulation*. 2020; 142: 2201–2204.
- [8] McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, *et al*. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *European Heart Journal*. 2021; 42: 3599–3726.
- [9] Verdonschot JAJ, Merlo M, Dominguez F, Wang P, Henkens MTHM, Adriaens ME, *et al*. Phenotypic clustering of dilated cardiomyopathy patients highlights important pathophysiological differences. *European Heart Journal*. 2021; 42: 162–174.
- [10] Schultheiss H, Fairweather D, Caforio ALP, Escher F, Hershberger RE, Lipshultz SE, *et al*. Dilated cardiomyopathy. *Nature Reviews Disease Primers*. 2019; 5: 32.
- [11] Fatkin D, Huttner IG, Kovacic JC, Seidman JG, Seidman CE. Precision Medicine in the Management of Dilated Cardiomy-

- opathy: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*. 2019; 74: 2921–2938.
- [12] Tayal U, Ware JS, Lakdawala NK, Heymans S, Prasad SK. Understanding the genetics of adult-onset dilated cardiomyopathy: what a clinician needs to know. *European Heart Journal*. 2021; 42: 2384–2396.
- [13] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015; 43: e47–e47.
- [14] Fernández-Delgado M, Sirsat MS, Cernadas E, Alawadi S, Barro S, Febrero-Bande M. An extensive experimental survey of regression methods. *Neural Networks*. 2019; 111: 11–34.
- [15] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. 2005; 102: 15545–15550.
- [16] Huang S, Wei YK, Kaliamurthi S, Cao Y, Nangraj AS, Sui X, *et al.* Circulating miR-1246 Targeting UBE2C, TNNT3, TRAIIP, UCHL1 Genes and Key Pathways as a Potential Biomarker for Lung Adenocarcinoma: Integrated Biological Network Analysis. *Journal of Personalized Medicine*. 2020; 10: 162.
- [17] Nangraj AS, Selvaraj G, Kaliamurthi S, Kaushik AC, Cho WC, Wei DQ. Integrated PPI- and WGCNA-Retrieval of Hub Gene Signatures Shared Between Barrett's Esophagus and Esophageal Adenocarcinoma. *Frontiers in Pharmacology*. 2020; 11: 881.
- [18] Ferrandi EE, Sayer C, De Rose SA, Guazzelli E, Marchesi C, Saneei V, *et al.* New Thermophilic α/β Class Epoxide Hydrolases Found in Metagenomes From Hot Environments. *Frontiers in Bioengineering and Biotechnology*. 2018; 6: 144.
- [19] Selvaraj G, Kaliamurthi S, Kaushik AC, Khan A, Wei Y, Cho WC, *et al.* Identification of target gene and prognostic evaluation for lung adenocarcinoma using gene expression meta-analysis, network analysis and neural network algorithms. *Journal of Biomedical Informatics*. 2018; 86: 120–134.
- [20] Yeh J, Sung P, Chiang JY, Sheu J, Huang C, Chu Y, *et al.* Early treatment with combination of SS31 and entresto effectively preserved the heart function in doxorubicin-induced dilated cardiomyopathic rat. *Biomedicine & Pharmacotherapy*. 2021; 141: 111886.
- [21] Becskei A. Tuning up Transcription Factors for Therapy. *Molecules*. 2020; 25: 1902.
- [22] Völkel S, Stielow B, Finkernagel F, Berger D, Stiewe T, Nist A, *et al.* Transcription factor Sp2 potentiates binding of the TALE homeoproteins Pbx1:Prep1 and the histone-fold domain protein Nf- γ to composite genomic sites. *Journal of Biological Chemistry*. 2018; 293: 19250–19262.
- [23] Ren G, Zhao K. CTCF and cellular heterogeneity. *Cell & Bioscience*. 2019; 9: 83.
- [24] Braccioli L, de Wit E. CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays in Biochemistry*. 2019; 63: 157–165.
- [25] Wang DC, Wang W, Zhang L, Wang X. A tour of 3D genome with a focus on CTCF. *Seminars in Cell & Developmental Biology*. 2019; 90: 4–11.
- [26] Stephen LA, Elmaghloob Y, Ismail S. Maintaining protein composition in cilia. *Biological Chemistry*. 2017; 399: 1–11.
- [27] Zhou C, Wang Y, Shi L, Wang Z, Ma Y, Li C, *et al.* GTPases Arf5 and ARL2 function partially distinctly during oocyte meiosis. *Journal of Cellular Biochemistry*. 2021; 122: 198–208.
- [28] Newman LE, Zhou CJ, Mudigonda S, Mattheyses AL, Paradies E, Marobio CM, *et al.* The ARL2 GTPase is required for mitochondrial morphology, motility, and maintenance of ATP levels. *PLoS ONE*. 2014; 9: e99270.
- [29] Schiavon CR, Turn RE, Newman LE, Kahn RA. ELMOD2 regulates mitochondrial fusion in a mitofusin-dependent manner, downstream of ARL2. *Molecular Biology of the Cell*. 2019; 30: 1198–1213.
- [30] Huang D, Liu C, Sun X, Sun X, Qu Y, Tang Y, *et al.* CRL4DCAF8 and USP11 oppositely regulate the stability of myeloid leukemia factors (MLFs) *Biochemical and Biophysical Research Communications*. 2020; 529: 127–132.
- [31] Sun Y, Chao J, Xu W, Pourpak A, Boyd K, Moshiah S, *et al.* MLF1 is a proapoptotic antagonist of HOP complex-mediated survival. *Biochimica Et Biophysica Acta (BBA) - Molecular Cell Research*. 2017; 1864: 719–727.
- [32] Wang X, Guan H, Liu W, Li H, Ding J, Feng Y, *et al.* Identification of Immune Markers in Dilated Cardiomyopathies with Heart Failure by Integrated Weighted Gene Coexpression Network Analysis. *Genes*. 2022; 13: 393.
- [33] Zheng Y, Liu Z, Yang X, Weng S, Xu H, Guo C, Xing Z, Liu L, Wang L, Dang Q, Qiu C. Exploring Key Genes to Construct a Diagnosis Model of Dilated Cardiomyopathy. *Frontiers in Cardiovascular Medicine*. 2022; 9: 865096.