Identification of hub genes and transcription factor regulatory network for heart failure using RNA-seq data and robust rank aggregation analysis

Dingyuan Tu¹, Chaoqun Ma¹, ZhenYu Zeng¹, Qiang Xu², Zhifu Guo*, Xiaowei Song* and Xianxian Zhao*

¹Department of Cardiology, Changhui Hospital, Naval Medical University, Shanghai, China, ²Department of Cardiology, Navy 905 Hospital, Naval Medical University, Shanghai, China

Background: Heart failure (HF) is the end stage of various cardiovascular diseases with a high mortality rate. Novel diagnostic and therapeutic biomarkers for HF are urgently required. Our research aims to identify HF-related hub genes and regulatory networks using bioinformatics and validation assays.

Methods: Using four RNA-seq datasets in the Gene Expression Omnibus (GEO) database, we screened differentially expressed genes (DEGs) of HF using Removal of Unwanted Variation from RNA-seq data (RUVSeq) and the robust rank aggregation (RRA) method. Then, hub genes were recognized using the STRING database and Cytoscape software with cytoHubba plug-in. Furthermore, reliable hub genes were validated by the GEO microarray datasets and quantitative reverse transcription polymerase chain reaction (qRT-PCR) using heart tissues from patients with HF and non-failing donors (NFDs). In addition, R packages “clusterProfiler” and “GSVA” were utilized for enrichment analysis. Moreover, the transcription factor (TF)–DEG regulatory network was constructed by Cytoscape and verified in a microarray dataset.

Results: A total of 201 robust DEGs were identified in patients with HF and NFDs. STRING and Cytoscape analysis recognized six hub genes, among which ASPN, COL1A1, and FMOD were confirmed as reliable hub genes through microarray datasets and qRT-PCR validation. Functional analysis showed that the DEGs and hub genes were enriched in T-cell-mediated immune response and myocardial glucose metabolism, which were closely associated with myocardial fibrosis. In addition, the TF–DEG regulatory network was constructed, and 13 significant TF–DEG pairs were finally identified.
Introduction

Heart failure (HF) is a complex clinical syndrome that results from dysfunction of ventricular filling or ejection, characterized by a variety of worsening symptoms and signs, including dyspnea, fatigue, and fluid retention (1). The occurrence of HF is predominantly caused by underlying myocardial diseases, while cardiac lesions from valves, vasculature, pericardium, heart rate/rhythm, or a combination of cardiac abnormalities may also result in cardiac malfunction (2). Despite the development of drug therapy and surgical interventional therapy, the morbidity and mortality of HF are increasing annually worldwide, which seriously threatens human health and quality of life (3, 4). Therefore, to improve the curative efficacy, it remains urgent to investigate the in-depth underlying molecular mechanisms of HF to facilitate its accurate diagnosis, early intervention, and precision therapy.

In recent years, the rapid progress of transcriptome sequencing technology provides new directions for the exploration of epigenetic changes and molecular mechanisms in different diseases, including neoplastic and non-neoplastic diseases (5, 6). Accordingly, an increasing volume of RNA sequencing (RNA-seq) and microarray datasets of HF has been uploaded in the Gene Expression Omnibus (GEO) database, providing opportunities for bioinformatics data mining of marker genes associated with HF (7). However, in comparison to cancer-related surgery, the number of heart transplantation surgeries is relatively small, which results in the small sample size and large batch effects of RNA sequencing or microarray datasets of HF. Therefore, to date, the bioinformatics data mining of HF still faces great challenges, especially regarding the integration and analysis of the RNA-seq data (RUVSeq) related to HF.

The robust rank aggregation (RRA) method, first proposed in 2012 by Kolde et al., is a rigorous approach using probabilistic models to analyze the significant probability of all elements in different sequencing or microarray datasets (8). Recently, the RRA algorithm has been extensively used to integrate data in different microarray platforms to screen the differentially expressed genes (DEGs) in multiple diseases, including thyroid carcinoma (9), prostate cancer (PCa) (10), and DCM (11). For example, Song et al. utilized the RRA method to integrate 10 eligible PCa microarray datasets from the GEO and identify four candidate biomarkers for prognosis of PCa (10). Ma et al. integrated four eligible dilated cardiomyopathy (DCM) microarray datasets from the GEO database and developed a 7-gene signature predictive of DCM by utilizing the RRA method (11). However, due to the greater difficulty in integrating sequencing data, the application of the previous RRA algorithm was limited to microarray data, and the RRA analysis of RUVSeq was still rarely reported. Removal of Unwanted Variation from RUVSeq, a Bioconductor package that generalizes a linear model to regress variance estimated from the expression of housekeeping genes, has been reported to be used to reduce batch effects due to different sequencing conditions (12), which provides a huge possibility for the combination of RUVSeq and the RRA method in integrating different RUVSeq sets and identifying HF-associated DEGs.

In the present study, RUVSeq and RRA analysis were performed for the first time based on four RNA-seq datasets in the GEO database to identify robust DEGs in HF samples and non-failing donor (NFD) samples, followed by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis for the DEGs. Moreover, three reliable HF-related hub genes with differential expression and excellent diagnostic efficiency, ASPN, COL1A1, and FMOD, were selected and validated using microarray datasets and human heart tissue assays. Gene set enrichment analysis (GSEA) and gene set variation analysis (GSVA) were further utilized to investigate potential functions of the hub genes. In addition, the transcription factor (TF)–DEG regulatory network was constructed based on the HF datasets and websites.

Materials and methods

Datasets search and inclusion criteria

The GEO database1 was searched to obtain the sequencing datasets based on the search terms of “heart failure” or/and “HF.” The search results and relevant datasets were filtered according to the following inclusion criteria: (i) the organism was filtered by “homo sapiens”; (ii) the study type was

---

1. www.ncbi.nlm.nih.gov/geo/
set as "expression profiling by high throughput sequencing"; (iii) RUVSeq for both HF samples and NFDs should be included in the dataset; (iv) the total number of samples should not be < 5; and (v) the raw data of the RNA-seq should be provided for reanalysis. Datasets that did not meet the aforementioned criteria were excluded. The selected HF sequencing datasets from the NCBI Sequence Read Archive (SRA) were downloaded as SRA files and converted to FASTQ files via the SRA toolkit.

Compilation of gene expression matrices

To obtain high-quality reads, raw data from the GEO dataset were pre-processed using the fastp tool, and sample quality was assessed by FastQC and MultiQC. The sequences were then aligned against the human reference genome hg38 using STAR. Furthermore, the expression values (count matrices) for either gene bodies or called peaks were generated by featureCounts.

Identification of robust differentially expressed genes by the RNA-seq data and robust rank aggregation method

For RNA-seq expression analysis, batch effects were adjusted using the R package RUVSeq, which applies a generalised linear model to regress out the variation estimated from the expression of the housekeeping gene. First, the initial DEGs were detected using the edgeR program package within a single RNA-seq dataset. Second, the RUVg function in RUVSeq was utilized to remove additional sources of unwanted variation (parameter \(k = 1\)) (17). The remaining non-DEGs were considered as negative control genes and used as housekeeping genes to correct for relative gene expression levels between different samples. Third, based on the corrected gene expression matrix, the corrected DEGs were further obtained by the edgeR package. Fourth, the RRA method-based R package "RobustRankAggreg" was used to integrate the results of RUVSeq analysis of each RNA-seq dataset to identify the final DEGs in patients with HF compared with NFDs. The threshold of DEGs was set as \(|\text{logFC}| > 1\), and the significance criterion was set as an adjusted \(p\)-value < 0.05.

Functional enrichment analysis

To further investigate the possible functions of DEGs identified by the RUVSeq and the RRA method, GO enrichment and KEGG pathway analyses were performed in the upregulated and downregulated DEGs separately, using the R package "clusterProfiler" (18). The GO term or KEGG pathway with adjusted \(p < 0.05\) was considered with significant enrichment. The results were visualized by dot plots using the "dotplot" function of the R package.

Identification of hub genes

The robust DEG list was uploaded to the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database, and the significant protein interaction was determined at the criterion of confidence (combined score) > 0.4. Next, we used Cytoscape software and cytoHubba plug-in to investigate node composition and pick out hub nodes with a high degree of connectivity in the network.

Validation of the hub genes using microarray datasets

RNA-seq datasets for HF samples are limited due to a small volume of heart transplant surgeries and the difficulty in obtaining human heart samples. Therefore, in our study, the four eligible HF sequencing datasets (GSE46224, GSE116250, GSE133054, and GSE135055), including 95 HF and NFD samples, were all used for the identification of DEGs, hub genes, and functional enrichment analysis. To further validate the analysis results, HF microarray datasets were acquired from the GEO database. The inclusion criterion was identical to the RUVSeq sources, except that the study type was set as "expression profiling by array." For the study, four microarray datasets were finally included for the validation: GSE16499, GSE26887, GSE57338, and GSE79962. The gene expression profiling was annotated using the annotation document of corresponding platforms, and the gene expression matrices were column-normalized by the R package "limma" and log-transformed, if necessary. Next, the differential expression of the identified hub genes between patients with HF and NFDs in the microarray datasets was validated and visualized by column graphs.

Validation of the hub genes using quantitative reverse transcription polymerase chain reaction

For further validation, total RNAs of the heart tissues from patients with HF and NFDs were extracted for the
qRT-PCR validation assay. Heart tissues from six patients with HF and eight NFDs were obtained from the Specimen Bank of Cardiovascular Surgery Laboratory and Department of Pathology of Changhai Hospital, Shanghai, China. Written informed consents were obtained from all patients or their family members, and the study was approved by the Institute Ethics Committee of Changhai Hospital.

Total RNAs from the heart tissues were isolated using TRizol reagent (TRizol™ Reagent, Invitrogen). RNAs were then reverse-transcribed into cDNAs using a TOYOBO
ReverTra Ace® qRT-PCR RT Kit (TOYOBO, Japan). SYBR®GREEN (TOYOBO, Japan) was used for qRT-PCR, and the primer sequences used are listed as follows: asporin (ASPN) forward, 5’-GGGTGACGGGTCTCATATC-3’ and reverse, 5’-TGGCACCCTGGACAGAAAG-3’; collagen type 1 alpha 1 chain (COL1A1) forward, 5’-TGGGAATGATGCTGCTA-3’ and reverse, 5’-ACCAGGTTCCACGCTGTTAC-3’; collagen type IX alpha 2 chain (COL9A2) forward, 5’-AAGAGCAACTGCGAGGTGC-3’ and reverse, 5’-GACCCTCGATCTCCATCCTT-3’; collagen type X alpha 1 chain (COL10A1) forward, 5’-TGGGACCCCTCTTGTTAGTG-3’ and reverse, 5’-GCCACACCTGGTCTCA TTTTCT-3’; cartilage oligomeric matrix protein (COMP) forward, 5’-CAGGACGACTTTGATGCAGA-3’ and reverse, 5’-AAGCTGAGCTGCTGTCTGA-3’; and fibromodulin (FMOD) forward, 5’-AGAGAGCTCCCATCTGACCA-3’ and reverse, 5’-GACGGCTGGTTGTAAGGAGG-3’. The expression levels of mRNAs relative to glyceraldehyde-3-phosphate dehydrogenase (GAPDH) were detected using the 2^(-ΔΔCT) method.

Gene set enrichment analysis and gene set variation analysis of the validated hub genes

To further explore potential functions of the hub genes in HE, we performed GSEA and GSV A in the microarray dataset with the maximum HF sample size (GSE57338). The flow of GSEA is as follows: First, correlation analyses were conducted between hub genes and other genes in the gene expression matrix of 54 patients with HF, and genes with the absolute value of correlation coefficient > 0.5 and p-value < 0.05 were defined as hub genes-related genes. Then, KEGG pathway enrichment analysis was conducted on these hub genes-related genes using the ClusterProfiler package. For GSA, 54 patients with HF in the GSE57338 dataset were divided into two groups based on the median expression level of each hub gene (high- and low-expression groups). Then, the “GSV” package was used to explore the pathways associated with the hub genes. The annotated gene set “c2.cp.kegg.v7.4.entrez.gmt” in the Molecular Signatures Database (MsigDB)5 was selected as the reference.

Construction of the transcription factor–differentially expressed gene regulatory network

It has been reported that binding of TFs to the regulatory regions of genes is a key transcriptional regulatory mechanism to control chromatin and transcription, forming a complex system that guides expression of the genome (26). The TF–DEG regulatory network is constructed by using the following methods: First, the NetworkAnalyst database (27)6 and the TF–gene interactions module from the JASPAR database (28) were utilized to explore the possible TFs that could bind to the RRA-identified DEGs. Second, a novel significant TF–DEG regulatory pair was defined in our study according to the following criteria: (i) both the TF and DEG were present in the TF regulatory network constructed by the JASPAR database, and there was predicted interaction between them; (ii) the TF was differentially expressed in patients with HF and NFD samples in the validation set GSE57338 (p < 0.05); and (iii) there was a statistically significant relationship between the expression level of TF and its target gene in the validation dataset GSE57338 (the absolute value of correlation coefficient > 0.5 and p < 0.05). Third, the constructed TF–DEG regulatory network was visualized using Cytoscape.

Statistical analysis

Independent two-sample t-tests were used to analyze variables with homogeneous variance and normal distribution, whereas Mann–Whitney non-parametric tests were used to analyze variables without homogeneous variance and normal distribution. P-values were adjusted for multiple testing by using the Benjamini–Hochberg method. The DEG threshold was set as |logFC| > 1, and the significance criterion was set as an adjusted p-value < 0.05. The hypergeometric test was used to calculate the statistical significance of enrichment analysis. An absolute value of the correlation coefficient |r| > 0.3 (p < 0.05) indicates a significant interaction relationship (29). All data analyses in the present study were performed by using R (version 3.5.3) and Rstudio (version 1.2.1335). Graphic representations were generated by using GraphPad Prism 9.0 (GraphPad, San Diego, CA, USA) and Cytoscape (Version 3.7.1).

Results

Characteristics of the screened heart failure RNA-seq datasets

Figure 1 depicts the flow diagram of our study. After screening and exclusion according to the aforementioned criteria, six datasets from the GEO database were finally included in this analysis: GSE46224 (30), GSE48166, GSE116250 (31), GSE120852 (32), GSE133054 (33), and GSE135055 (34). The characteristics of these six datasets are summarized in

5 http://www.gsea-msigdb.org/gsea/downloads.jsp
6 https://www.networkanalyst.ca/
Supplementary Table 1, including the GSE accession number, study country, number of patients with HF and NFDs, and sequencing platform.

Pre-processing of RNA-seq data

After the quality-filtering using the fastp tool, the reads with a base quality < 20 or the sequence length ≤ 36 nt were discarded. Then, FastQC was used to assess the sequence quality of the dataset. The final all-in-one quality control report of each dataset was generated using MultiQC. The per base sequence quality and per sequence GC content across all samples of 6 RNA-seq datasets are demonstrated in Figure 2.

Determination of the selected datasets

Reads were mapped to the human genome (UCSC, hg38) using STAR, and the unique alignments were filtered and presented in Supplementary Table 2. Samples from each dataset were characterized by principal component analysis (PCA) after normalization and adjustment for batch effects using the RUVSeq package. 2D plots of PCA distribution showed that complete separation between samples of patients with HF and NFD samples was observed in five datasets (GSE46224, GSE116250, GSE120852, GSE133054, and GSE135055), except GSE48166 (Figure 3). Hence, dataset GSE48166 was excluded from subsequent analysis. Next, the expression difference and diagnostic efficacy of the four cardiac function markers, namely, natriuretic peptide A (NPPA), natriuretic peptide B (NPPB), myosin heavy chain 6 (MYH6), and myosin heavy chain 7 (MYH7), were examined between samples of patients with HF and NFD samples in the five sequencing databases. As shown in Figure 4, the markers showed no differential expression and poor diagnostic performance between the two groups in dataset GSE120852, which was also eliminated from further analysis.

Identification of robust differentially expressed genes by RNA-seq data and robust rank aggregation method

Using the RUVSeq package, DEGs (patients with HF vs. NFDs) were screened for adjusted \( p < 0.05 \) and \( |\log FC| > 1 \) in the four identified datasets, respectively, which were visualized by volcano plots (Figures 5A–D). Furthermore, an integrated analysis was performed using the R package “RobustRankAggreg” to generate the differentially expressed ranked gene list. A total of 201 highly ranked DEGs were identified in patients with HF vs. NFD samples, and Supplementary Table 3 exhibits the top 50 upregulated and the top 50 downregulated DEGs. The top 20 upregulated and the 20 most downregulated genes consistently expressed across all datasets were visualized by heatmap, as shown in Figure 5E.

Functional enrichment analysis of differentially expressed genes

To explore the potential biological functions of these DEGs, GO term enrichment and KEGG pathway analyses were performed. The upregulated genes were significantly enriched in extracellular structure organization, skeletal system development, extracellular matrix organization, T-cell activation, and connective tissue development in the biological process (BP) term; the extracellular matrix, collagen-containing extracellular matrix, endoplasmic reticulum lumen, basement membrane, and collagen trimer in the cellular component (CC) term; and extracellular matrix structural constituent, glycosaminoglycan binding, heparin binding, growth factor activity, and extracellular matrix structural constituent conferring tensile strength in the molecular function (MF) term (Figure 6A). For the downregulated genes, the enriched GO functions included purine ribonucleotide metabolic process, coenzyme metabolic process, energy derivation by oxidation of organic compounds, cellular respiration, and citrate metabolic process in the BP category; organelle inner membrane, mitochondrial inner membrane, mitochondrial matrix, mitochondrial protein complex, and mitochondrial membrane part in the CC category; and cofactor binding, coenzyme binding, and NAD binding in the MF category (Figure 6B).

Regarding KEGG pathway analysis, the MAPK signaling pathway, TGF-β signaling pathway, T-cell receptor signaling pathway, Th17 cell differentiation, and ECM–receptor interaction were mostly associated with the upregulated genes (Figure 6C), while the downregulated genes were most enriched in the calcium signaling pathway, carbon metabolism, valine, leucine and isoleucine degradation, citrate cycle, and propanoate metabolism (Figure 6D).

Hub gene determination

The PPI network of the 201 DEGs in patients with HF was constructed by using the STRING database (Figure 7A). Next, the PPI network was loaded into Cytoscape to screen hub genes by degree using the cytoHubba plug-in. As shown in Figure 7B, genes in the inner concentric circles have higher degrees, while genes in the outer concentric circles have relatively lower degrees. Therefore, hub genes were the six genes with the highest degree of connectivity (degree ≥ 10) in the innermost concentric circle: COL1A1, COMP, ASPN, COL10A1, FMOD, and COL9A2.

Furthermore, the relative expression of the identified hub genes in patients with HF and NFD samples was assessed in the
**FIGURE 2**
Quality assessment and GC count evaluation of the data from six RNA sequencing datasets. (A–F) Per base sequence quality across all samples of GSE46224 (A), GSE48166 (B), GSE116250 (C), GSE120852 (D), GSE133054 (E), and GSE135055 (F). (G–L) Per sequence GC content across all samples of GSE46224 (G), GSE48166 (H), GSE116250 (I), GSE120852 (J), GSE133054 (K), and GSE135055 (L).
FIGURE 3
PCA plots of the six RNA-seq datasets in GEO database. PCA distribution plots showed that complete separation between patients with HF and NFD samples was observed in five datasets after RUVseq correction, except GSE48166 (C,D), namely, GSE46224 (A,B), GSE116250 (E,F), GSE120852 (G,H), GSE135055 (K,L). PCA, principal component analysis; GEO, Gene Expression Omnibus; HF, heart failure; NFD, non-failing donor.

Hub gene validation

After normalization, four microarray datasets (GSE16499, GSE26887, GSE57338, and GSE79962) containing human left ventricular samples of HF and NFDs were used to validate the expression of these hub genes (Supplementary Table 4 and Supplementary Figure 1). As shown in Figures 8A–D, the expression of ASPN or FMOD in HF samples was significantly higher than that in the NFD samples in all four datasets, and COL1A1 or COMP showed the similar upregulation in three datasets. However, the expression of COL9A2 or COL10A1 was not statistically different in the HF and NFD samples in these datasets. Consistently, the diagnostic values of the hub genes suggested by the ROC curves revealed the same trend (Figures 8E–H).

In addition to the microarray datasets, the expression of hub genes was further validated by qRT-PCR experiments using 14 heart tissues from patients with HF or NFDs. As described in Figure 9, ASPN, COL1A1, and FMOD were significantly upregulated in the six heart tissues of patients with HF compared with NFDs. Taken together, these validation results confirmed the differential expression and diagnostic value of ASPN, COL1A1, and FMOD as reliable hub genes in HF development.

Gene set enrichment analysis and gene set variation analysis reveal a close relationship between hub genes and glucose metabolism-related pathways

To reveal the underlying mechanism of the three reliable hub genes (ASPN, COL1A1, and FMOD) involved in HF, GSEA was conducted to explore significantly enriched pathways associated with the hub genes in the validation dataset GSE57338. As shown in Figures 10A–C, the top three signaling pathways enriched by the DEGs between subgroups were identified, among which citrate cycle (TCA cycle) and propanoate metabolism pathways were significantly enriched in the subgroups of all the three hub genes. In addition, the enrichment in glucose metabolism-related pathways was further confirmed by GSVA (Figures 10D–F).
Identification of signatures of transcription factor–differentially expressed gene regulatory network

To determine the potential roles of TFs in regulating the transcriptional expression of DEGs, the specific TF–gene regulatory network was established based on the top 20 upregulated and the 20 most downregulated integrated DEGs (Figure 11A). As demonstrated in Figures 11B,C, several TFs, including CEBPB, MEF2A, PPARG, BRCA1, TEAD1, TFAP2A, TP63, SREBF1, and PDX1, showed significant correlation with multiple DEGs and were differentially expressed in patients...
FIGURE 5
Identification of DEGs by RUVSeq and the RRA method. (A–D) Using RUVSeq package, DEGs (patients with HF vs. NFDs) were screened in the four selected RNA-seq datasets GSE46224 (A), GSE116250 (B), GSE133054 (C), and GSE135055 (D), as visualized by volcano plots. Adjusted p < 0.05 and |logFC| > 1. (E) Heatmap of the top 20 upregulated and the 20 most downregulated DEGs screened from the four selected RNA-seq datasets using RRA analysis. DEGs, differentially expressed genes; RRA, robust rank aggregation; HF, heart failure; NFD, non-failing donor; FC, fold change.
with HF and NFDs in GSE57338 ($p < 0.05$). According to the defining criteria of the significant TF–DEG regulation pair, we identified TP63–SERTM1/SYTL5/UNC80, PPARG-XG, BRCA1-NRG1, MEF2A-LSAMP, SREBF1-NPPA/HOOK1/CENPA, TEAD1-CA14/MYH6/PENK, and PDX1-SEC14L5 as significant TF–DEG pairs (Figure 11D).

**Discussion**

In the present study, four HF RNA-seq GEO datasets (GSE46224, GSE116250, GSE133054, and GSE135055) were finally included, involving a total of 100 patients with HF and 38 NFDs. In total, 201 robust HF-related DEGs were obtained utilizing RUVSeq and RRA method, and ASPN, COL1A1, COL9A2, COL10A1, COMP, and FMOD were identified as hub genes with the highest degree of connectivity using STRING database and cytoHubba plug-in. Among them, ASPN, COL1A1, and FMOD exhibited differential expressions and excellent diagnostic efficiency in all four RNA-seq datasets, which were further validated using data from the four screened HF microarray datasets (GSE16499, GSE26887, GSE57338, and GSE79962). Moreover, the significant upregulation of ASPN, COL1A1, and FMOD was experimentally confirmed by qRT-PCR using the heart tissues of patients with HF and NFD samples. In addition, functional enrichment analysis showed that myocardial fibrosis-related pathways resulted from T-cell-mediated immune response and myocardial glucose metabolism were closely associated with the onset and progression of HF. In addition to this, the TF–DEG regulatory network was established, and 13 significant TF–DEG pairs were identified.

Despite the great advancement in HF medical treatment, it remains the major and growing public health problem that leads to considerable morbidity and mortality (35). Robust biomarkers for early diagnosis of HF are the key for novel targeted pharmacological approaches and for improving the prognosis of patients (36). Consistently, serum type B natriuretic peptide (BNP) has been recognized as a well-established biomarker for the diagnosis of HF. However, a recent study reported that a subset (4.9%) of hospitalized patients with confirmed HF had unexpectedly low BNP levels (<50 pg/ml), and a small proportion (0.1–1.1%) had BNP levels even below
FIGURE 7
Identification of HF-related hub genes from the four selected RNA-seq datasets. (A) PPI network of the 201 DEGs in patients with HF was constructed by using the STRING database. The nodes represent proteins, and the edges represent the interactions. (B) Robust hub genes were screened by degree using the cytoHubba plug-in in Cytoscape. The inner the concentric circles, the larger the degree values of the genes. (C–F) Relative expression of the six identified hub genes in GSE46224 (C), GSE116250 (D), GSE133054 (E), and GSE135055 (F), respectively. ns, not significant vs. the NFD group; *p < 0.05 vs. the NFD group; **p < 0.01 vs. the NFD group; ***p < 0.001 vs. the NFD group; ****p < 0.0001 vs. the NFD group. (G–J) Diagnostic values of the six identified hub genes in GSE46224 (G), GSE116250 (H), GSE133054 (I), and GSE135055 (J), respectively, as determined by ROC curves. HF, heart failure; RNA-seq, RNA sequencing; PPI, protein–protein interaction; DEGs, differentially expressed genes; STRING, search tool for the retrieval of interacting genes; NFD, non-failing donor; ROC, receiver operating characteristic.

detection limits (37). Therefore, it remains urgent to explore novel molecules with potentially new mechanisms for the development of HF.

Recently, gene mining using microarrays or RNA-seq datasets has been widely used to generate the transcriptomic profiles of HF development. Zhang et al. identified six hub genes (BCL3, HCK, PPIF, S100A9, SERPINA1, and TBC1D9B) as potential biomarkers of HF by using the weighted gene co-expression network analysis (WGCNA) method through three HF datasets, namely, GSE59867, GSE1869, and GSE42955 (38). Tian et al. constructed a random forest algorithm and artificial neural network and detected six hub genes by mining of two HF datasets (GSE57345 and GSE141910) (39). However, the aforementioned studies were based on the integration of DEGs, rather than raw data from different datasets. To date, the inconsistency between different platforms and datasets remains the major hurdle blocking the bioinformatics mining of HF-related genes, especially for RNA-seq datasets.

The RRA method, a recently emerging analysis method, has been widely used to integrate different datasets and produce a ranked list of the DEGs (40). For example, Ma et al. utilized the RRA method to integrate four eligible DCM microarray datasets from the GEO and developed a 7-gene signature predictive model of DCM (11). While in the present study, using RUVSeq to substantially decrease batch effects, we integrated, for the first time, the different RNA-seq datasets of the GEO database to
Validation of the six hub genes using four normalized HF-related microarray datasets from the GEO database. (A–D) Relative expression of the six hub genes in GSE16499 (A), GSE26887 (B), GSE57338 (C), and GSE79962 (D), respectively. ns, not significant vs. the NFD group; *p < 0.05 vs. the NFD group; **p < 0.01 vs. the NFD group; ***p < 0.001 vs. the NFD group; ****p < 0.0001 vs. the NFD group. (E–H) Diagnostic values of the six identified hub genes in GSE16499 (E), GSE26887 (F), GSE57338 (G), and GSE79962 (H), respectively, as determined by ROC curves. HF, heart failure; GEO, Gene Expression Omnibus; NFD, non-failing donor; ROC, receiver operating characteristic.
explore DEGs and hub genes associated with HF by using the RRA method. Through internal RNA-seq dataset and external microarray dataset validation, ASPN, COL1A1, and FMOD were finally identified as real hub genes of HF, which were further confirmed by qRT-PCR using the heart tissues from patients with HF and NFDs.

Interestingly, the identified hub genes ASPN (41), COL1A1 (42), and FMOD (43), all belong to the type I collagen members in the extracellular matrix (ECM) composition and have been reported to play important roles in the development and progression of various diseases, especially malignant tumors. For example, ASPN was reported to enhance tumor invasion and cancer-associated fibroblasts via activation of the CD44-Rac1 pathway in gastric cancer (41). Ma et al. highlighted the role of COL1A1 as a potential diagnostic biomarker and therapeutic target in early development and metastasis of hepatocellular carcinoma (42). Ao et al. revealed that FMOD could promote tumor angiogenesis by upregulating the expression of angiogenic factors in human small-cell lung cancer (43). Regarding the function of hub genes in HF development, a multi-level transcriptomic study conducted by Hua et al. suggested that COL1A1 might be a plasma biomarker of HF and associated with HF progression, especially to predict the 1-year survival from HF onset to transplantation. A COL1A1 content ≥ 256.5ng/ml in plasma was found to be associated with poor survival within 1 year of heart transplantation from HF (34). In the study conducted by Andenæs et al., FMOD was found 3–10-fold upregulated in hearts of patients with HF and mice, and FMOD-KO mice showed a relatively mild hypertrophic phenotype (44). However, to the best of our knowledge, there are no experimental studies focusing on the role of ASPN in HF. Therefore, our multi-dataset RRA analysis, followed by microarray dataset and experimental validation, provides more robust and comprehensive evidence for the value of the three ECM-related genes, namely, COL1A1, FMOD, and ASPN, in HF development.

Recent advances have highlighted the crucial role of immune activation in the development and progression of HF. A study...
by Aghajanian et al. demonstrated that adoptive transfer of T cells that express a chimeric antigen receptor against fibroblast activation protein can inhibit myocardial fibrosis and improve cardiac function in mice (45). Consistently, according to the GO term analysis in our study, the upregulated HF-related DEGs were enriched in T-cell activation of the “BP” term, the extracellular matrix of “CC” terms, and the extracellular matrix structural constituent of “MF” terms. Moreover, regarding the KEGG pathway analysis, the T-cell receptor signaling pathway and ECM–receptor interaction were identified as the significantly enriched pathways of the upregulated DEGs. Considering that all the three hub genes—ASPN, COL1A1, and FMOD—are closely associated with the ECM, we thus speculate a potentially key pathway in the development of HF, that is, T-cell-mediated immune responses lead to the imbalance in ECM anabolism and catabolism, ultimately resulting in myocardial fibrosis and HF.

To further explore the potential mechanism of ASPN, COL1A1, and FMOD in HF, we performed GSEA and GSVA on the validation dataset of GSE57338. Results showed that
ASPAN-, COL1A1-, or FMOD-related DEGs were enriched in the "citric acid cycle (TCA cycle)" and "propionic acid metabolism" pathways, both of which are closely associated with glucose metabolism (46, 47). Notably, targeting cardiac glucose metabolism has been recognized as a promising therapeutic strategy for HF treatment. Liu et al. reported that dichloroacetate, a pyruvate dehydrogenase kinase inhibitor, could alter glucose metabolism in cardiomyocytes by...
stimulating the activity of pyruvate dehydrogenase complex, thereby improving cardiac efficiency (48). In addition, inhibitors of fatty acid oxidation such as trimetazidine (49), perhexiline (50), and etomoxir (51) can improve cardiac function in patients with HF by increasing glucose oxidation.

Aberrant regulation of TFs is strongly associated with the onset and progression of HF (52). Therefore, in our research, we further investigated the TF–gene interactions to detect the transcriptional regulators of the robust DEGs. Among the seven identified significant TFs, MEF2A (53) and PPARG (54) have been reported to play a role in cardiac remodeling and water retention in HF, respectively. Liu et al. found that suppressing expression of TEAD1, the Hippo signaling effector, could activate the necroptotic pathway and induce massive cardiomyocyte necroptosis, ultimately leading to impaired cardiac function (55). Moreover, loss of BRCA1 in mouse cardiomyocytes resulted in adverse cardiac remodeling and poor ventricular function (56). Although the functions of these TFs in HF have been partially reported, the regulatory relationship of the TF–DEG pairs and the in-depth molecular mechanisms remain to be further validated through HF-related experimental studies.

Our study has several limitations. First, the sample size of patients with HF is relatively small, although we have included as many datasets that met the criteria as possible. Future studies with larger sample sizes are needed to confirm these findings. Second, this study is mainly based on bioinformatics analysis and qRT-PCR validation of hub gene expression. Further experimental research is needed to clarify the in-depth mechanism of the hub gene-related HF regulation. Third, information about disease grades, treatment methods, and prognosis of patients with HF is not available in the database, leading to the failure to analyze correlation between hub genes and clinical characteristics or prognosis of HF. Fourth, the etiology of HF is complex, involving multiple environmental factors in addition to genetic factors (57), such as behavioral factors, socioeconomic and psychosocial factors, air quality, and meteorological factors (58–60). Horton et al. reported that the influence of modifiable lifestyle factors cannot be ignored in the development of direct-to-consumer (DTC) genetic tests (61). In recent years, emerging evidence has shown that gene–environment interactions play an important role in complex disease progression. Bentley et al. revealed that the genetic associations with lipids could be modified by smoking (62). Therefore, future research needs to further focus on the role of environmental factors and gene–environment interactions in HF.

Data availability statement
The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material.

Ethics statement
The studies involving human participants were reviewed and approved by the Institute Ethics Committee of Changhai Hospital. The patients/participants provided their written informed consent to participate in this study.

Author contributions
ZG, XS, and XZ contributed to the conception of the study. DT and CM performed the study execution and experiments. ZZ and QX contributed to part of bioinformatic analysis. DT, CM, and XS prepared the manuscript. XZ and ZG contributed to the funding of the study. All authors reviewed the manuscript, provided critical revision, and have approved the final version for publication.

Funding
This work was supported in part by the grants from the Chinese Natural Science Foundation (82170233 to XZ) and 234 Discipline Promotion Foundation of Changhai (2020YXK057 to ZG).

Acknowledgments
We thank Sirui Zhang and Xiaojuan Fan for their technical guidance on the analysis. In addition, we also thank the donors and their families for research organ donations.
Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcvm.2022.916429/full#supplementary-material

References

1. McMurray JJ, Pfeffer MA. Heart failure. Lancet. (2005) 365:1877–89. doi: 10.1016/s0140-6736(05)66621-4
2. Tanai F, Frantz S. Pathophysiology of heart failure. Compr Physiol. (2015) 6:187–214. doi: 10.1002/cphy.c140005
3. Snipelly D, Chaudhry SP, Stewart GC. The many faces of heart failure. Card Electrophysiol Clin. (2019) 11:11–20. doi: 10.1016/j.ccep.2018.11.001
4. Moster A, Hons AW. Clinical epidemiology of heart failure. Heart. (2007) 93:1137–46. doi: 10.1136/hrt.2003.025270
5. Saeedian AH, Yousefsfan L, Vahidnejhad H, Uitto J. Research techniques made simple: whole-transcriptome sequencing by rna-seq for diagnosis of monogenic disorders. J Invest Dermatol. (2020) 140:1117–26.e1. doi: 10.1016/j.jid.2020.02.032
6. Hanauer DA, Rhodes DR, Sinha-Kumar C, Chinnaiyan AM. Bioinformatics approaches in the study of cancer. Curr Mol Med. (2007) 7:133–41. doi: 10.2174/156652407779943431
7. Joshi A, Rienks M, Thieleflatus K, Mayr M. Systems biology in cardiovascular disease: a multiteomics approach. Nat Rev Cardioiol. (2021) 18:313–30. doi: 10.1038/s41569-020-00477-1
8. Kolde R, Laur S, Adler P, Vílio J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. (2012) 28:573–80. doi: 10.1093/bioinformatics/bts570
9. Liu Y, Chen TY, Yang ZY, Fang W, Wu Q, Zhang C. Identification of hub genes in papillary thyroid carcinoma: robust rank aggregation and weighted gene co-expression network analysis. J Transl Med. (2020) 18:170. doi: 10.1186/s12967-020-02327-7
10. Song ZY, Chao F, Zhou Z, Ma Z, Li W, Chen G. Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis. Aging. (2019) 11:4736–56. doi: 10.18632/aging.102087
11. Ma X, Mo C, Huang L, Cao P, Shen L, Gui C. An robust rank aggregation and least absolute shrinkage and selection operator analysis of novel gene signatures in dilated cardiomyopathy. Front Cardiovasc Med. (2021) 8:747803. doi: 10.3389/fcvm.2021.747803
12. Rissou D, Ngai J, Speed TP, Dudoit S. Normalization of rna-seq data using factor analysis of control genes or samples. Nat Biotechnol. (2014) 32:896–902. doi: 10.1038/nbt.2933
13. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics. (2018) 34:4884–90. doi: 10.1093/bioinformatics/bty560
14. Ewels P, Magnusson M, Lundin S, Käller M. Multiqc: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. (2016) 32:3047–8. doi: 10.1093/bioinformatics/btw354
15. Dohin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. Star: ultrafast universal rna-seq aligner. Bioinformatics. (2013) 29:15–21. doi: 10.1093/bioinformatics/btt635
16. Liao Y, Smyth GK, Shi W. Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. (2014) 30:923–30. doi: 10.1093/bioinformatics/btt656
17. Pang B, Xu J, Hu J, Guo F, Wan L, Cheng M, et al. Single-cell rna-seq reveals the invasive trajectory and molecular cascades underlying glioblastoma progression. Mol Oncol. (2019) 13:2588–603. doi: 10.1016/j.molonc.2019.12.016
18. Yu G, Wang LG, Han Y, He QY. Clusterprofiler: an r package for comparing biological themes among gene clusters. OMICS. (2012) 16:284–7. doi: 10.1089/omi.2011.0118
19. Salazarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. (2019) 47:D607–13. doi: 10.1093/nar/gky1131
20. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. Cytobubble: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. (2014) 8(Suppl 4):S1. doi: 10.1186/1752-0509-8-S4-s1
21. Kong SW, Hu YW, Ho JW, Ikeda S, Polster S, John R, et al. Heart failure-associated changes in rna splicing of sarcomere genes. Circ Cardiovasc Genet. (2010) 3:138–46. doi: 10.1161/circgenetics.109.104698
22. Greco S, Fasano R, Castelvecchio S, D’Alessandra Y, Arcelli D, Di Donato M, et al. MicroRNA dysregulation in diabetic ischemic heart failure patients. Diabetes. (2012) 61:1633–41. doi: 10.2337/db11-0952
23. Liu Y, Morley M, Brandimarto J, Hannenhalli S, Hu Y, Ashley EA, et al. Rna-seq identifies novel myocardial gene expression signatures of heart failure. Genomics. (2015) 105:83–9. doi: 10.1016/j.ygeno.2014.12.002
24. Markovich SJ, Al Khiami B, Efimov JR, Evans S, Vader J, Jain A, et al. Widespread down-regulation of cardiac mitochondrial and sarcomeric genes in patients with sepsis. Crit Care Med. (2017) 45:407–14. doi: 10.1097/ccm.0000000000002207
25. 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 Res. (2015) 43:e47. doi: 10.1093/nar/gkv007
26. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. (2018) 172:650–65. doi: 10.1016/j.cell.2018.01.029
27. 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 Res. (2019) 47:W234–41. doi: 10.1093/nar/gkae240
28. Khan A, Fornes O, Stigland A, Gheorge M, Castro-Mondragon JA, van der Lee R, et al. Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. (2018) 46:D226–6. doi: 10.1093/nar/gkx1126
29. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Med J. (2012) 24:69–71.
30. Yang KC, Yamada KA, Patel AV, Topkara VK, George I, Cheema FH, et al. Deep rna sequencing reveals dynamic regulation of myocardial noncoding rnas in failing human heart and remodeling with mechanical circulatory support. Circulation. (2014) 129:1099–21. doi: 10.1161/circulationaha.113.038863
31. Sweet ME, Coccio A, Slavov D, Jones KL, Sweet JR, Graw SL, et al. Transcriptional analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure. BMC Genom. (2018) 19:812. doi: 10.1186/s12864-018-5213-9
Collagen 1a1 (Col1a1) is a reliable biomarker and putative therapeutic target for heart failure progression. *BMC Med.* (2020) 18:2. doi: 10.1186/s12916-019-1469-4

50. Schmidt-Schweda S, Holubarsch C. First clinical trial with etomoxir in patients with chronic congestive heart failure. *Clin Sci.* (2000) 99:27–35.

51. Minerath RA, Hall DD, Grueter CE. Targeting transcriptional machinery to inhibit enhancer-driven gene expression in heart failure. *Heart Fail. Rev.* (2019) 24:725–41. doi: 10.1007/s10741-019-09792-3

52. Minerath RA, Hall DD, Grueter CE. Targeting transcriptional machinery to inhibit enhancer-driven gene expression in heart failure. *Heart Fail. Rev.* (2019) 24:725–41. doi: 10.1007/s10741-019-09792-3

53. Aghajanian H, Kimura T, Rurik JG, Hancock AS, Leibowitz MS, Li L, et al. Metabolism. *Sci Rep.* (2017) 7:4476. doi: 10.1038/s41598-017-04762-x

54. Bao LZ, Sheng M, Qudirat H, Shi JB, Yu T, Song JW, et al. Obestatin ameliorates water retention in chronic heart failure by downregulating renal aquaporin 2 through Gpr39, V2r and Pparg signaling. *Cell Death Differ.* (2021) 28:2045–59. doi: 10.1038/s41418-020-00732-5

55. Liu J, Wu Y, Dong K, He X, Zhou H, Shen J, et al. Tead1 protects against necroptosis in postmitotic cardiomyocytes through regulation of nuclear DNA-encoded mitochondrial genes. *Cell Death Differ.* (2021) 28:2045–59. doi: 10.1038/s41418-020-00732-5

56. Shukla PC, Singh KK, Quan A, Al-Omran M, Teoh H, Lovren F, et al. Brcal is an essential regulator of heart function and survival following myocardial infarction. *Nat Commun.* (2021) 2:593. doi: 10.1038/ncomms1601

57. Wald NJ, Old R. The illusion of polygenic disease risk prediction. *Genet Med.* (2019) 21:1705–7. doi: 10.1038/s41436-018-0418-5

58. Yang Q, Zhang Z, Gregg EW, Flanders WD, Merritt R, Hu FB. Added sugar intake and cardiovascular diseases mortality among us adults. *JAMA Internal Med.* (2014) 174:516–24. doi: 10.1001/jamainternalmed.2013.15363

59. Zeng J, Zhang X, Yang J, Iann J, Xiang H, Dear K, et al. Humidity may modify the relationship between temperature and cardiovascular mortality in Zhejiang province, China. *Int J Environ Res Public Health.* (2017) 14:1383. doi: 10.3390/ijerph14111383

60. Yuval S, Joseph P, Rangarajan S, Islam S, Mente A, Hystad P, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (pure): a prospective cohort study. *Lancet.* (2020) 395:795–808. doi: 10.1016/s0140-6736(19)32008-2

61. Horton R, Crawford G, Freeman L, Fenwick A, Wright CF, Lucassen A. Direct-to-consumer genetic testing. *BMJ.* (2019) 367:l5688. doi: 10.1136/bmj.l5688

62. Bentley AR, Sung YJ, Brown MR, Winkler TW, Kraja AT, NtallA I, et al. Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat Genet.* (2019) 51:636–48. doi: 10.1038/s41438-019-0378-y