# Integrative transcriptomic, proteomic, and machine learning approach to identifying feature genes of atrial fibrillation

**Yaozhong Liu**
Department of Cardiovascular Medicine, Second Xiangya Hospital, Central South University, Hunan Province, China.

**Fan Bai**
Department of Cardiovascular Medicine, Second Xiangya Hospital, Central South University, Hunan Province, China.

**Zhenwei Tang**
Department of Dermatology, Xiangya Hospital, Central South University, Hunan Province, China.

**Na Liu**
Department of Cardiovascular Medicine, Second Xiangya Hospital, Central South University, Hunan Province, China.

**Qiming Liu** ( qimingliu@csu.edu.cn )
Department of Cardiovascular Medicine, Second Xiangya Hospital, Central South University, Hunan Province, China.

**Research Article**

**Keywords:** atrial fibrillation, transcriptomic, proteomic, machine learning, feature gene

**Posted Date:** November 11th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-104305/v1

# Abstract

**Background:** Atrial fibrillation (AF) is the most common arrhythmia with poorly understood mechanisms. We aimed to investigate the biological mechanism of AF and to discover feature genes by analyzing multi-omics data and by applying a machine learning approach.

**Methods:** At the transcriptomic level, four microarray datasets (GSE41177, GSE79768, GSE115574, GSE14975) were downloaded from the Gene Expression Omnibus database, which included 130 available atrial samples from AF and sinus rhythm (SR) group. Microarray meta-analysis was adopted to identified differentially expressed genes (DEGs). At the proteomic level, a qualitative and quantitative analysis of proteomics in the left atrial appendage of 18 patients (9 with AF and 9 with SR) was conducted. The machine learning correlation-based feature selection (CSF) method was introduced to selected feature genes of AF using the training set of 130 samples involved in the microarray meta-analysis. The Naive Bayes (NB) based classifier constructed using training set was evaluated on an independent validation test set GSE2240.

**Results:** 863 DEGs with a FDR<0.05 and 482 differentially expressed proteins (DEPs) with a FDR<0.1 and fold change >1.2 were obtained from the transcriptomic and proteomic study, respectively. The DEGs and DEPs were then analyzed together which identified 30 biomarkers with consistent trends. Further, 10 feature, including 8 upregulated genes (CD44, CHGB, FHL2, GGT5, IGFBP2, NRAP, SEPTIN6, YWHAQ) and 2 downregulated genes (TNNT1, TRDN) were selected from the 30 biomarkers through machine learning CFS method using training set. The NB based classifier constructed using the training set accurately and reliably classify AF from SR samples in the validation test set with a precision of 87.5% and AUC of 0.995.

**Conclusion:** Taken together, our present work might provide novel insights into the molecular mechanism and provide some promising diagnostic and therapeutic targets of AF.


# 1. Background

Atrial fibrillation (AF) is the most common cardiac arrhythmia and is a leading cause of stroke, heart failure, and dementia[1]. AF currently affects over 30 million individuals worldwide[2], and this number is projected to grow dramatically over the next 20 years[3]. Despite > 100 years of basic and clinical research, the fundamental mechanisms of AF remain poorly understood.

Microarray expression analysis of atrial tissues can provide a global unbiased framework to characterize the transcriptional changes associated with AF. Advancement of high-throughput microarray technology is producing a large number of gene expression data, which are powerful tools for discovering and studying novel biomarkers for AF. However, analysis based on high throughput data may face the ‘curse of dimensionality’. This refers to the phenomena that the amount of dependent variables increases greatly while the amount of samples is relatively small, increasing statistical errors[4].

Recently, Integrated transcriptomic and quantitative proteomic analyses have been widely used to promote a better understanding of the molecular mechanisms driving biological processes in cells and tissues[5]. Advances in mass-spectrometry (MS) provide an unprecedented opportunity for antibody-independent proteome profiling with approximately 80% of all proteins in major human tissues quantifiable by this technique[6]. By integrating the transcriptomic and proteomic data, the 'curse of dimensionality' can be solved through cross-validation in the two levels. Besides, combining datasets from different origins by meta-analysis to increase the power of identification and using some machine learning algorithms can also improve the problems caused by this 'curse' [7].

Here, our objective was to elucidate a more complete understanding of molecular mechanisms underlying AF and to find potential diagnostic and therapeutic targets. The integration of multi-omics data, along with the application of the machine learning approach, vouched for the identification of key pathways and feature genes in AF, which may help to investigate the underlying mechanism of AF and to discover potential diagnostic and therapeutic targets.

# 2. Methods

### 2.1 Microarray data collection and preprocessing

For the meta-analysis, AF microarray expression data sets were collected from NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). Only microarray data that met the following criteria were included: (1) Datasets were produced by Genome-wide mRNA expression profiling by microarray; (2) The experimental platform was GPL570 (Affymetrix Human Genome U133 Plus 2.0 microarray); (3) Data sets should be gene expression profiles of human atria tissues between AF and sinus rhythm (SR); (4) The minimum number of cases and controls was three. Then, the raw CEL files were downloaded and preprocessed using robust multi array average (RMA) algorithm with 'affy' package[8] implemented in R software. The quality of individual samples was assessed using the 'arrayQualitymetrics' packages[9]. The outlier samples were excluded if it was detected by array intensity distribution criteria. After that, raw CEL files of the rest samples were preprocessed again using RMA algorithm for background correction, quantile normalization, and summarization.

We then reannotated the probes of GPL570 as it improves accuracy and makes it possible to identify new transcripts. In brief, the probe sequences were downloaded from Affymetrix (affymetrix.com) and were remapped to the human genome (GRCh38 release 99 primary assembly) using the R package 'Rsubread'[10]. Then, the chromosomal positions of these probes were matched to the corresponding genome annotation database in Ensembl using the R package 'GenomicRanges'[11]. Probe sets that were mapped to >1 gene were removed to ensure the reliability of the reannotation. The median expression values among all multiple probe IDs were selected to represent the corresponding gene symbol. After that, 19557 unique genes were retained. The normalized and annotated datasets containing 19557 rows and 130 columns were used for further meta-analysis.

GSE2240, which contained microarray expression profiles from 10 AF and 20 SR atrial samples, were preprocessed using RMA algorithm and annotated using ‘annotate’ and ‘hgu133a.db’ packages. The median expression values among multiple probe IDs were selected to represent the corresponding gene symbol.

### 2.2 Microarray meta-analysis using GeneMeta

‘GeneMeta’ Bioconductor package[12] in R was used to perform a microarray meta-analysis of data sets from different ‘origins’. This package is based on the meta-analysis method proposed by Choi et al.[12] in which an overall ranked gene list is produced based on the false discovery rate (FDR) of each gene. In this study, samples regarded as the same ‘origin’ must come from the same tissue (left atria, right atria, etc) and the same microarray study. The Random effect model (REM) was used[13]. The false discovery rate (FDR) for each gene was obtained with the function “ZscoreFDR” using 1000 permutations. Genes with FDR<0.05 were considered as differentially expressed genes (DEGs).

### 2.3 Proteomics study

18 left atrial appendage (LAA) tissue samples were obtained as surgical specimens from patients with mitral stenosis undergoing cardiac surgery at the Second Xiangya Hospital of Central South University, including 9 with chronic AF and 9 with SR. The characteristics of all patients are presented in Table 2. For each clinical group, three samples were mixed into one pooled sample. Qualitative and quantitative proteomic analysis was performed using dimethyl label-coupled high performance liquid chromatography-tandem mass spectrometry (HPLC-MS/MS) and MaxQuant software[14]. Benjamini-Hochberg's method was used to calculate the FDR. DEPs were identified using a criterion of FDR<0.1 and fold change >1.2. The detailed procedure for proteomic study was described in Supplementary Material 1.

### 2.4 Pathway enrichment analysis

Metascape (https://Metascape.org/) is a web-based portal designed to provide a comprehensive gene list annotation and analysis resource for biologists[15]. It is one of the most effective tools to conducted muti-omics level enrichment analysis. To gain more insights of biological roles of identified DEGs and DEPs, we conducted pathway enrichment analysis of Gene Ontology biological process (GO BP), Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and Canonical pathway in Metascape tools. By inputting the lists of DEGs and DEPs simultaneously, Metascape online tools can identify commonly-enriched and selectively-enriched pathways from two levels, which enables a comprehensive assessment of the molecular features of the biological process.

### 2.5 Cross-validation between the transcriptomic and proteomic study

The DEGs and DEPs were further analyzed using VennDiagram to compare and identify the shared genes. To make the selected biomarkers more significant, we only select genes that have consistent expression trends (upregulated or downregulated) between the transcriptomic and proteomic levels for further analysis.

### 2.6 Feature selection and classification algorithm

The 130 samples involved in the meta-analysis were selected as the training set. The correlation-based selection (CFS) method[16] implemented in WEKA solfware[17] was used using the training set to select feature genes. Three popular state-of-the-art supervised classification methods (NB, Naive Bayes; SMO, sequential minimal optimization; and RF, random forest) were used for generating the classification models using WEKA with the default parameter settings. The three supervised machine learning algorithms were trained with the training set and were further validated by 6 fold cross-validation. The generated training models were compared based on their accuracy. The best trained classifier generated in the training set was further validated on the independent test set GSE2240. The performance of the classifier was evaluated using criteria including precision, recall, F-measure, Matthews correlation coefficient (MCC), AUC (area under receiver operating curve), and auPRC (area under precision-recall curve), true positive rate, false positive rate, and Kappa statistic.

## 3. Results

### 3.1 Microarray data description and preprocessing

In the transcriptomic meta-analysis study, four microarray data sets were included containing a total of 54 SR and 79 AF atrial samples (Table 1). The included raw CEL files were pre-processed and quality control analysis of the data sets (after normalization) led to the removal of 3 samples including GSM1005420, GSM3182694, and GSM3182707. After removing the outliers and reprocessing, the normalized data sets consisting of 130 samples were taken for further meta-analysis approach.

### 3.2 Identification of DEGs

As shown in Table 1, we only considered samples from the same study and the same tissue as the same ‘origin’, which led to a total of 7 different origins. We then performed a meta-analysis by using the R package ‘GeneMeta’ and DEGs were detected by comparing the differential expression levels between the AF and SR group. The results identified 863 genes as DEGs (FDR < 0.05; 485 up-regulated: z-score > 0; 378 down-regulated: z-score < 0) (Supplementary Table SI).

**Table 1.** Characteristics of publically available GEO data sets used in the microarray meta-analysis

| Accession number | Organism | Platform | Number of samples (SR/AF) | Origin |
|---|---|---|---|---|
| GSE41177 | Homo sapiens | Affymetrix Human Genome U133 Plus 2.0 | Left atrial appendage: 3/16 Left atrial junction: 3/16 | 1 2 |
| GSE79768 | Homo sapiens | Affymetrix Human Genome U133 Plus 2.0 | Left atrial specimen: 6/7 Right atrial specimen:6/7 | 3 4 |
| GSE115574 | Homo sapiens | Affymetrix Human Genome U133 Plus 2.0 | Left atrial tissue: 15/14 Right atrial tissue: 16/14 | 5 6 |
| GSE14975 | Homo sapiens | Affymetrix Human Genome U133 Plus 2.0 | Left atrial appendage: 5/5 | 7 |

### 3.3 **Results of proteomic study**

The characteristics of the patients included in the proteomic study were balanced between the two groups, except for the left atrial (LA) size (Table 2). Figure 1A shows the procedure of the proteomic study. Pearson's correlation analysis indicated good repeatability between the samples (Figure 1B). The mass accuracy of the MS data met the requirement (Figure 1C) and the distribution of peptides' length agreed with the properties of tryptic peptides (Figure 1D). In total, we identified 4489 proteins including 3606 quantifiable proteins(Figure 1E). Proteins with FDR<0.1 and fold change >1.2 were considered significant, which led to the identification of 482 DEPs (301 upregulated and 181 downregulated) (Figure 1E, F) (Supplementary Material 2).

**Table 2.** Characteristics of the patients involved in proteomic study

| | SR(n=9) | AF(n=9) | P |
|---|---|---|---|
| Male (n, %) | 5 (55.6%) | 4 (44.4%) | 1 |
| Age (year) | 50.5 ± 6.5 | 55.5 ± 9.0 | 0.195 |
| BMI (kg/m2) | 22.2 ± 2.0 | 22.7 ± 1.8 | 0.489 |
| Hypertension (n, %) | 4 (44.4%) | 6 (66.7%) | 0.637 |
| Hemoglobin (g/L) | 135.7±16.8 | 128.1±22.4 | 0.546 |
| WBC (109/L) | 6.2±2.1 | 6.7±1.6 | 0.546 |
| Platelet(109/L) | 225.8±86.4 | 205.2±44.0 | 0.931 |
| ALT (u/L) | 18.7±11.4 | 19.6±7.7 | 0.666 |
| AST (u/L) | 20.2±4.9 | 24.8±12.4 | 0.605 |
| ALB (g/L) | 37.4±1.9 | 39.1±4.2 | 0.489 |
| Serum creatinine (umol/L) | 64.5±21.0 | 69.1±15.6 | 0.222 |
| NT-proBNP (pg/mL) | 161.4±77.7 | 201.8±138.7 | 0.546 |
| Fasting blood glucose (mmol/L) | 5.0 ± 0.4 | 5.2 ± 0.4 | 0.489 |
| Total cholesterol (mmol/L) | 4.6 ± 0.5 | 4.3 ± 0.5 | 0.269 |
| RA size (mm) | 33.0±4.3 | 33.0±4.3 | 0.796 |
| LA size (mm) * | 37.9 ± 3.1 | 49.4 ± 8.0 | 0.001 |
| RV size (mm) | 30.1±4.9 | 33.8±8.1 | 0.489 |
| LV size (mm) | 46.9±10.6 | 54.1±10.8 | 0.161 |
| EF (%) | 62.9 ± 8.6 | 61.3 ± 8.8 | 0.711 |
| Mitral valve area ($cm^2$) | 1.8 ± 0.3 | 1.9 ± 0.3 | 0.746 |
| NYHA class (I/II) | 9/0 | 6/3 | |

*: $p<0.05$

### 3.4 Pathway enrichment analysis and visualization

Pathway enrichment analysis helps researchers gain mechanistic insight into gene lists generated from genome-scale (omics) experiments. This method identifies biological pathways that are enriched in a gene list more than would be expected by chance. Metascape helps to integrate different omics data such as genomics, transcriptomics, and proteomics, which enables a comprehensive understanding of a

biological process. Unlike other methods, Metascape clusters enriched terms into non-redundant groups that will be critical for informing future studies. We visualized the top 20 clusters and chose the most significant (lowest p-value) term within each of the 20 clusters to represent the cluster. For the upregulated proteins and mRNAs, most of the top 20 clusters (19) were enriched in both protein and mRNA levels, which highly suggested the importance of these pathways in AF pathogenesis (Figure 2A). While for the down-regulated ones, the top 20 clusters were mainly involved in energy metabolism-related pathways, and these pathways were only enriched in the protein level (Figure 2B). To further capture the relationships between the terms, we selected a subset of representative terms from each of the 20 clusters (up to the 10 best scoring terms) and convert them into a network layout which was visualized within Cytospace (Figure 2, right part).

### 3.5 Cross-validation

To make the selected biomarkers more significant, we only select genes that have consistent expression trends (upregulated or downregulated) between the transcriptomic and proteomic levels for further analysis. As VeneDiagram showed (Figure 3), 23 up-regulated genes/proteins, and 7 down-regulated genes/proteins were identified to have consistent trends from two-level. These 30 genes/proteins were considered as important biomarkers for AF.

### 3.6 Performance evaluation of AF classifier

After feature selection using training set, the number of features reduced from 30 to 10 including CD44, CHGB, FHL2, GGT5, IGFBP2, NRAP, SEPTIN6, TNNII, and TRDN. After removing the bath effect using ‘sva’ packages in the R solfware, the expression values of these 10 features were used to develop the classification models using three supervised machine learning algorithms—NB, SMO, and RF by 6-fold cross-validation to classify AF and SR samples on training set. All classifiers performed well with a precision of 86.9% for NB, 86.3% for SMO, and 76.8% for RF (Table 3). Base on a comprehensive evaluation of precision and other popular measures, the NB classifier performed best and the constructed NB classifier using the training set was further evaluated in the independent test set. Among the 30 atrial samples, 24 of them (80%) were correctly classified. The performance criteria including precision, recall, F-measure, MCC, AUC, auPRC, true positive rate, false positive rate, and Kappa statistic were 87.5%, 0.8, 0.805, 0.661, 0.995, 0.995, 0.8, 0.1, and 0.609, respectively. Therefore, the overall measures of high accuracy confirmed the efficacy of the classifier to distinguish AF from SR samples, which further proved that the 10 gene feature are important biomarkers for AF.

**Table 3.** Performance of different prediction models generated by 6-fold cross-validation on the training data set

| Classifier | Precision | Recall | F-Measure | MCC | AUC | auPRC | TP Rate | FP Rate | Kappa statistic |
|---|---|---|---|---|---|---|---|---|---|
| NB | 0.869 | 0.869 | 0.869 | 0.729 | 0.925 | 0.920 | 0.869 | 0.143 | 0.728 |
| SMO | 0.863 | 0.862 | 0.862 | 0.715 | 0.860 | 0.814 | 0.862 | 0.142 | 0.715 |
| RF | 0.768 | 0.769 | 0.768 | 0.518 | 0.887 | 0.881 | 0.769 | 0.259 | 0.516 |

Abbreviations: NB, Naive Bayes; SMO, sequential minimal optimization; RF, random forest; MCC, Matthews correlation coefficient; AUC, area under receiver operating curve; auPRC, area under precision recall curve; TP, true positive; FP, false positive.

# 4. Discussion

To our knowledge, this is the first integrated transcriptomic and proteomic analysis of human AF atrial tissue, and the first to identify feature genes of AF using machine learningPrevious transcriptomic studies have provided insights into the pathogenesis of AF[18, 19]. However, these experiments are generally analyzed through a single data source or restricted to a fewer sample which can lead to biological and technical biases. Thus, the microarray meta-analysis was used in this study to integrate four microarray data sets of AF from GEO which led to the identification of 863 DEGs. To elucidate a more complete understanding of AF pathogenesis, we also conducted a proteomic study of local atrial tissue which identified 482 DEPs.

Pathway enrichment analysis can help to characterize physiological and functional changes associated with the changes in mRNA and protein expression in AF atrial tissues. For the upregulated mRNAs or proteins, the top 19 scoring items were enriched in both transcriptomic and proteomic levels, which vouched for the importance and significance of these pathways. Some of the items, such as ‘PDGFRB PATHWAY’, ‘activation of immune response’, ‘muscle structure development’, ‘regulation of actin cytoskeleton’, and ‘leukocyte degranulation’, have been proved to play key roles in AF progression[3, 20]. For the downregulated mRNAs or proteins, the top 19 scoring items were only enriched in the proteomic level, and these pathways were mainly involved in metabolism regulation, such as ‘mitochondrial respiratory chain complex assembly’, ‘TP53 regulates metabolic genes’, and ‘response to oxidative stress’. Besides, the ‘Metabolism of lipids’ pathway was enriched in two levels. These are in accord with the recent studies which highlighted the role of metabolic remodeling in AF[21-23]. The reason why these pathways are only identified in the protein level may be caused by some post-transcriptional and translational regulations.

After cross-validation between the two omics data. We identified 30 genes or proteins with the same trends between two levels. To make the selected features more significant and informative, the machine learning CFS feature selection method was adopted in the training set which led to the final 10 features, wherein 8 are upregulated (CD44, CHGB, FHL2, GGT5, IGFBP2, NRAP, SEPTIN6, YWHAQ) and 2 are downregulated (TNNI1, TRDN). The NB classifier base on the expression values of these features in the

training set can classify AF and SR samples with a precision of 87.5% and AUC of 0.995 in the independent test set.

Some of these feature genes have been reported to be associated with AF or its related pathogenesis. The CD44 related pathways including CD44/STAT3 and CD44/NOX4 signaling pathways can lead to atrial fibrosis[24] and $Ca^{2+}$-handling abnormalities[25] during AF. Secretogranin-1 (CHGB) presents in the secretory granules in atrial myoendocrine cells and is co-localized with atrial natriuretic peptide (ANP) while CHGB genetic variation results in oxidative stress[26] and hypertension[27]. The four and a half LIM domains protein 2 (FHL2) is a component of the hypertrophic response and is found to be protective in cardiac hypertrophic through inhibiting MAPK/ERK signaling[28]. MAPK has been proved to function in AF context by mediating oxidative stress[29, 30], epicardial adipose tissue remodeling[31], atrial fibrosis[32], load-induced hypertrophic response[33], and ionic channel remodeling[34]. Gamma-glutamyltransferase-5 (GGT5) is confirmed to be closely associated with immune cell activation[35] and oxidative stress[36, 37] and can be a potential biomarker of myocardial infarction[38]. Insulin-like growth factor-binding protein 2 (IGFBP2) belongs to the insulin-like growth factor-binding protein (IGFBP) family. Two recent studies observed a higher hazard of incident AF associated with higher mean levels of plasma IGFBP1 protein[39] and IGFBP3 protein[40]. Nebulin related anchoring protein (NRAP) is present in myofibril precursors during myofibrillogenesis and thought to be involved in myofibril assembly[41], and its genetic variance is associated with cariomyopathy[42]. Septin-6 (SEPTIN6) is invovled in extracellular matrix remodeling[43]. 14-3-3 protein theta (YWHAQ) is a gene in the P53 network and has been shown to promote apoptosis directly upon genotoxic stress[44]. Another proteomic also identified YWHAQ as an important biomarker in AF[44]. TNNI1 encodes a troponin-I protein that is the dominant form of troponin-I expressed in the fetal/neonatal/infant heart, and its participants in AF remains unknown. Triadin (TRDN) is a stable subunit of the ryanodine receptor 2 (RyR2) and is involved in the regulation of $Ca^{2+}$ release[45]. The loss or dysfunction of RyR2 stable subunits was demonstrated to cause the occurrence of spontaneous calcium elevation in AF atrial cells[46]. Our present study further proved and emphasized the importance of these markers.

There are some limitations to the current study. Firstly, the number of samples included in the microarray meta-analysis remains relatively small (n=130), which is caused by the limited number of available studies in the GEO database. Secondly, there is no corresponding clinical information of the samples, we were not able to make a prognostic analysis of these biomarkers. Third, the samples used in the proteomic study came from patients with mitral stenosis. The psychophysiology of AF in mitral stenosis may have some differences from those with non-valvular AF. Finally, the transcriptomic and proteomic can only indicate the potential causes for a phenotypic response, but they cannot predict what will happen at the next level. Thus, one should consider the metabolomic that provides a functional view of an organism as determined by the sum of its genes, RNA, proteins, and environmental factors[47]. Nonetheless, the integrated analysis of multi-omics data along with the machine learning method makes sure the selected genes as important features for AF. Further studies are needed to clarify their functions in AF pathogenesis.

# 5. Conclusions

In conclusion, the current study identified a list of significantly dysregulated feature genes associated with AF using a multi-omics analysis. The machine learning feature selection identified 10 feature genes. Naive Bayes prediction model built in the training set using the expression profiles of 10 features performed accurately and reliably classified AF from SR samples in the independent test set. These findings could provide novel insight into the pathogenesis of AF and suggested that the feature genes might be diagnostic and therapeutic targets for AF.

# Abbreviations

**AF**, atrial fibrillation; **GEO**, Gene Expression Omnibus; **MS:** mass spectrometry; **DEGs,** and differentially expressed proteins; **SR,** sinus rhythm; **RMA,** robust multiarray average; **REM,** random effect model; **FDR,** false discovery rate; **DEGs:** differentially expressed genes; **LAA,** left atrial appendage; **HPLC-MS/MS:** high-performance liquid chromatography-tandem mass spectrometry; **GO BP**, Gene Ontology biological process; **KEGG,** Kyoto Encyclopedia of Genes and Genomes; **CSF,** correlation-based selection; **NB,** Naive Bayes; **SMO,** sequential minimal optimization; **RF,** random forest; **AUC,** area under receiver operating curve; **MCC,** Matthews correlation coefficient; **auPRC,** area under precision-recall curve.

# Declarations

**Ethics approval and consent to participate:** The proteomic study was approved by the Ethics Committee of the Second Xiangya Hospital of Central South University. The research was carried out in accordance with the World Medical Association Declaration of Helsinki. Informed written consent was obtained from all patients.

**Consent for publication:** Not applicable.

**Availability of data and materials:** The microarray datasets analyzed during the present study are available from the Gene Expression Omnibus repository (https://www.ncbi.nlm.nih.gov/geo). The results of proteomic study were submitted as supplementary material.

**Competing interests:** The authors declare that they have no competing interests.

**Funding:** This work was supported by grants from the National Natural Science Foundation of China (no.81770337). They had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Authors' contributions:** YL and FB performed the bioinformatic analysis and was a major contributor in writing the manuscript. ZT and NL made important modifications to the manuscript. YL and QL designed the research project and created the final revision of the manuscript. All authors read and approved the final version of the manuscript.

**Acknowledgments:** Liu Yaozhong would like to thank Miss Wan Ziwei for her love.

# References

[1] K. P, B. S, K. D, A. A, A. D, C. B, C. M, D. HC, H. H, H. J, H. G, M. AS, O. J, P. BA, S. U, V.P. B, V. P, 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS, European heart journal 37(38) (2016) 2893-2962.

[2] S.S. Chugh, R. Havmoeller, K. Narayanan, D. Singh, M. Rienstra, E.J. Benjamin, R.F. Gillum, Y.H. Kim, J.H. McAnulty, Jr., Z.J. Zheng, M.H. Forouzanfar, M. Naghavi, G.A. Mensah, M. Ezzati, C.J. Murray, Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study, Circulation 129(8) (2014) 837-47.

[3] U. Schotten, S. Verheule, P. Kirchhof, A. Goette, Pathophysiological mechanisms of atrial fibrillation: a translational appraisal, Physiological reviews 91(1) (2011) 265-325.

[4] N. Loris, B. Sheryl, L. Alessandra, Combining multiple approaches for gene microarray classification, Bioinformatics (8) (2012) 8.

[5] A. Ghazalpour, B. Bennett, V.A. Petyuk, L. Orozco, R. Hagopian, I.N. Mungrue, C.R. Farber, J. Sinsheimer, H.M. Kang, N. Furlotte, C.C. Park, P.Z. Wen, H. Brewer, K. Weitz, D.G. Camp, 2nd, C. Pan, R. Yordanova, I. Neuhaus, C. Tilford, N. Siemers, P. Gargalovic, E. Eskin, T. Kirchgessner, D.J. Smith, R.D. Smith, A.J. Lusis, Comparative analysis of proteome and transcriptome variation in mouse, PLoS genetics 7(6) (2011) e1001393.

[6] M.-S. Kim, S.M. Pinto, D. Getnet, R.S. Nirujogi, S.S. Manda, R. Chaerkady, A.K. Madugundu, D.S. Kelkar, R. Isserlin, S. Jain, J.K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N.A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L.D.N. Selvan, A.H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S.K. Sreenivasamurthy, A. Marimuthu, G.J. Sathe, S. Chavan, K.K. Datta, Y. Subbannayya, A. Sahu, S.D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K.R. Murthy, N. Syed, R. Goel, A.A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.-C. Huang, J. Zhong, X. Wu, P.G. Shaw, D. Freed, M.S. Zahari, K.K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C.J. Mitchell, S.K. Shankar, P. Satishchandra, J.T. Schroeder, R. Sirdeshmukh, A. Maitra, S.D. Leach, C.G. Drake, M.K. Halushka, T.S.K. Prasad, R.H. Hruban, C.L. Kerr, G.D. Bader, C.A. Iacobuzio-Donahue, H. Gowda, A. Pandey, A draft map of the human proteome, Nature 509(7502) (2014) 575-581.

[7] A. Ramasamy, A. Mondry, C.C. Holmes, D.G. Altman, Key issues in conducting a meta-analysis of gene expression microarray datasets, PLoS medicine 5(9) (2008) e184.

[8] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry, affy-analysis of Affymetrix GeneChip data at the probe level, Bioinformatics 20(3) (2004) p. 307-315.

[9] K. Audrey, G. Robert, H. Wolfgang, arrayQualityMetrics—a bioconductor package for quality assessment of microarray data, Bioinformatics (3) (2008) 3.

[10] Y. Liao, G. Smyth, W. Shi, The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads, Nucleic acids research 47(8) (2019) e47.

[11] M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. Morgan, V. Carey, Software for computing and annotating genomic ranges, PLoS computational biology 9(8) (2013) e1003118.

[12] C. JK, Y. U, K. S, Y. OJ, Combining multiple microarray studies and modeling interstudy variation, Bioinformatics (Oxford, England) (2003) i84-90.

[13] C.J. Kyoon, U. Yu, K. Sangsoo, Y.O. Joon, Combining multiple microarray studies and modeling interstudy variation, Bioinformatics (suppl_1) (2003) suppl_1.

[14] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nature Biotechnology 26(12) (2008) 1367-1372.

[15] Y. Zhou, B. Zhou, L. Pache, M. Chang, A.H. Khodabakhshi, O. Tanaseichuk, C. Benner, S.K. Chanda, Metascape provides a biologist-oriented resource for the analysis of systems-level datasets, Nat Commun 10(1) (2019) 1523.

[16] Y. Lei, H. Liu, Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA, 2003.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl. 11(1) (2009) 10–18.

[18] A.S. Barth, S. Merk, E. Arnoldi, L. Zwermann, P. Kloos, M. Gebauer, K. Steinmeyer, M. Bleich, S. Kääb, M. Hinterseer, Reprogramming of the human atrial transcriptome in permanent atrial fibrillation: Expression of a ventricular-like genomic signature, Circulation Research 96(9) (2005) 1022-9.

[19] D. A, B. J, S. H, N. D, C. L, P. G, J. D, R. E, G. AM, M. K, M. C, S. JD, V.W. DR, C. MK, Left atrial transcriptional changes associated with atrial fibrillation susceptibility and persistence, Circulation. Arrhythmia and electrophysiology 8(1) (2015) 32-41.

[20] L. Y, S. Q, M. Y, L. Q, The role of immune cells in atrial fibrillation, Journal of molecular and cellular cardiology 123 (2018) 198-208.

[21] D. Opacic, K.A. van Bragt, H.M. Nasrallah, U. Schotten, S. Verheule, Atrial metabolism and tissue perfusion as determinants of electrical and structural remodelling in atrial fibrillation, Cardiovascular

Research 109(4) (2016) 527-541.

[22] L. Y, B. F, L. N, O. F, L. Q, The Warburg effect: A new insight into atrial fibrillation, Clinica chimica acta; international journal of clinical chemistry 499 (2019) 4-12.

[23] F. Bai, T. Tu, F. Qin, Y. Ma, N. Liu, Y. Liu, X. Liao, S. Zhou, Q. Liu, Quantitative proteomics of changes in succinylated proteins expression profiling in left appendages tissue from valvular heart disease patients with atrial fibrillation, Clinica Chimica Acta 495 (2019) 345-354.

[24] S.H. Chang, Y.H. Yeh, J.L. Lee, Y.J. Hsu, C.T. Kuo, W.J. Chen, Transforming growth factor-beta-mediated CD44/STAT3 signaling contributes to the development of atrial fibrosis and fibrillation, Basic research in cardiology 112(5) (2017) 58.

[25] W.J. Chen, S.H. Chang, Y.H. Chan, J.L. Lee, Y.J. Lai, G.J. Chang, F.C. Tsai, Y.H. Yeh, Tachycardia-induced CD44/NOX4 signaling is involved in the development of atrial remodeling, J Mol Cell Cardiol 135 (2019) 67-78.

[26] F. Rao, K. Zhang, S. Khandrika, M. Mahata, M.M. Fung, M.G. Ziegler, B.K. Rana, D.T. O'Connor, Isoprostane, an "intermediate phenotype" for oxidative stress heritability, risk trait associations, and the influence of chromogranin B polymorphism, Journal of the American College of Cardiology 56(16) (2010) 1338-50.

[27] K. Zhang, F. Rao, L. Wang, B.K. Rana, S. Ghosh, M. Mahata, R.M. Salem, J.L. Rodriguez-Flores, M.M. Fung, J. Waalen, B. Tayo, L. Taupenot, S.K. Mahata, D.T. O'Connor, Common functional genetic variants in catecholamine storage vesicle protein promoter motifs interact to trigger systemic hypertension, Journal of the American College of Cardiology 55(14) (2010) 1463-75.

[28] Y. Liang, W.H. Bradford, J. Zhang, F. Sheikh, Four and a half LIM domain protein signaling and cardiomyopathy, Biophys Rev 10(4) (2018) 1073-1085.

[29] L. Rochette, J. Lorin, M. Zeller, J.C. Guilland, L. Lorgis, Y. Cottin, C. Vergely, Nitric oxide synthase inhibition and oxidative stress in cardiovascular diseases: possible therapeutic targets?, Pharmacology & therapeutics 140(3) (2013) 239-57.

[30] X. Liang, Q. Zhang, X. Wang, M. Yuan, Y. Zhang, Z. Xu, G. Li, T. Liu, Reactive oxygen species mediated oxidative stress links diabetes and atrial fibrillation, Molecular medicine reports 17(4) (2018) 4933-4940.

[31] N. Suffee, T. Moore-Morris, P. Farahmand, C. Rucker-Martin, G. Dilanian, M. Fradet, D. Sawaki, G. Derumeaux, P. LePrince, K. Clement, I. Dugail, M. Puceat, S.N. Hatem, Atrial natriuretic peptide regulates adipose tissue accumulation in adult atria, Proceedings of the National Academy of Sciences of the United States of America 114(5) (2017) E771-e780.

[32] J. Fan, L. Zou, K. Cui, K. Woo, H. Du, S. Chen, Z. Ling, Q. Zhang, B. Zhang, X. Lan, L. Su, B. Zrenner, Y. Yin, Atrial overexpression of angiotensin-converting enzyme 2 improves the canine rapid atrial pacing-

induced structural and electrical remodeling. Fan, ACE2 improves atrial substrate remodeling, Basic research in cardiology 110(4) (2015) 45.

[33] R. Kerkela, M. Ilves, S. Pikkarainen, H. Tokola, V.P. Ronkainen, T. Majalahti, J. Leppaluoto, O. Vuolteenaho, H. Ruskoaho, Key roles of endothelin-1 and p38 MAPK in the regulation of atrial stretch response, American journal of physiology. Regulatory, integrative and comparative physiology 300(1) (2011) R140-9.

[34] W. Cheng, Y. Zhu, H. Wang, The MAPK pathway is involved in the regulation of rapid pacing-induced ionic channel remodeling in rat atrial myocytes, Molecular medicine reports 13(3) (2016) 2677-82.

[35] E. Lu, F.D. Wolfreys, J.R. Muppidi, Y. Xu, J.G. Cyster, S-Geranylgeranyl-L-glutathione is a ligand for human B cell-confinement receptor P2RY8, Nature 567(7747) (2019) 244-248.

[36] W. Li, Z.Q. Wu, S. Zhang, R. Cao, J. Zhao, Z.J. Sun, W. Zou, Augmented expression of gamma-glutamyl transferase 5 (GGT5) impairs testicular steroidogenesis by deregulating local oxidative stress, Cell and tissue research 366(2) (2016) 467-481.

[37] R. Dhingra, P. Gona, T.J. Wang, C.S. Fox, R.B. D'Agostino, Sr., R.S. Vasan, Serum gamma-glutamyl transferase and risk of heart failure in the community, Arteriosclerosis, thrombosis, and vascular biology 30(9) (2010) 1855-60.

[38] A. Sharma, M. Ghatge, L. Mundkur, R. Vangala, Translational informatics approach for identifying the functional molecular communicators linking coronary artery disease, infection and inflammation, Molecular medicine reports (2016).

[39] L. Staerk, S.R. Preis, H. Lin, S.A. Lubitz, P.T. Ellinor, D. Levy, E.J. Benjamin, L. Trinquart, Protein Biomarkers and Risk of Atrial Fibrillation: The FHS, Circ Arrhythm Electrophysiol 13(2) (2020) e007607.

[40] M. Busch, A. Kruger, S. Gross, T. Ittermann, N. Friedrich, M. Nauck, M. Dorr, S.B. Felix, Relation of IGF-1 and IGFBP-3 with prevalent and incident atrial fibrillation in a population-based study, Heart rhythm 16(9) (2019) 1314-1319.

[41] M.L. Bang, J. Chen, Roles of Nebulin Family Members in the Heart, Circulation journal : official journal of the Japanese Circulation Society 79(10) (2015) 2081-7.

[42] C. Vasilescu, T.H. Ojala, V. Brilhante, S. Ojanen, H.M. Hinterding, E. Palin, T.P. Alastalo, J. Koskenvuo, A. Hiippala, E. Jokinen, T. Jahnukainen, J. Lohi, J. Pihkala, T.A. Tyni, C.J. Carroll, A. Suomalainen, Genetic Basis of Severe Childhood-Onset Cardiomyopathies, Journal of the American College of Cardiology 72(19) (2018) 2324-2338.

[43] K.B. Collins, H. Kang, J. Matsche, J.E. Klomp, J. Rehman, A.B. Malik, A.V. Karginov, Septin2 mediates podosome maturation and endothelial cell invasion associated with angiogenesis, The Journal of cell biology 219(2) (2020).

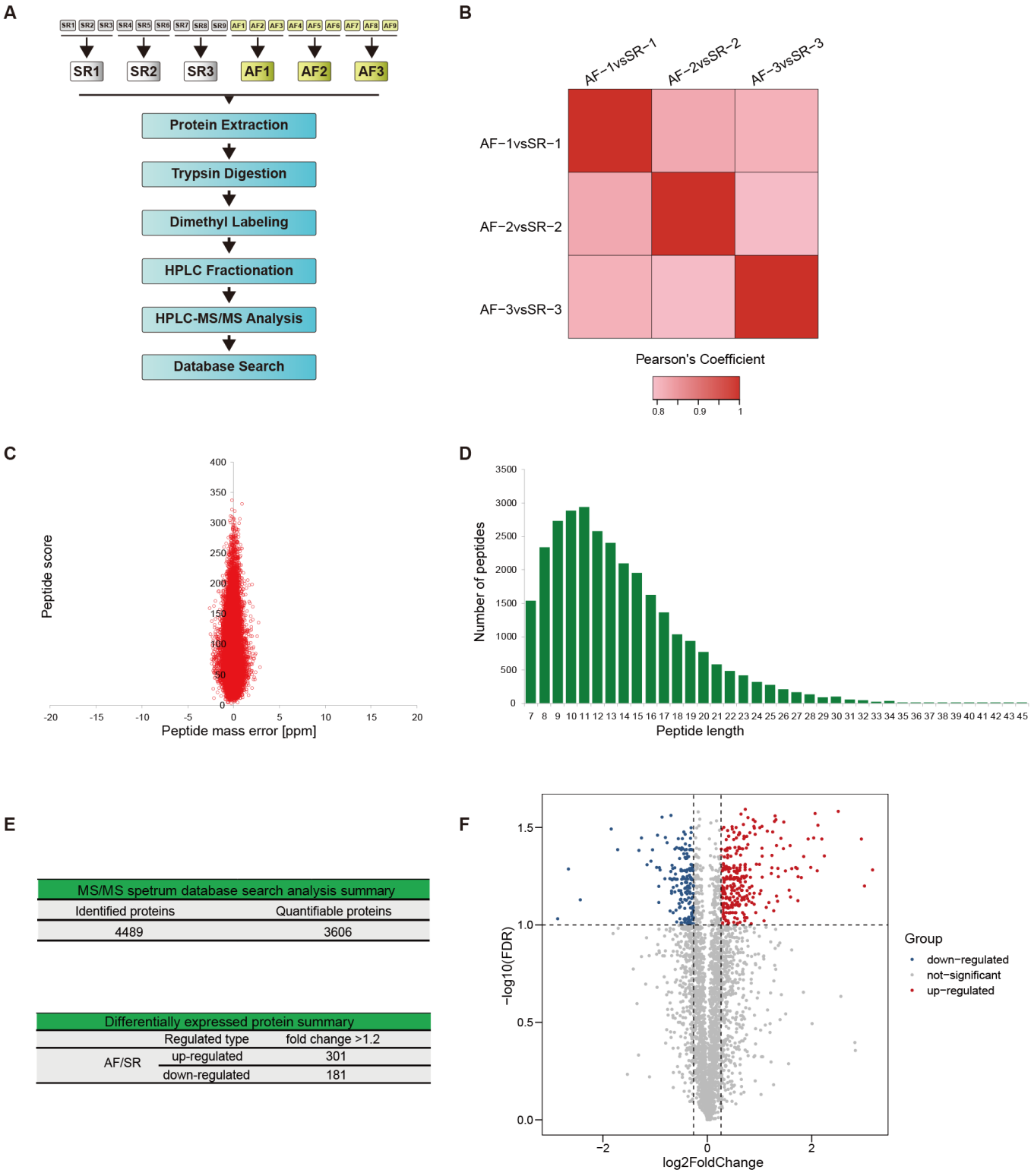
[44] A. Vazquez, L.F. Grochola, E.E. Bond, A.J. Levine, H. Taubert, T.H. Müller, P. Würl, G.L. Bond, Chemosensitivity profiles identify polymorphisms in the p53 network genes 14-3-3tau and CD44 that affect sarcoma incidence and survival, Cancer research 70(1) (2010) 172-80.

[45] C. Franzini-Armstrong, F. Protasi, P. Tijskens, The assembly of calcium release units in cardiac muscle, Annals of the New York Academy of Sciences 1047 (2005) 76-85.

[46] J.C. Zhang, H.L. Wu, Q. Chen, X.T. Xie, T. Zou, C. Zhu, Y. Dong, G.J. Xiang, L. Ye, Y. Li, P.L. Zhu, Calcium-Mediated Oscillation in Membrane Potentials and Atrial-Triggered Activity in Atrial Cells of Casq2(R33Q/R33Q) Mutation Mice, Frontiers in physiology 9 (2018) 1447.

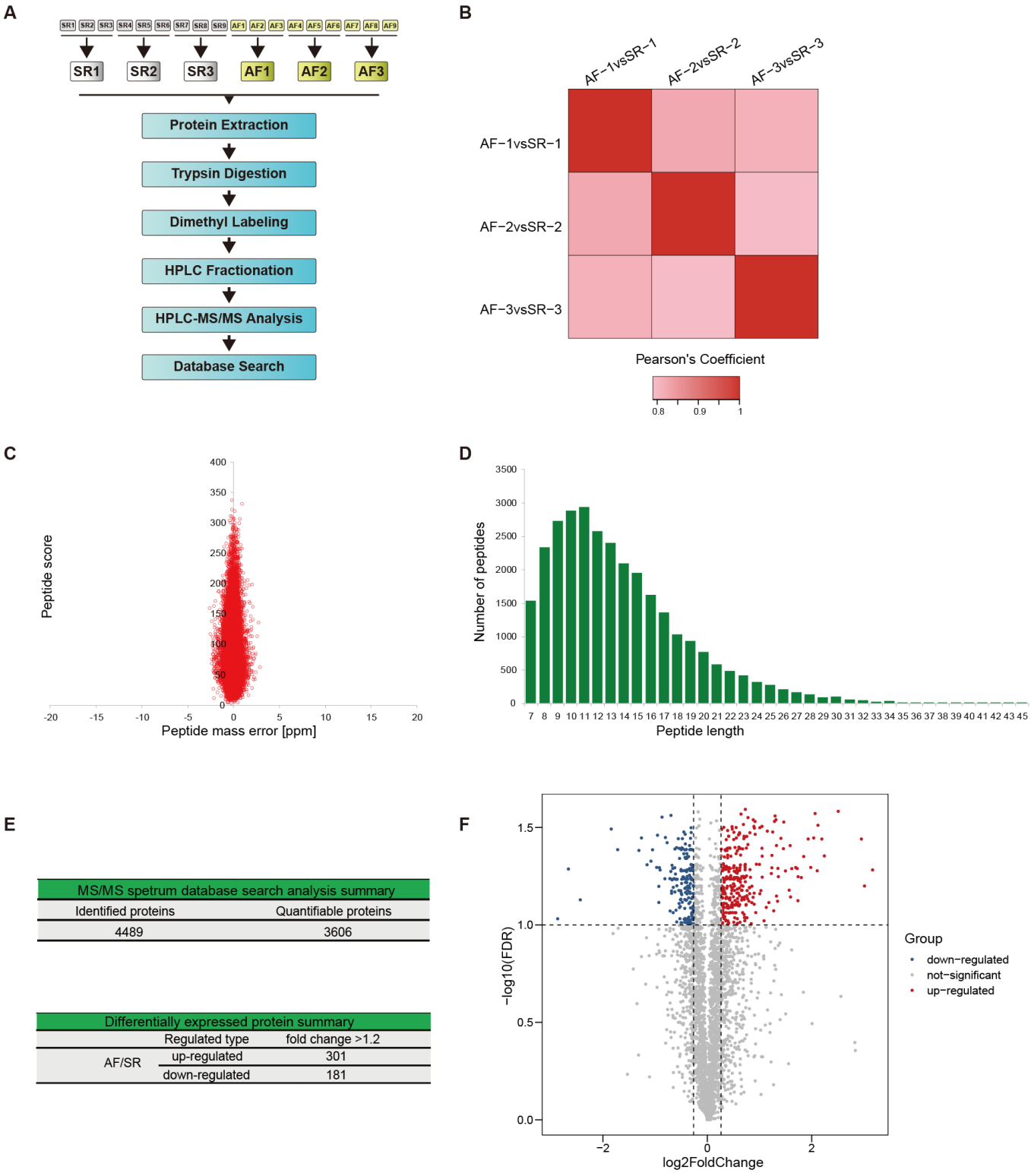
[47] G. Mercuro, P. Bassareo, M. Deidda, C. Cadeddu, L. Barberini, L. Atzori, Metabolomics: a new era in cardiology?, Journal of cardiovascular medicine (Hagerstown, Md.) 12(11) (2011) 800-5.

# Figures



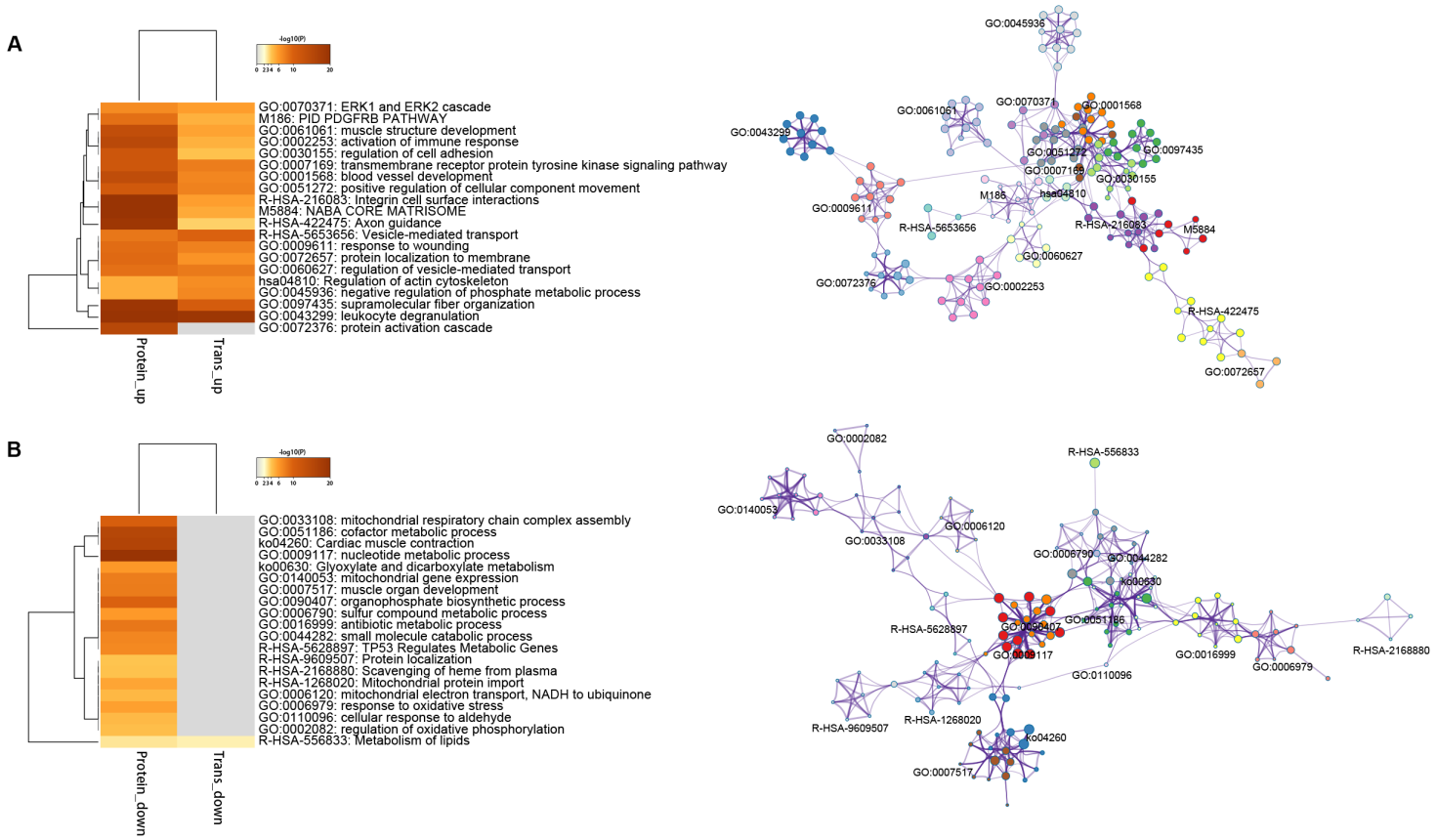

## Figure 1

Quantitative proteomic analysis of AF and SR tissue samples. A. Experimental process; B. Reproducibility of the quantitative proteomic analysis; C. QC validation of MS data. Mass error indicates the distribution of all identified peptides. D. Peptide length distribution identified by quantitative proteomic analysis; E. Identified and quantified proteins; F. Volcano plot of differentiated expressed proteins.



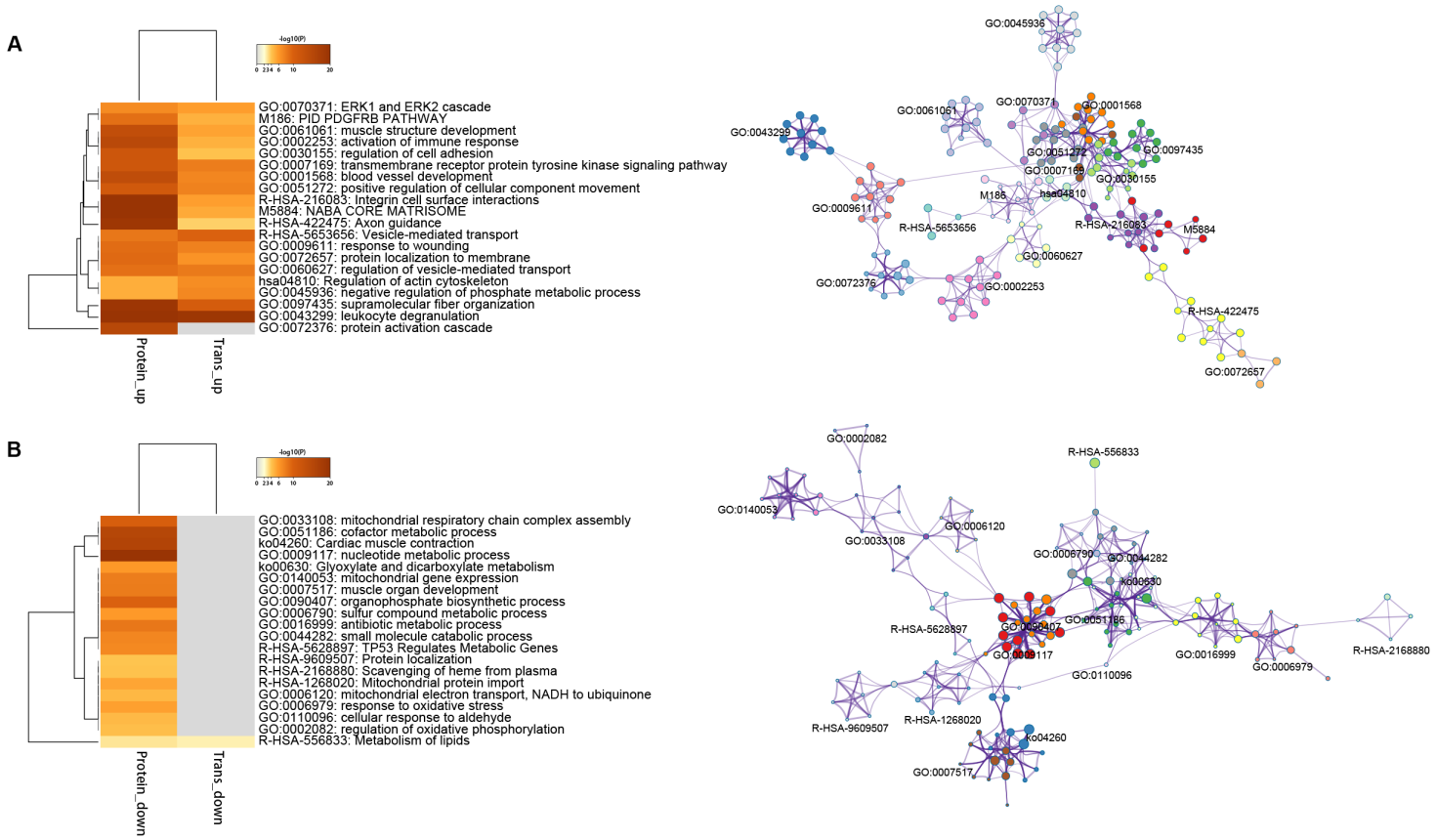

## Figure 1

Quantitative proteomic analysis of AF and SR tissue samples. A. Experimental process; B. Reproducibility of the quantitative proteomic analysis; C. QC validation of MS data. Mass error indicates the distribution of all identified peptides. D. Peptide length distribution identified by quantitative proteomic analysis; E. Identified and quantified proteins; F. Volcano plot of differentiated expressed proteins.



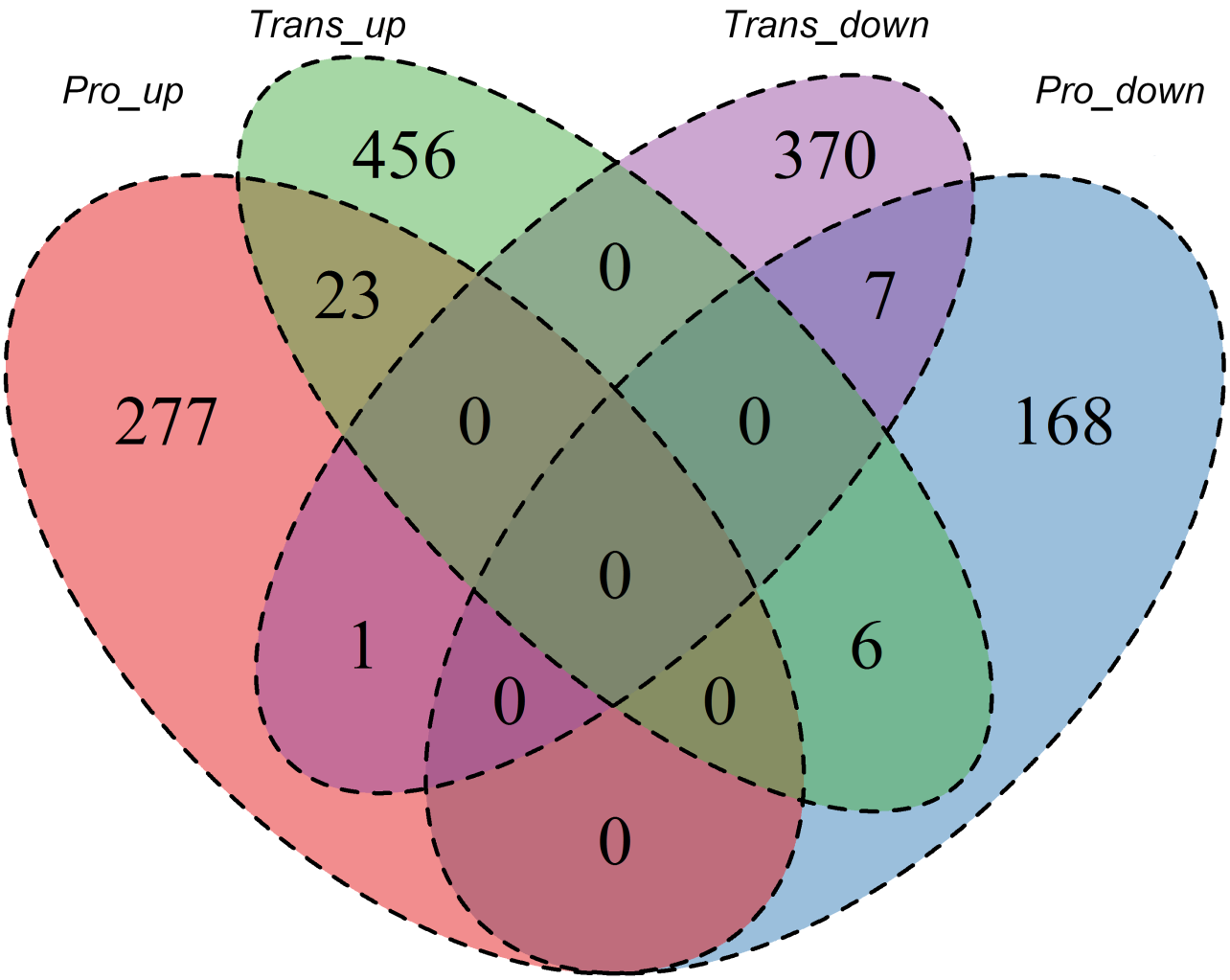

## Figure 2

Pathway enrichment analysis. A. Top 20 clusters with the smallest p-value of upregulated mRNAs/proteins; B. Top 20 clusters with the smallest p-value of downregulated mRNAs/proteins right. The right part displays the network of selected enriched terms. Each term is represented by a circle node, where its size is proportional to the number of input genes that fall into that term, and its color represents its cluster identity (i.e., nodes of the same color belong to the same cluster). Terms with a similarity score > 0.3 are linked by an edge (the thickness of the edge represents the similarity score)




## Figure 2

Pathway enrichment analysis. A. Top 20 clusters with the smallest p-value of upregulated mRNAs/proteins; B. Top 20 clusters with the smallest p-value of downregulated mRNAs/proteins right. The right part displays the network of selected enriched terms. Each term is represented by a circle node, where its size is proportional to the number of input genes that fall into that term, and its color represents its cluster identity (i.e., nodes of the same color belong to the same cluster). Terms with a similarity score > 0.3 are linked by an edge (the thickness of the edge represents the similarity score)

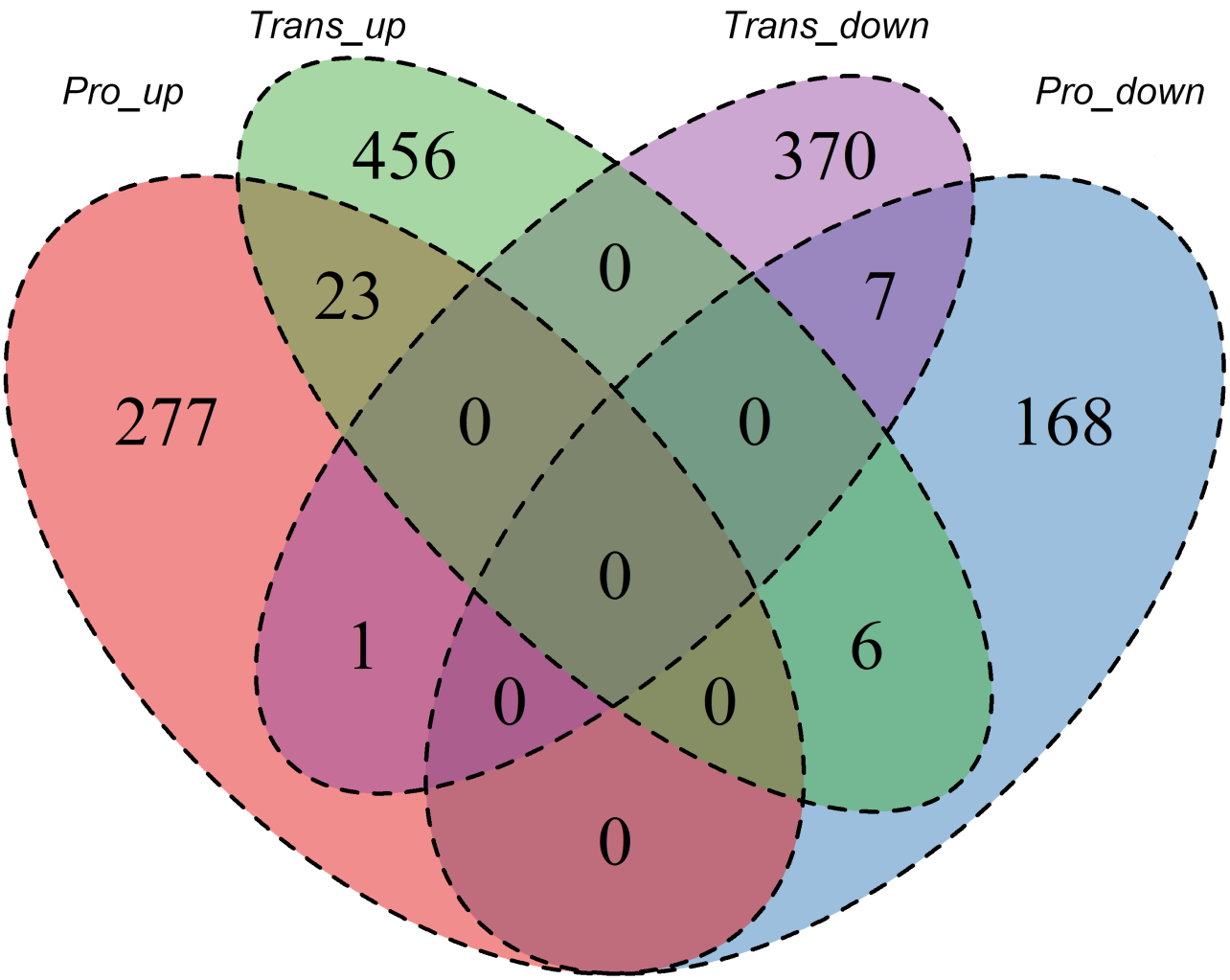



*23 upregulated: BGN, C1QC, CAP1, CD44, CDK4, CHGB, CMYA5, COL12A1, COL4A2, CORO1A, CTSZ, DBNL, FHL2, GGT5, HAPLN1, HLA-DRA, IGFBP2, NRAP, SEPTIN6, SORBS2, THBS4, TSTA3, YWHAQ*

*7 downregulated: ACSF2, ISOC1, MYH6, SLC27A6, TMEM143, TNNI1, TRDN*

**Figure 3**

Vene diagram of DEGs and DEPs.




*23 upregulated: BGN, C1QC, CAP1, CD44, CDK4, CHGB, CMYA5, COL12A1, COL4A2, CORO1A, CTSZ, DBNL, FHL2, GGT5, HAPLN1, HLA-DRA, IGFBP2, NRAP, SEPTIN6, SORBS2, THBS4, TSTA3, YWHAQ*

*7 downregulated: ACSF2, ISOC1, MYH6, SLC27A6, TMEM143, TNNI1, TRDN*

**Figure 3**

Vene diagram of DEGs and DEPs.