Mathematical Biosciences and Engineering

MBE, 18(6): 8622–8640.
DOI: 10.3934/mbe.2021427
Received: 09 August 2021
Accepted: 16 September 2021
Published: 09 October 2021

http://www.aimspress.com/journal/MBE

Research article

Construction of the gene expression subgroups of patients with coronary artery disease through bioinformatics approach

Bin Zhang[1,†], Kuan Zeng[1,†], Rongzhen Li[2,†], Huiqi Jiang[1], Minnan Gao[1], Lu Zhang[1], Jianfen Li[1], Ruicong Guan[1], Yuqiang Liu[1], Yongjia Qiang[1] and Yanqi Yang[1,*]

[1] Department of Cardiovascular Surgery, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China
[2] Department of Radiation Oncology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center of Cancer Medicine, Guangzhou 510000, China

† The authors contributed equally to this work.

* **Correspondence:** Email: yangyq@mail.sysu.edu.cn; Tel: +8615237106082.

Abstract: Coronary artery disease (CAD) is a heterogeneous disease that has placed a heavy burden on public health due to its considerable morbidity, mortality and high costs. Better understanding of the genetic drivers and gene expression clustering behind CAD will be helpful for the development of genetic diagnosis of CAD patients. The transcriptome of 352 CAD patients and 263 normal controls were obtained from the Gene Expression Omnibus (GEO) database. We performed a modified unsupervised machine learning algorithm to group CAD patients. The relationship between gene modules obtained through weighted gene co-expression network analysis (WGCNA) and clinical features was identified by the Pearson correlation analysis. The annotation of gene modules and subgroups was done by the gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. Three gene expression subgroups with the clustering score of greater than 0.75 were constructed. Subgroup I may experience coronary artery disease of an in-creased severity, while subgroup III is milder. Subgroup I was found to be closely related to the upregulation of the mitochondrial autophagy pathway, whereas the genes of subgroup II were shown to be related to the upregulation of the ribosome pathway. The high expression of APOE, NOS1 and NOS3 in the subgroup I suggested that the patients had more severe coronary artery disease. The construction of genetic subgroups of CAD patients has enabled clinicians to improve their understanding of CAD pathogenesis and provides potential tools for disease diagnosis, classification and assessment of prognosis.

Keywords: coronary artery disease; coronary heart disease; myocardial infarction; gene; RNA-Seq; subgroup

Abbreviations: CAD: Coronary artery disease; GEO: Gene expression omnibus; WGCNA: Weighted gene co-expression network analysis; GO: Gene ontology; KEGG: Kyoto encyclopedia of genes and genomes; MI: Myocardial infarction; GSEA: Gene set enrichment analysis; CC: Consensus clustering; PPI: Protein-protein interaction; DEGs: Pifferentially expressed genes; TOM: Topological overlap matrix; PCA: Principal component analysis; Tfrc: Transferring receptor; Sirt6: Stress-responsive proautophagic histone deacetylase; APOE: Apolipoprotein E; NOS1: Nitric oxide synthase 1; NOS3: Nitric oxide synthase 3; RXRA: Retinoid X receptors A; TXNDC9: Thioredoxin domain-containing 9; WDTC1: WD40 and tetratricopeptide repeats 1

1. Introduction

Coronary artery disease (CAD) is one of the major causes of cardiovascular diseases, such as myocardial infarction (MI), ischemic cardiomyopathy, and arrhythmia [1,2]. In the year 2020, statistics from the American Heart Association showed that the prevalence of coronary artery disease (CAD) in U.S. adults who were ≥ 20 years of age was 6.7%, while the annual incidence of MI (heart attack) in the United States totaled up to 805,000 cases, with 605,000 being new cases and 200,000 being recurrent cases [3]. CAD mortality and any-mention mortality were 365,914 and 541,008 respectively; whereas MI mortality and any-mention mortality were 110,346 and 149,028 respectively [3]. The morbidity and mortality of CAD in low- and middle-income countries around Asia and Africa had gradually risen to a level close to that of the Western society, turning it into a global issue [4].

The efficient management of CAD diseases depends on scientific classification and targeted treatment. At present, classifications based on pathological characteristics, disease progression and clinical symptoms have gradually matured, while genotypic subgroup is currently developing at a slower pace [5]. Increasing evidences suggest the involvement of inheritance in the initiation or progression of CAD [6,7]. With the commercialization of DNA microarray and high-throughput sequencing, the study of CAD diseases at the genetic level has been greatly facilitated [8–10]. Sinnaeve et al. of Duke University tested the peripheral blood transcriptome of 110 CAD patients and 112 healthy controls, and 160 genes were found to be related to the CAD index (indicators for assessing the severity of CAD) [11]. Han et al. also analyzed the relationship between atherosclerotic plaque and immune infiltration in CAD patients by using sequencing results from the GEO database [12]. Nevertheless, the establishment of gene expression subgroups to classify CAD patients based on sequencing data is still a field less researched.

A better understanding of the genetic drivers of CAD will help to promote the development of relevant drugs, and even catalyze the inception of a novel gene therapy strategy. Therefore, the transcriptome data of 352 CAD cases were summarized, a consensus clustering analysis was established and the differences between the subgroups were compared in this study. The gene expression subgroup was constructed via the consensus clustering method that was based on an unsupervised learning algorithm. The WGCNA and Gene Set Enrichment Analysis (GSEA) was combined to elaborate the correlations between genes. Henceforth, the results from this study provide the basis for CAD gene grouping.

2. Materials and methods

2.1. Acquisition and processing of GEO microarray data

Gene expression datasets (GSE12288, GSE20680 and GSE20681) of CAD and healthy controls were downloaded from the GEO database of the NCBI (https://www.ncbi.nlm.nih.gov/geo/). GSE12288 came from GPL96, while GSE60280 and GSE60281 were based on GPL4133. Annotations of the three datasets were done using the corresponding platform files. The data were then normalized with the "limma" and "sva" packages using the ComBat method of the R software to eliminate the batch effect [13,14]. The homogenized data were subsequently combined to obtain a final gene expression profile of the 352 CAD patients and 263 normal controls. Clinical information such as age, CAD index, gender and smoking status were also collated into the data. Duke Coronary Artery Disease Index (CAD index) is a prognostic assessment of the extent of coronary artery disease, reflecting the number and severity of lesions and diseased vessels, as well as the involvement of left anterior descending branch and left main stem lesions [15,16]. For example, the CAD index is 0 in patients without coronary artery disease and 23 in patients with at least one stenosis greater than 50%; therefore, the greater the number of vascular lesions and the greater the severity of the stenosis, the greater the CAD index [15].

2.2. Construction of subgroups based on consensus clustering

The gene expression matrix of all 352 CAD patients was extracted for grouping purposes. The algorithm of unsupervised class discovery was implemented with minor modifications to precisely identify CAD patients with shared genetic features. In particular, an estimation of the number of unsupervised classes in the data set can be obtained through quantitative and visual approach via consensus clustering (CC), a kind of unsupervised learning algorithm. ConsensusClusterPlus is a method of extended CC in the R language that exhibits more functions and visualizations, such as project tracking, project consensus, and generating cluster consensus graphs [17]. In short, the homogenized gene matrix was passed to the consensus clustering algorithm (input parameters k = 2–10) to generate the cluster membership of each CAD sample. Upon running the R software, the cophenetic coefficient for the k = 2 to k = 10 clusters and the silhouette values for the "best cluster" (k = 3) were obtained. As shown in Figure 2A, when the clustering variable was 3, all the CAD cases were divided into 3 subgroups, the clustering score of each subgroup was greater than 0.75 with better clustering consensus. The core idea of the consensus cluster was to generate multiple partitions from the dataset to establish an ideally more meaningful consensus subgroup for the data. The clustering score greater than 0.75 in this study indicated that these three subgroups have high similarity in gene expression and can be classified into three subgroups. Further, each subgroup with homogeneity can be regarded as a cluster for in-depth analysis.

2.3. Comparison of clinical characteristics of different subgroups

According to the series matrix file retrieved from the three gene sets of GSE12288, GSE20680 and GSE20681 in the GEO database, clinical information such as age, gender, CAD index, and smoking status were collected. The age and CAD index acted as the continuous variables, and were compared by means and standard deviations before subsequently presented using box plots. The

proportions of patients who were male and smoking were analyzed as categorical variables by their ratio and were presented in a histogram. Patients who are currently smoking, or who had quit smoking within the last two months in the original data were defined as smoking.

2.4. Top gene identification and protein-protein interaction (PPI) network construction

The corresponding differential genes of each subgroup were obtained by comparing between the normal group and the other subgroups. The selection criteria were mean differences that were greater than 0.2 and adjusted P-value of less than 0.05. The top 10 genes with the upregulated and unique in subgroups I, II and III were listed in Table 1. The online PPI analysis website STRING (https://string-db.org/) was used to analyze the top 10 genes of the three subgroups to identify the relationship between each subgroup in terms of protein linkage. The minimum interaction score required was 0.4 (the default parameter). The horizontal histogram shows the 10 genes with the most relationship pairs.

2.5. GSEA analysis of each subgroup

The comparison files of the control group and the subgroups I, II and III were transformed into gene list files and gene set files as required by the Perl software (Version 5) to run the GSEA analysis. These lists and set files were uploaded to the GSEA software for analysis. Run options (max size) were set to 5000 in order to meet the data criteria of a large gene set.

2.6. Construction and analysis of WGCNA

The conversion of data from gene expression profile to scale-free network was carried out by the WGCNA package of R Software [18]. The optimal soft threshold power (soft power = 9) was screened in reference to the standard scale-free network analysis. The adjacency values among all differentially expressed genes (DEGs) and correlation matrices were calculated by the power function. Then, the topological overlap matrix (TOM) and corresponding dissimilarity (1-TOM) values were computed using the adjacency values obtained in the previous step. The identification and stabilization of modules were achieved through the dynamic tree cut method and module preservation function, respectively [19].

The Pearson correlation analysis was used to analyze the correlations between clinical characteristics and gene modules to identify the association between biological modules and age, CAD index, male, and smoking status. The expression of gene modules obtained from the WGCNA analysis in subgroups I, II and III was produced by the heatmap package of R software and was presented in the form of a heatmap [20].

2.7. Functional enrichment analysis of each module and subgroup

The enrichment analysis of the seven selected gene modules on biological processes, cellular components, molecular function and molecular biological pathway were analyzed using the GO database (http://geneontology.org/) and KEGG database (https://www.kegg.jp/kegg/pathway.html). The data obtained from the database analysis was imported into R software, and the results were displayed through the visual bubble diagram generated by the clusterProfiler and enrichplot packages [21,22]. The seven pathways (mitophagy-animal, ribosome, neuroactive ligand-receptor interaction, ovarian

steroidogenesis, steroid biosynthesis, shigellosis and legionellosis) with the most significant upregulation were screened from the KEGG pathways of each module for further analysis. Furthermore, pathway correlation heat map was used to demonstrate detailed gene enrichment in these seven pathways for the control group and the three subgroups.

2.8. Statistical analysis

The statistical software SPSS 25.0 (SPSS Inc., Chicago, IL, USA) and R software (Version 4.0.2) were used for statistical analysis. Mean ± standard deviation was used to describe the unity and discreteness of continuous variables. Chi-square analysis was used to evaluate percentage differences in discrete variables. The data is considered as statistically significant when P-value < 0.05 in a two-tailed test.

3. Results

3.1. Workflow and batch effect removal

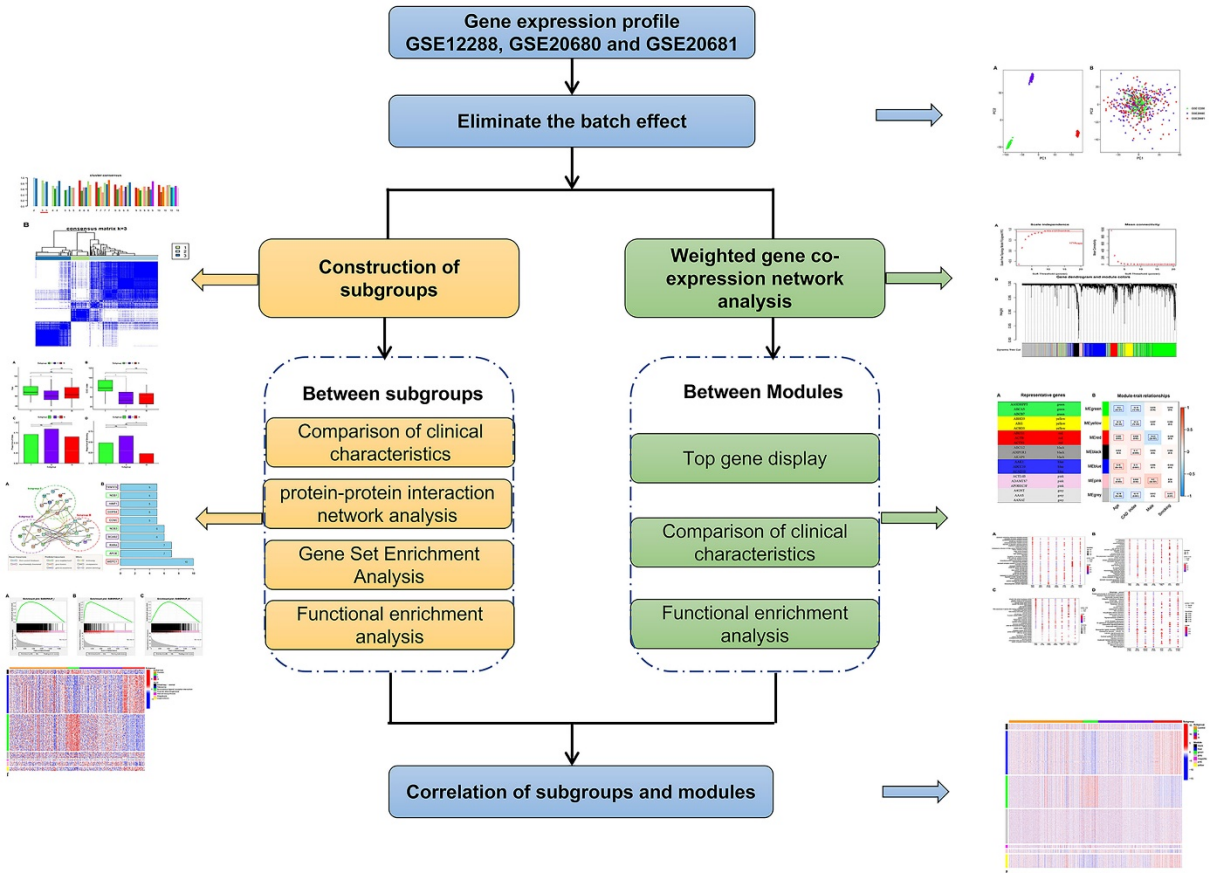

Figure 1. Workflow for the whole study.

The workflow is shown in Figure 1. The batch effect between GSE12288, GSE20680 and GSE20681 was assessed and visualized using a principal component analysis (PCA) cluster diagram, with results showing that the data of the three gene sets were gathered in three different regions,

indicating a batch effect among them (Figure S1A). The gene sets were then normalized by the sav package of R software to remove these batch effects. Figure S1B shows the data from the three gene sets that was processed by the PCA cluster diagram (Figure S1B). The information from GSE12288, GSE20680 and GSE20681 was uniformly distributed and regarded as a synthetic gene set.

3.2. Sample grouping based on the relevance of transcriptome information

The consensus clustering analysis with all 352 CAD samples in the GSE12288, GSE20680 and GSE20681 data sets was performed to explore the connections between each gene subgroup and clinical characteristics of CAD. The clustering variable (k) was set from 2 to 10, resulting in a total of 9 clusters (Figure 2A). The high intragroup correlations and low intergroup correlations between the three subgroups indicated that the 352 CAD patients could be well-divided into 3 clusters based on the transcriptome gene (Figure 2B). Finally, consensus clustering analysis yielded three subgroups, with 54, 196 and 102 cases in subgroups I, II and III respectively, which had significantly distinguished expression patterns.

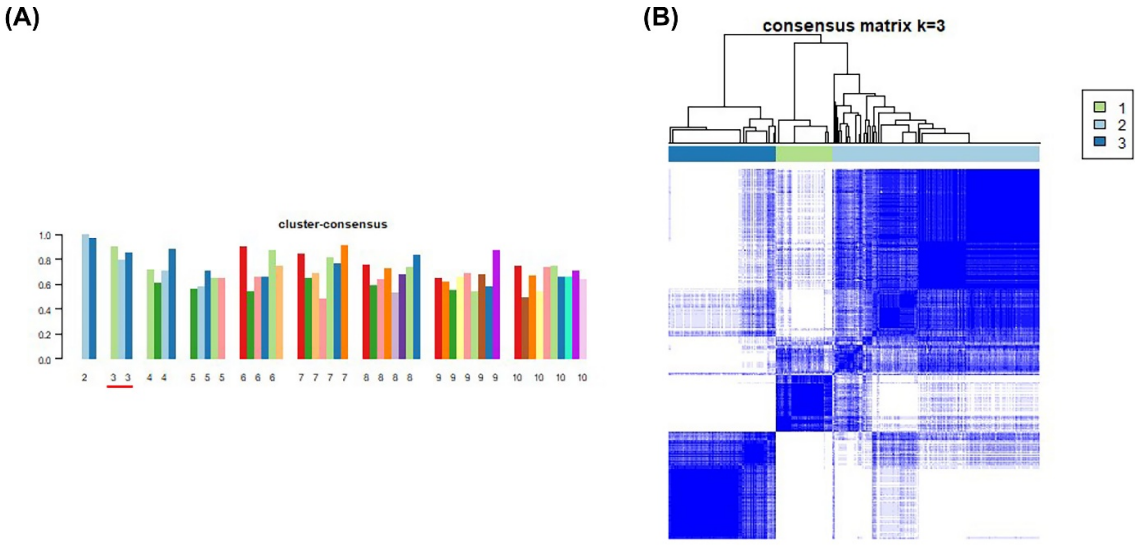

Figure 2. Classification of CAD samples based on transcriptome correlation. (A) The cluster consensus scores of different subgroups include 9 classification methods from 2 to 10 subgroups. The ordinate shows the cluster consensus score. The abscissa shows different grouping situations. (B) Gene expression clustering and correlation map of 3 subgroups. The regularity and color depth of the blue rectangles represent the correlation of genes within different subgroups.

3.3. Analysis of clinical characteristics among the subgroups

The clinical characteristics of CAD patients, such as age, CAD index, gender, and smoking status, were extracted from the platform. The mean and standard deviation for age of the three subgroups were 69.87 ± 12.79 (subgroup I), 62.63 ± 12.90 (subgroup II) and 66.54 ± 15.51 (subgroup III) respectively. As shown in Figure 3A, the statistically significant difference was observed between subgroups I and II (Figure 3A). The mean and standard deviation for CAD index of subgroups I, II and

III were 73.83 ± 25.44, 47.68 ± 19.34 and 44.10 ± 19.92, respectively. The contrast between subgroups I and II, and subgroups I and III were both significant (Figure 3B). Figure 3C,D demonstrate that the proportion of males (84 and 63%) and smoking (64 and 21%) between subgroup II and subgroup III were also statistically valid (Figure 3C,D).

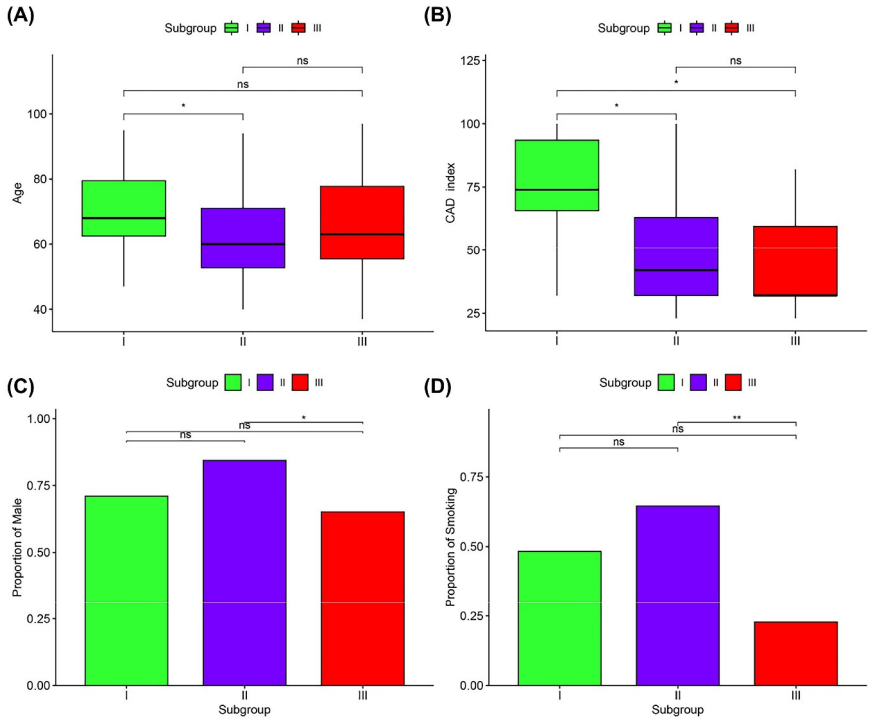

Figure 3. Analysis of differences in clinical characteristics of each subgroup. (A) Box plot of age differences in the 3 subgroups. (B) Box plot of CAD index differences in the 3 subgroups. (C) Histogram of the proportion of male CAD patients in the 3 subgroups. (D) Histogram of the proportion of smoking CAD patients in the 3 subgroups. P-values were showed as: ns: $P > 0.05$; *: $P < 0.05$; **: $P < 0.01$.

3.4. Differential gene display and PPI network analysis of the subgroups

Table 1. Top 10 genes specifically upregulated in the three subgroups.

Subguoup I	Subguoup II	Subguoup III
APOE	RXRA	EED
VGF	ACTN1	CCNC
NOS1	TOM1	CCDC90B
FOXA2	BCKDK	POT1
NOS3	TAGLN2	USP1
GHSR	GNB2	COPS4
CLDN11	CEBPA	PRKRIR
MMP19	MBOAT7	PIBF1
OR1F2P	ZYX	PTPN2
MMP28	CTSA	TXNDC9

The top 10 DEGs that were upregulated and unique among the 3 subgroups are presented in Table 1. The PPI network analysis of the top 10 DEGs in each subgroup is presented in Figure 4A. Thirty nodes selected nodes and approximately forty-six protein pairs were obtained when the confidence coefficient was set to 0.4 (Figure 4A). WDTC1, APOE, RXRA, BCAS2, NOS3, CCNC, COPS4, HINT1, NOS1 and TXNDC9 were found to interact most closely among the 3 subgroups (Figure 4B).

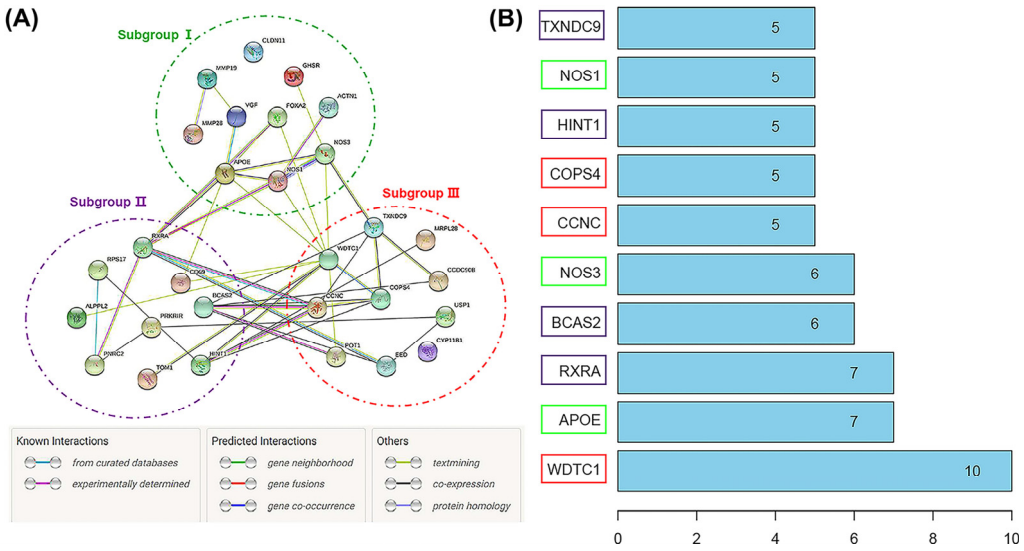

Figure 4. PPI network of the top 10 DEGs in each subgroup. (A) The green dashed circle contains the DEGs of subgroup I. The purple and red dashed circles contain the DEGs of the subgroup II and subgroup III, respectively. The ball and the line represent DEGs and their relationships, respectively. (B) The histogram of the core nodes and their numbers of proteins pairs. The ordinate is the top 10 core node gene, while the abscissa is the number of relationship pairs.

3.5. GSEA analysis of the subgroups

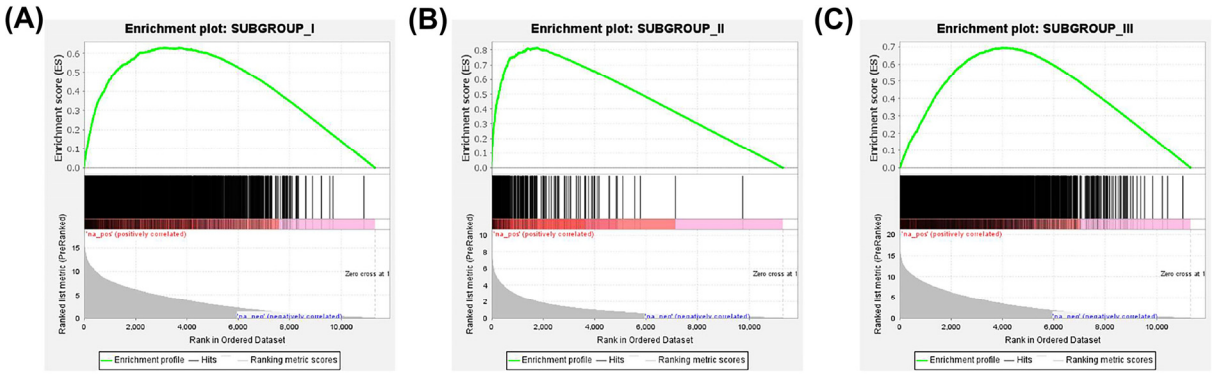

Figure 5. GSEA analysis of the three subgroups. (A) GSEA analysis of subgroup I. The green curve is the value of the enrichment score. The black vertical line is the position of each gene in the gene ranking list. The gray area reflects the signal-to-noise ratio between subgroup I and the control group. (B) GSEA analysis of subgroup II. (C) GSEA analysis of subgroup III.

The black vertical lines in Figure 5 represent the unique DEGs among the subgroups, while the gray vertical lines represent the DEGs between the subgroups and the normal samples (Figure 5). Both sets of data were clustered on the left side of the image. The P-value and FDR value of all 3 subgroups were far less than 0.01. The analysis results of GSEA proved that the unique DEGs between the different subgroups and the DEGs between the subgroups and the normal samples were consistent.

3.6. WGCNA analysis of CAD patients

In the integrated data sets, 352 CAD samples and 263 control samples with 11,314 genes expression profiles were included in the WGCNA. After selecting a soft threshold of 9 (Figure 6A), the weighted co-expression network was constructed based on the determined genes. Seven modules were derived from the gene clustering tree (Figure 6B) based on gene–gene non-ω similarity.

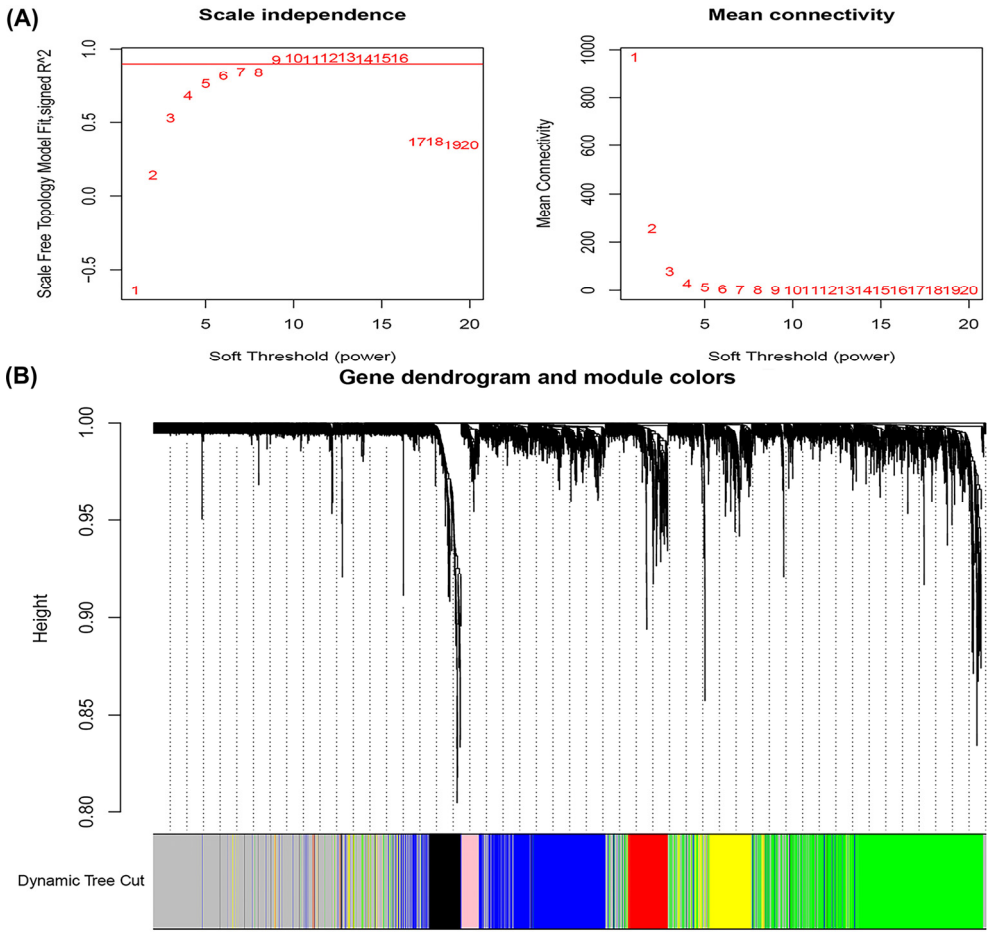

Figure 6. Weighted gene co-expression network analysis. (A) Dissection of network topology for various soft-thresholding powers. The left plot shows the scale-free fitting index (y-axis) as a function of the soft-thresholding power (x-axis). The right plot displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). (B) Clustering dendrogram of genes. The colored row beneath the dendrogram shows the module membership as determined by the dynamic tree cut process, as well as allocated merged module colors and original module colors.

The top three genes with the most significant differences in each gene module are listed in Figure 7A. In addition, the association between each module and the clinical information of CAD cases was examined by mapping clinical data to samples (Figure 7B). The genes in the green, yellow and gray modules were negatively correlated with age and CAD index. The synergistic positive correlation was reflected between the red, black, blue and pink gene modules and both the factors of age and CAD index. Female CAD patients expressed more genes in the red gene module, while male CAD patients expressed more genes that were found in the pink gene module. Surprisingly, smoking had a weak positive correlation with the gray gene module.

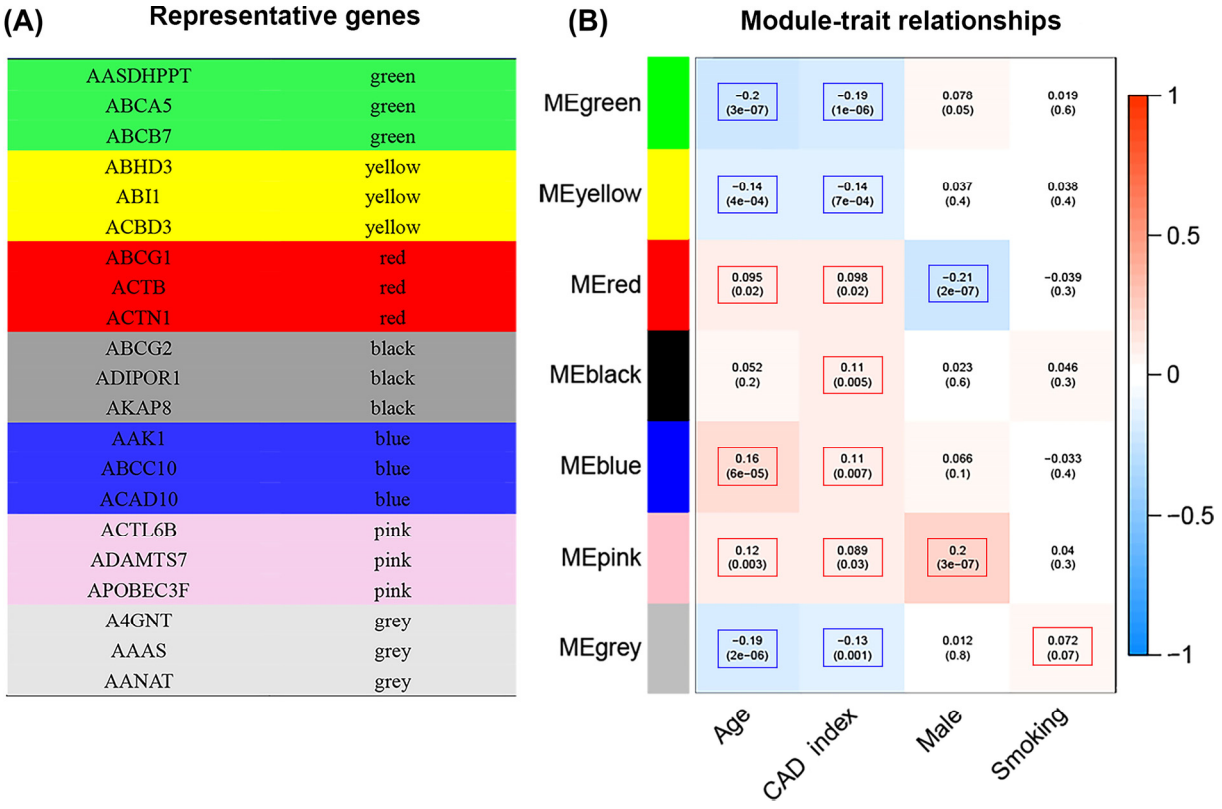

Figure 7. Representative genes and module-trait relationships of the seven gene modules. (A) The three representative genes with the most significant differences in each gene module. (B) The relationships between each module and clinical information. Red represents genes that were upregulated, whereas blue represents genes that were down-regulated. The numbers in the rectangles represent the degree of difference and the p-value.

The correlation between different gene subgroups and the WGCNA modules is shown in Figure 8. In subgroup I, expression of genes was low in the blue and yellow modules, and high in the green module. In contrast, genes in the blue and yellow modules were highly expressed among CAD patients in subgroup III, while low expression was observed for genes in the green module. The gene expression of samples from the healthy control group and subgroup II did not differ significantly in each gene module.

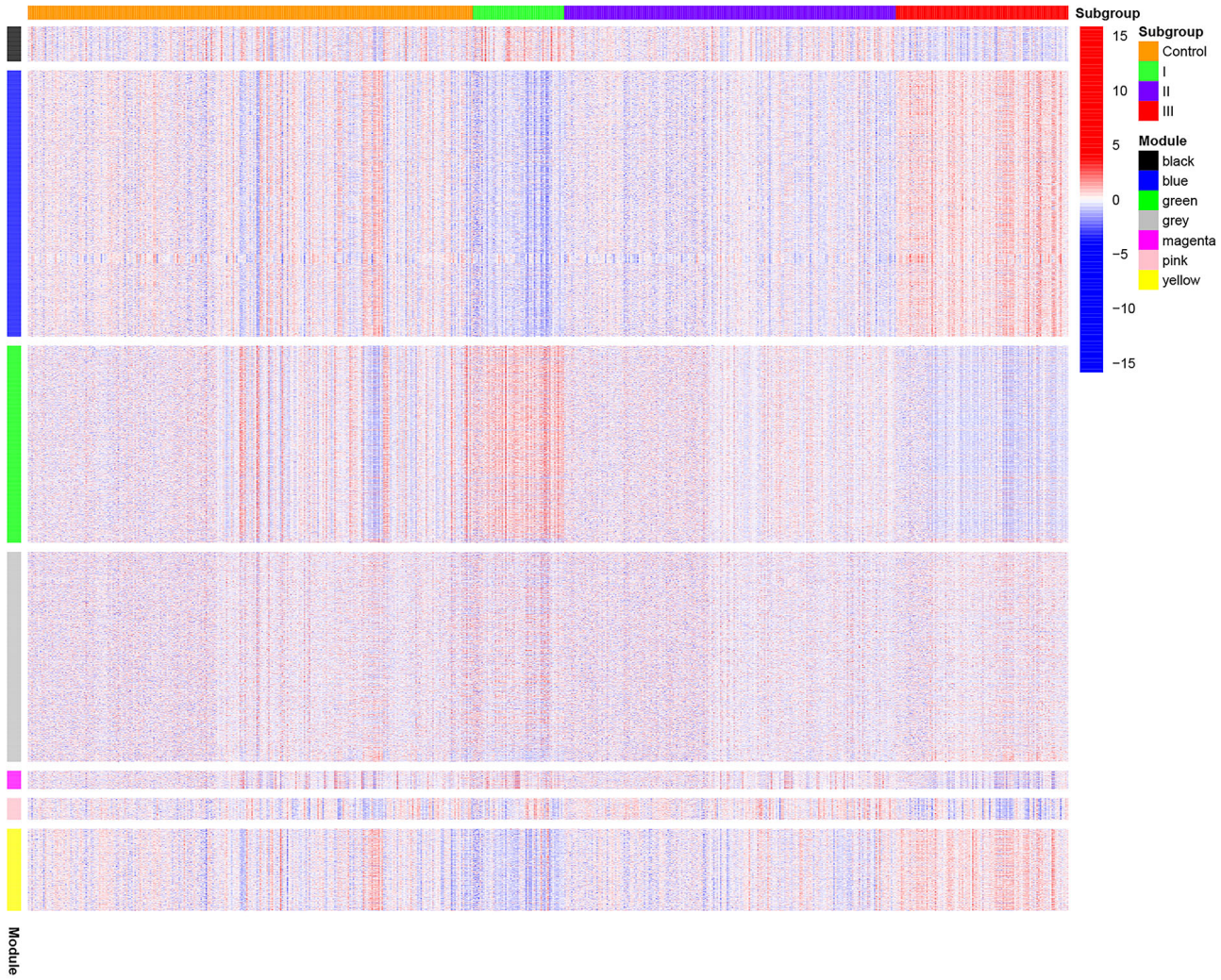

Figure 8. Correlation heat map of different subgroups and gene modules. The orange, green, purple and red modules on the horizontal axis represent the control group, subgroup I, subgroup II and subgroup III, respectively. The seven colors on the vertical axis represent the seven gene modules obtained through WGCNA analysis. Red represents genes that were upregulated, whereas blue represents genes that were down-regulated in the heat map.

3.7. GO and KEGG functional enrichment analysis

Enrichment analysis of the biological process (Figure 9A), cellular component (Figure 9B) and molecular function by GO, (Figure 9C) and the 7 selected gene modules by the KEGG pathway (Figure 9D) were displayed through the improved bubble charts. In the GO analysis, seven gene modules were not enriched in one term at the same time, indicating that the clustering effect of WGCNA was ideal. Genes in the red module were strongly correlated with the biological processes of neutrophils activation, degranulation and immune response. As for cellular component, focal adhesion, cell-substrate junction, secretory granule lumen, cytoplasmic vesicle lumen and secretory granule membrane were related to genes in the pink module. The molecular functions of actin binding, actin filament binding, cell adhesion molecule binding and cadherin binding were

highly related to genes in the pink module as well. KEGG pathway analysis mapped module genes into calcium signaling pathway, autophagy-animal, protein processing in endoplasmic reticulum, neuroactive ligand-receptor interaction, etc. (Figure 9D).

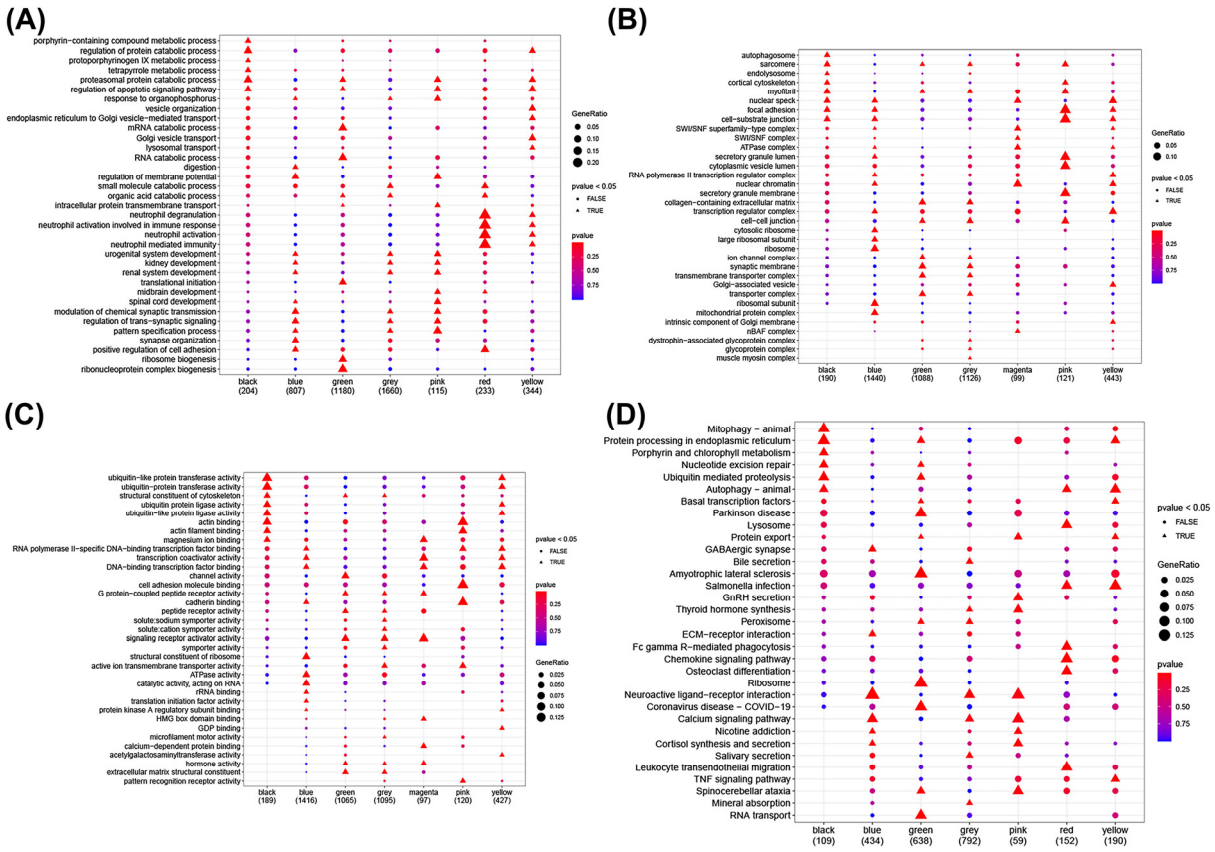

Figure 9. Functional enrichment analysis of module genes. (A) GO enrichment analysis of biological process. The x-axis shows the gene modules and the number of genes they contain. The y-axis represents the terms of biological process. The triangle represents statistically significant enrichment results ($P < 0.05$). The circle indicates that the enrichment results are not statistically significant. The size and color of the triangle and circle represent the gene ratio and P-value respectively. The larger the shape, the higher the gene ratio; the redder the color, the lower the P-value. (B) GO enrichment analysis of cellular components. (C) GO enrichment analysis of molecular function. (D) Enrichment analysis of the KEGG pathway.

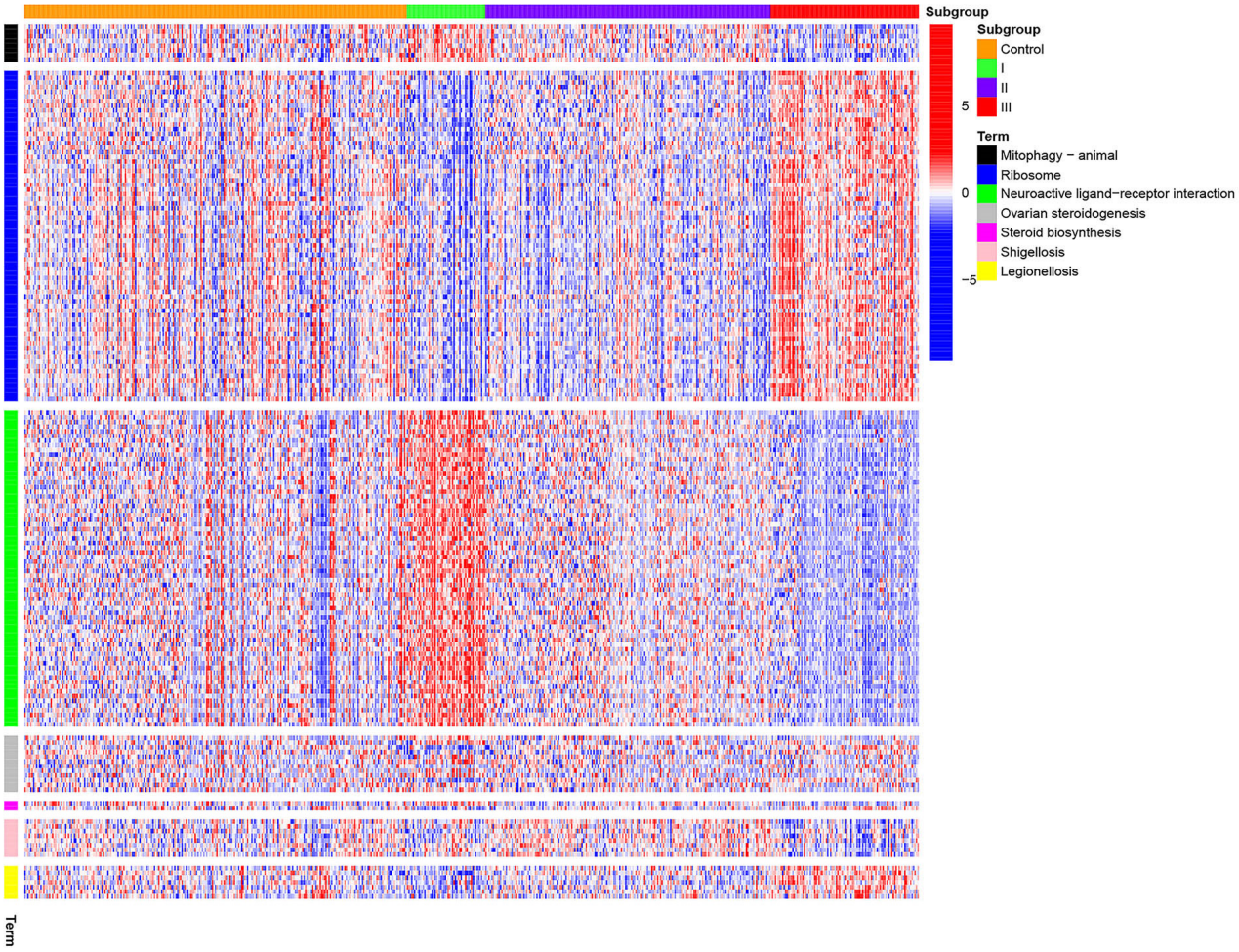

Figure 10. Correlation heat map of different subgroups and KEGG enrichment results. The orange, green, purple and red modules on the horizontal axis represent the control group, subgroup I, subgroup II and subgroup III respectively. The seven colors on the vertical axis represent the pathways related to mitophagy-animal, ribosome, neuroactive ligand-receptor interaction, ovarian steroidogenesis, steroid biosynthesis, shigellosis and legionellosis. In the heat map, red indicates genes that were upregulated, whereas blue shows genes that were downregulated in the related pathways.

To further investigate the KEGG pathway enrichment in different subgroups of CAD patients and control samples, KEGG heat maps (Figure 10) were presented. The normal samples had no statistical significance in selected pathways, mitophagy-animal, ribosome, neuroactive ligand-receptor interaction, ovarian steroidogenesis, steroid biosynthesis, shigellosis and legionellosis. The genes involved in the mitophagy-animal pathway were highly expressed in subgroup I and lowly expressed in subgroup III. The genes associated with the ribosome pathway experienced a low expression in subgroup I and high expression in subgroup III. The genes related to the neuroactive ligand-receptor interaction pathway were significantly active in subgroup I of CAD patients, while very inactive in subgroup III. Genes in subgroup II may be inhibited along the ribosome pathway and were not specifically found in other pathways.

4. Discussion

Coronary artery disease is a multifactorial disease that imposes tremendous economic burden on modern society besides causing anxiety among cardiologists [23]. Based on the transcriptome of coronary artery disease, this study utilized unsupervised learning algorithm to divide CAD patients into different subgroups, and then studied the differences between each subgroup. The correlation between each subgroup and the clinical characteristics of CDA, and the correlation between transcriptome differences among each subgroup were elaborated through combining the WGCNA analysis and the GSEA analysis. The subgroup analysis in this study may improve our understanding of coronary artery disease and provide a theoretical basis for clinical classification based on transcriptomics.

Many significant strides have been made in understanding the genetics of CAD, which put the development of CAD-related gene groupings on the agenda [24–26]. Our research made a challenging attempt based on the transcriptome data of 352 CAD patients. Peng et al. conducted a study on the molecular subtypes of CAD patients for the first time in 2019, but the deep associations between the subgroups need to be further explored [27]. In this study, the high expression of subgroup I, alongside high age and CAD index, in the animal mitophagy and neuroactive ligand-receptor interaction pathways may suggest dysfunctions related to mitochondrial function and nerve activation in this group of CAD patients. The protective effect of mitophagy in cardiovascular disease has been well-reviewed [28]. Tfrc (transferring receptor), Fbxo32 (atrogin-1/MAFbx) and Sirt6 (stress-responsive proautophagic histone deacetylase) are proteins involved in the process of autophagy [29–31]. In animal experiments, Tfrc-/-, Fbxo32-/- and Sirt6-/- mice spontaneously developed mitochondrial respiration failure, cardiac hypertrophy, poor cardiac function, heart failure and even death [32–34]. Interestingly, this phenotype could be partially alleviated through iron supplementation and the administration of nicotinamide riboside (which potently induces autophagy and mitophagy). According to Figure 3, it is believed that older age and higher CAD index in subgroup I indicate coronary heart disease of higher severity. In other words, differential expression of genes associated with age and CAD index contributed to the establishment of the subgroup I. In Figure 10, the high expression of mitophagy pathways in subgroup I proved that mitophagy was activated to maintain the mitochondrial quality control system, thereby resisting the damage inflicted on cardiovascular tissues by severe atherosclerosis [35]. Likewise, the activity of the neuroactive ligand-receptor interaction pathway in subgroup I also showed that this pathway was involved in cardiovascular protection. This conclusion had been notably proven by the experiment of Wang et al. in 2017 [36]. Looking at subgroup III with the low CAD index (Figure 3), the contrasting performance of mitophagy pathway and neuroactive ligand-receptor interaction pathway once again confirmed our inference.

In this study, the transcriptome analysis of CAD patients was carried out by unsupervised deep learning algorithm with the WGCNA and GSEA methods. The DEGs of the 3 subgroups were divided by consensus clustering, and the unique genes of each subgroup were shown in Table 1. The activation of APOE (apolipoprotein E), NOS1 (nitric oxide synthase 1) and NOS3 (nitric oxide synthase 3) was the characteristic of subgroup I. APOE, a key regulator of plasma lipids, positively correlates with the content of triglycerides in the blood of CAD patients [37,38]. The significance of different APOE polymorphisms on CAD risk remained unclear, while ε4+ in APOE may increase the severity of CAD [39,40]. In addition, APOE knockout mice fed a Western diet were also classic mouse model of atherosclerosis [41]. NOS1 (neuronal-NOS, nNOS) and NOS3 (endothelial-NOS, eNOS) were both calcium-dependent and

belong to the constitutive class of NOS [42]. The NOS1 released from parasympathetic postganglionic (nitrergic) neurons could produce NO to reduce vascular resistance and increase blood flow [43]. The NOS3 protein synthesizes NO by converting L-arginine to L-citrulline [44].

RXRA (retinoid X receptors A), a member of the RXR family which belongs to the nuclear receptor superfamily, could be activated by sterol. As reviewed by Ahuja et al., RXRs act as key regulators in the metabolism of glucose, fatty acid and cholesterol, as well as in metabolic disorders such as type 2 diabetes, hyperlipidemia and atherosclerosis by activating multiple nuclear receptors [45]. Lima et al. conducted a free routine blood test in 622 healthy subjects of European ancestry, excluding diabetes and secondary dyslipidemia due to renal, liver or thyroid disease, to evaluate the influence of the RXRA polymorphisms on lipid and lipoprotein levels [46]. Therefore, the slight elevation of RXRA in subgroup II deserves more attention due to the possible effects on numerous metabolic signaling pathways.

TXNDC9 (Thioredoxin domain-containing 9) belongs to the thioredoxins family and was the most diverse gene in subgroup II. Recently, research have shown the role of TXDCT9 in both the fields of cardiovascular diseases and cancer. The research by Zhou et al. found that TDX could regulate the homeostasis of colorectal cancer cells during apoptosis and autophagy, thereby affecting tumor development [47]. The regulation of oxidative stress-induced androgen receptor signaling led to the progression of prostate cancer, which may be related to the higher number of males in subgroup II (Figure 3C) [48]. The precise role of TXNDC9 in apoptosis, autophagy, oxidative stress and androgen in coronary heart disease is worth finding out.

WDTC1 (WD40 and tetratricopeptide repeats 1), the mammalian homolog of Adp, was lowly expressed in subgroup III besides being the gene with the most correlation pairs in the PPI network. Adipose (Adp) is an evolutionarily conserved gene that can be isolated from naturally occurring obese flies with homozygous adp mutation [49]. The deletion of a single WDTC1 allele caused poor metabolic profiles and insulin resistance in obese mice. Conversely, transgenic expression of WDTC1 in fat cells yielded lean mice [50]. Lai et al. examined 935 and 1115 adults of 2 ethnically diverse U.S. populations for polymorphisms in the WDTC1 gene [51]. The results suggested that WDTC1 variants may be an important risk factor for obesity in these populations. The aforementioned evidences suggested that subgroup III with lower expression of WTDC1 may have more obese patients, which in turn led to coronary heart disease.

The findings of this study comprehensively elaborated the possible subtypes of CAD patients based on analyses at the molecular level and introduced the characteristics of each subtype. Nevertheless, these results should first be validated by prospective studies in larger populations before they can be used in clinical practice. Some limitations in this study need to be addressed too. Firstly, the consistency of the CAD subgroups requires verification by more data. Secondly, the specific genes of each subgroup need to be verified by cytology, zoology and even human tissue specimens. Thirdly, this study involved only transcriptome data. The addition of more omics, such as proteomics and metabolomics, may improve the precision of the consensus clustering. Lastly, the GEO database lacks detailed clinical features, which would be conducive to the integration of gene grouping and clinical typing.

5. Conclusions

The present study provided an outline for gene groupings in CAD patients, analyzed the differences between each subgroup and annotated the unique genes of each group. The results of this

research will be helpful in the clinical application of transcriptome-based CAD patient classification. The high expression of APOE, NOS1 and NOS3 in the subgroup I suggested that the patients had more severe coronary artery disease.

6. Research highlights

1) Patients grouped according to transcriptome have distinct characteristics.

2) Subgroup I may experience coronary artery disease of an increased severity, while subgroup III is milder.

3) Genes in group I (APOE, NOS1 and NOS3), group II (RXRA and TXNDC9) and group III (WDTC1) play a major role in the development of coronary artery disease.

4) Bioinformatics provides a new perspective for the study of pathogenesis of CAD.

Acknowledgements

We acknowledge Gene Expression Omnibus (GEO) database for the platform and the respective contributors for uploading their meaningful datasets. This work was supported the 3 × 3 Clinical Scientist Fund of Sun Yat-sen Memorial Hospital (1320900026), the National Natural Science Foundation for Young Scientists of China (81600245), and the Guangdong Science and Technology Department (2020B1212060018).

Conflict of interest

The authors declared that they have no conflict of interest.

References

1. J. Knuuti, W. Wijns, A. Saraste, D. Capodanno, E. Barbato, C. Funck-Brentano, et al., 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes, *Eur. Heart J.*, **41** (2020), 407–477.
2. C. Weber, H. Noels, Atherosclerosis: current pathogenesis and therapeutic options, *Nat. Med.*, **17** (2011), 1410–1422.
3. S. S. Virani, A. Alonso, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, et al., Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association, *Circulation*, **141** (2020), e139–e596.
4. S. A. Sherif, O. O. Tok, Ö. Taşköylü, O. Goktekin, I. D. Kilic, Coronary Artery Aneurysms: A Review of the Epidemiology, Pathophysiology, Diagnosis, and Treatment, *Front. Cardiovasc. Med.*, **4** (2017), 24.
5. A. Davies, K. Fox, A. R. Galassi, S. Banai, S. Ylä-Herttuala, T. F. Lüscher, Management of refractory angina: an update, *Eur. Heart J.*, **42** (2021), 269–283.
6. A. V. Khera, S. Kathiresan, Genetics of coronary artery disease: discovery, biology and clinical translation, *Nat. Rev. Genet.*, **18** (2017), 331–344.
7. S. Kang, Y. Ye, G. Xia, H. B. Liu, Coronary artery disease: differential expression of ceRNAs and interaction analyses, *Ann. Transl. Med.*, **9** (2021), 229.

8. Y. Y. Li, H. Wang, X. X. Yang, H. Y. Geng, G. Gong, X. Z. Lu, PCSK9 Gene E670G Polymorphism and Coronary Artery Disease: An Updated Meta-Analysis of 5,484 Subjects, *Front. Cardiovasc. Med.*, **7** (2020), 582865.
9. K. Musunuru, S. Kathiresan, Genetics of Common, Complex Coronary Artery Disease, *Cell*, **177** (2019), 132–145.
10. M. Franchini, Genetics of the acute coronary syndrome, *Ann. Transl. Med.*, **4** (2016), 192.
11. P. R. Sinnaeve, M. P. Donahue, P. Grass, D. Seo, J. Vonderscher, S. D. Chibout, et al., Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease, *PLoS One*, **4** (2009), e7037.
12. H. Han, R. Du, P. Cheng, J. Zhang, Y. Chen, G. Li, Comprehensive Analysis of the Immune Infiltrates and Aberrant Pathways Activation in Atherosclerotic Plaque, *Front. Cardiovasc. Med.*, **7** (2020), 602345.
13. M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, et al., limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Res.*, **43** (2015), e47.
14. J. T. Leek, J. D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genet.*, **3** (2007), 1724–1735.
15. D. B. Mark, C. L. Nelson, R. M. Califf, F. E. Harrell, K. L. Lee, R. H. Jones, et al., Continuing evolution of therapy for coronary artery disease. Initial results from the era of coronary angioplasty, *Circulation*, **89** (1994), 2015–2025.
16. G. M. Felker, L. K. Shaw, C. M. O'Connor, A standardized definition of ischemic cardiomyopathy for use in clinical research, *J. Am. Coll Cardiol.*, **39** (2002), 210–218.
17. M. D. Wilkerson, D. N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, *Bioinformatics*, **26** (2010), 1572–1573.
18. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics*, **9** (2008), 559.
19. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis, *Stat. Appl. Genet. Mol. Biol.*, **4** (2005), 17.
20. G. M. Li, C. L. Zhang, R. P. Rui, B. Sun, W. Guo, Bioinformatics analysis of common differential genes of coronary artery disease and ischemic cardiomyopathy, *Eur. Rev. Med. Pharmacol. Sci.*, **22** (2018), 3553–3569.
21. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, *Omics*, **16** (2012), 284–287.
22. G. Yu, Gene Ontology Semantic Similarity Analysis Using GOSemSim, *Methods Mol. Biol.*, **2117** (2020), 207–215.
23. R. Bauersachs, U. Zeymer, J. B. Brière, C. Marre, K. Bowrin, M. Huelsebeck, Burden of Coronary Artery Disease and Peripheral Artery Disease: A Literature Review, *Cardiovasc. Ther.*, **2019** (2019), 8295054.
24. H. Turpeinen, E. Raitoharju, A. Oksanen, N. Oksala, M. Levula, L. P. Lyytikäinen, et al., Proprotein convertases in human atherosclerotic plaques: the overexpression of FURIN and its substrate cytokines BAFF and APRIL, *Atherosclerosis*, **219** (2011), 799–806.
25. Y. Li, D. W. Wang, Y. Chen, C. Chen, J. Guo, S. Zhang, et al., Genome-Wide Association and Functional Studies Identify SCML4 and THSD7A as Novel Susceptibility Genes for Coronary Artery Disease, *Arterioscl. Thromb. Vasc. Biol.*, **38** (2018), 964–975.

26. A. Busch, S. M. Eken, L. Maegdefessel, Prospective and therapeutic screening value of non-coding RNA as biomarkers in cardiovascular disease, *Ann. Transl. Med.*, **4** (2016), 236.
27. X. Y. Peng, Y. Wang, H. Hu, X. J. Zhang, Q. Li, Identification of the molecular subgroups in coronary artery disease by gene expression profiles, *J. Cell Physiol.*, 2019.
28. J. M. B. Pedro, G. Kroemer, L. Galluzzi, Autophagy and Mitophagy in Cardiovascular Disease, *Circ. Res.*, **120** (2017), 1812–1824.
29. G. Salazar, A. Cullen, J. Huang, Y. Zhao, A. Serino, L. Hilenski, et al., SQSTM1/p62 and PPARGC1A/PGC-1alpha at the interface of autophagy and vascular senescence, *Autophagy*, **16** (2020), 1092–1110.
30. J. D. Murdoch, C. M. Rostosky, S. Gowrisankaran, A. S. Arora, S. F. Soukup, R. Vidal, et al., Endophilin-A Deficiency Induces the Foxo3a-Fbxo32 Network in the Brain and Causes Dysregulation of Autophagy and the Ubiquitin-Proteasome System, *Cell Rep.*, **17** (2016), 1071–1086.
31. Y. Chen, Y. Zhao, W. Chen, L. Xie, Z. A. Zhao, J. Yang, et al., MicroRNA-133 overexpression promotes the therapeutic efficacy of mesenchymal stem cells on acute myocardial infarction, *Stem. Cell Res. Ther.*, **8** (2017), 268.
32. W. Xu, T. Barrientos, L. Mao, H. A. Rockman, A. A. Sauve, N. C. Andrews, Lethal Cardiomyopathy in Mice Lacking Transferrin Receptor in the Heart, *Cell Rep.*, **13** (2015), 533–545.
33. T. Zaglia, G. Milan, A. Ruhs, M. Franzoso, E. Bertaggia, N. Pianca, et al., Atrogin-1 deficiency promotes cardiomyopathy and premature death via impaired autophagy, *J. Clin. Invest.*, **124** (2014), 2410–2424.
34. N. R. Sundaresan, P. Vasudevan, L. Zhong, G. Kim, S. Samant, V. Parekh, et al., The sirtuin SIRT6 blocks IGF-Akt signaling and development of cardiac hypertrophy by targeting c-Jun, *Nat. Med.*, **18** (2012), 1643–1650.
35. R. R. Bartz, H. B. Suliman, C. A. Piantadosi, Redox mechanisms of cardiomyocyte mitochondrial protection, *Front. Physiol.*, **6** (2015), 291.
36. J. Wang, J. Cheng, C. Zhang, X. Li, Cardioprotection Effects of Sevoflurane by Regulating the Pathway of Neuroactive Ligand-Receptor Interaction in Patients Undergoing Coronary Artery Bypass Graft Surgery, *Comput. Math. Methods Med.*, **2017** (2017), 3618213.
37. C. J. Willer, S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, et al., Newly identified loci that influence lipid concentrations and risk of coronary artery disease, *Nat. Genet.*, **40** (2008), 161–169.
38. H. Tada, M. A. Kawashiri, A. Nomura, R. Teramoto, K. Hosomichi, A. Nohara, et al., Oligogenic familial hypercholesterolemia, LDL cholesterol, and coronary artery disease, *J. Clin. Lipidol.*, **12** (2018), 1436–1444.
39. J. P. Karjalainen, N. Mononen, N. Hutri-Kähönen, M. Lehtimäki, M. Hilvo, D. Kauhanen, et al., New evidence from plasma ceramides links apoE polymorphism to greater risk of coronary artery disease in Finnish adults, *J. Lipid Res.*, **60** (2019), 1622–1629.
40. Y. Long, X. T. Zhao, C. Liu, Y. Y. Sun, Y. T. Ma, X. Y. Liu, et al., A Case-Control Study of the Association of the Polymorphisms of MTHFR and APOE with Risk Factors and the Severity of Coronary Artery Disease, *Cardiology*, **142** (2019), 149–157.
41. J. Li, S. Lin, P. M. Vanhoutte, C. W. Woo, A. Xu, Akkermansia Muciniphila Protects Against Atherosclerosis by Preventing Metabolic Endotoxemia-Induced Inflammation in Apoe-/- Mice, *Circulation*, **133** (2016), 2434–2446.

42. N. Toda, S. Tanabe, S. Nakanishi, Nitric oxide-mediated coronary flow regulation in patients with coronary artery disease: recent advances, *Int. J. Angiol.*, **20** (2011), 121–134.
43. N. Toda, T. Okamura, The pharmacology of nitric oxide in the peripheral nervous system of blood vessels, *Pharmacol. Rev.*, **55** (2003), 271–324.
44. J. Qian, D. Fulton, Post-translational regulation of endothelial nitric oxide synthase in vascular endothelium, *Front. Physiol*., **4** (2013), 347.
45. H. S. Ahuja, A. Szanto, L. Nagy, P. J. Davies, The retinoid X receptor and its ligands: versatile regulators of metabolic function, cell differentiation and cell death, *J. Biol. Regul. Homeost. Agents*, **17** (2003), 29–45.
46. L. O. Lima, S. Almeida, M. H. Hutz, M. Fiegenbaum, PPARA, RXRA, NR1I2 and NR1I3 gene polymorphisms and lipid and lipoprotein levels in a Southern Brazilian population, *Mol. Biol. Rep.*, **40** (2013), 1241–1247.
47. W. Zhou, C. Fang, L. Zhang, Q. Wang, D. Li, D. Zhu, Thioredoxin domain-containing protein 9 (TXNDC9) contributes to oxaliplatin resistance through regulation of autophagy-apoptosis in colorectal adenocarcinoma, *Biochem. Biophys. Res. Commun.*, **524** (2020), 582–588.
48. T. Feng, R. Zhao, F. Sun, Q. Lu, X. Wang, J. Hu, et al., TXNDC9 regulates oxidative stress-induced androgen receptor signaling to promote prostate cancer progression, *Oncogene*, **39** (2020), 356–367.
49. W. W. Doane, Developmental physiology of the mutant female sterile(2)adipose of Drosophila melanogaster. I. Adult morphology, longevity, egg production, and egg lethality, *J. Exp. Zool.*, **145** (1960), 1–21.
50. J. M. Suh, D. Zeve, R. McKay, J. Seo, Z. Salo, R. Li, et al., Adipose is a conserved dosage-sensitive antiobesity gene, *Cell Metab.*, **6** (2007), 195–207.
51. C. Q. Lai, L. D. Parnell, D. K. Arnett, B. García-Bailo, M. Y. Tsai, E. K. Kabagambe, et al., WDTC1, the ortholog of Drosophila adipose gene, associates with human obesity, modulated by MUFA intake, *Obesity (Silver Spring).*, **17** (2009), 593–600.

AIMS Press