BioData Mining

RESEARCH

Open Access

The biomedical knowledge graph of symptom phenotype in coronary artery plaque: machine learning-based analysis of real-world clinical data

Jia-Ming Huan[1], Xiao-Jie Wang[1], Yuan Li[1], Shi-Jun Zhang[1], Yuan-Long Hu[1] and Yun-Lun Li[1,2,3*]

*Correspondence:
li.yunlun@163.com

[1] First School of Clinical Medicine, Shandong University of Traditional Chinese Medicine, Jinan 250355, China
[2] Department of Cardiovascular, Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan 250014, China
[3] Precision Diagnosis and Treatment of Cardiovascular Diseases with Traditional Chinese Medicine Shandong Engineering Research Center, Jinan 250355, China

Abstract

A knowledge graph can effectively showcase the essential characteristics of data and is increasingly emerging as a significant means of integrating information in the field of artificial intelligence. Coronary artery plaque represents a significant etiology of cardiovascular events, posing a diagnostic challenge for clinicians who are confronted with a multitude of nonspecific symptoms. To visualize the hierarchical relationship network graph of the molecular mechanisms underlying plaque properties and symptom phenotypes, patient symptomatology was extracted from electronic health record data from real-world clinical settings. Phenotypic networks were constructed utilizing clinical data and protein–protein interaction networks. Machine learning techniques, including convolutional neural networks, Dijkstra's algorithm, and gene ontology semantic similarity, were employed to quantify clinical and biological features within the network. The resulting features were then utilized to train a K-nearest neighbor model, yielding 23 symptoms, 41 association rules, and 61 hub genes across the three types of plaques studied, achieving an area under the curve of 92.5%. Weighted correlation network analysis and pathway enrichment were subsequently utilized to identify lipid status-related genes and inflammation-associated pathways that could help explain the differences in plaque properties. To confirm the validity of the network graph model, we conducted coexpression analysis of the hub genes to evaluate their potential diagnostic value. Additionally, we investigated immune cell infiltration, examined the correlations between hub genes and immune cells, and validated the reliability of the identified biological pathways. By integrating clinical data and molecular network information, this biomedical knowledge graph model effectively elucidated the potential molecular mechanisms that collude symptoms, diseases, and molecules.

Keywords: Coronary artery plate, Biomedical knowledge graph, Symptom phenotypes, Machine learning, Network analysis, Clinical decision support

Introduction

There is mounting evidence that adverse cardiovascular events in patients with chronic ischemic heart disease are linked to the overall burden of atherosclerosis [1–3]. Coronary artery stenosis caused by myocardial ischemia is a common manifestation of coronary artery disease (CAD). Despite secondary preventive treatment, ischemic heart disease remains the leading cause of mortality and morbidity, with a high incidence of cardiovascular events [4]. While plaques typically do not rupture in the early stages of percutaneous coronary intervention (PCI) treatment, the risk in subsequent years still largely stems from coronary artery disease [5]. Early diagnosis and treatment of coronary atherosclerosis can thus significantly alleviate the disease burden of patients.

The symptom phenotype, which includes symptoms and signs, reflects the clinical characteristics of diseases and plays a vital role in disease diagnosis and treatment. In clinical practice, doctors mainly rely on the symptom information provided by patients to diagnose CAD. Junior doctors with limited clinical experience often rely on causal knowledge of the disease to make diagnoses, while senior doctors rely more on clinical experience, including the recollection of specific clinical cases [6]. Thus, the investigation of symptom combinations in patients with specific diseases as symptom phenotypes is of utmost importance. However, most clinical guidelines tend to describe more common symptoms in disease populations rather than at the individual level [7].

With the progress of technology, high-throughput sequencing and other techniques have been widely used in clinical research, but the potential molecular mechanism of symptom phenotypes has not been widely investigated; in particular, the use of nonspecific symptoms, such as fatigue, dizziness, headache and other symptoms, is still a challenge in the diagnosis and treatment of patients with CAD. Nonspecific symptoms are also part of the disease, but they are not enough to explain the pathological theory of persistent discomfort. One of the reasons why it is difficult to carry out this test is that the symptoms of these patients are clinical, and it is difficult to combine clinical information with experimental data effectively. Most of the existing common methods are to establish a multiplex network of clinical information and experimental information. Random walk with restart (RWR) and other classic algorithms are used to measure the importance of nodes [8, 9].

In modern times, the storage and transmission of clinical information and molecular biological information have become increasingly convenient. This has facilitated the integration and advancement of medical knowledge at various levels. As a visual representation of information structure, knowledge graphs are increasingly being utilized in the medical field. A medical knowledge graph can encompass a vast array of disease symptom characteristics and molecular biological characteristics. It offers broader coverage of entities and a wider range of semantic relations. Consequently, it serves as a valuable foundation for training machine learning models [10, 11]. Moreover, with the wide application of machine learning algorithms, in-depth mining technology for network graph information is constantly being developed [12–14].

The concept of knowledge graphs was introduced by Google in 2012 with the initial purpose of optimizing search engine results and enhancing user search quality and experience. Essentially, a knowledge graph is a technical approach that utilizes graph models to describe knowledge and model the relationships among entities in

the world [15]. It is a symbolic representation of entities and their relationships in the objective world, aiming to achieve structured semantic knowledge storage.

The knowledge generated from medical activities is vast, encompassing not only the explicit knowledge found in patients' clinical diagnosis and treatment records but also the implicit knowledge governed by the micro mechanisms of the human body. Patient data constitute an essential part of the information required for clinical decision-making, along with the understanding of etiology, pathological processes, and the effectiveness of drugs or other therapeutic measures [16]. With the exponential growth of biomedical knowledge [17], there is a need to integrate medical clinical knowledge into microbiological systems to facilitate multidimensional representations of medical knowledge. The process of knowledge discovery and translation into practice involves extracting and isolating knowledge units from information sources and establishing appropriate representation models to achieve knowledge correlation.

Therefore, as shown in Fig. 1, based on real-world electronic medical records, this study obtained information on symptoms and coronary artery plaque combined with relevant protein–protein interaction information. Machine models and algorithms such as convolutional neural networks (CNNs) and K-nearest neighbors (KNNs) have been used to analyze clinical-molecular biomedical knowledge graphs and explain the clinical characteristics and internal mechanisms of different plaques from many perspectives.

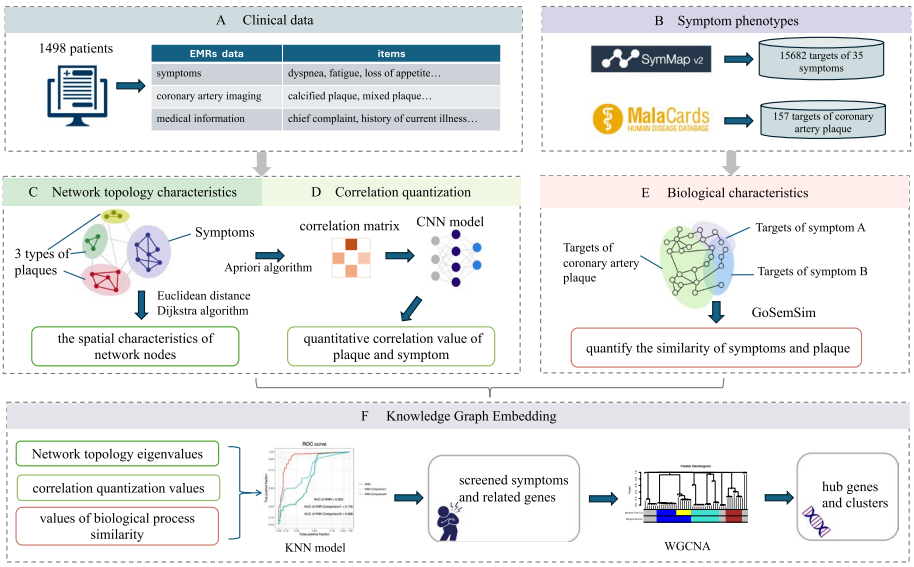

Fig. 1 Association mining between plaque characteristics and symptom phenotypes: **A** Electronic medical record data from 1498 clinical patients were utilized to extract symptoms, coronary artery imaging, and medical information. **B** Symptom phenotypes and relevant targets of coronary artery plaque were collected from the SymMap and MaleCards databases. **C** A network was established with symptoms and three types of plaques as nodes, followed by network topology analysis. **D** A correlation matrix between symptoms and plaques was constructed, and a CNN model was trained to obtain correlation values. **E** Biological process similarity between symptoms and plaques was computed in PPIN. **F** The KNN model was used to screen symptoms and related genes, and hub genes were obtained using WGCNA

Materials and methods

Data preparation

Clinical data

Clinical medical records were collected from 1498 patients who underwent coronary computed tomography angiography at the Affiliated Hospital of Shandong University of Traditional Chinese Medicine from August 2014 to June 2019. Demographic information, current symptoms, coronary artery imaging, and blood test results were collected from electronic medical records. Symptoms were standardized using Medical Subject Headings (MeSH). The characteristics of coronary artery plaques, including calcified plaque, noncalcified plaque, mixed plaque, and no plaque, were recorded. This study excluded plaques treated with coronary artery stents. Quality control was performed by a chief cardiovascular physician. In quality control, the physician is responsible for evaluating records, and the criteria for including patients are as follows:

The inclusion criteria were as follows: 1) Patients who had undergone chest spiral CT scans to obtain coronary artery images. 2) The images included the coronary arteries, left main trunk, left anterior descending branch, left circumflex branch, and right main trunk.

The exclusion criteria were as follows: 1) had a history of coronary artery bypass grafting. 2) Presence of myocardial bridging. 3) Poor image quality, including respiratory misalignment artifacts, discontinuities, and slice thicknesses greater than 1 mm.

Data on symptom phenotypes

The SymMap database contains various types of phenotypes related to diseases, including 1148 symptom terms and associated targets. It is currently being utilized in the research and development of new drugs, particularly natural drugs, for various chronic diseases, such as coronary artery atherosclerosis, Alzheimer's disease, and chronic atrophic gastritis [18–21]. The MalaCards human disease database integrates annotated disease information from 68 data sources, including symptoms, therapeutic drugs, and genes associated with diseases. It has been widely applied in genomic data annotation and drug repurposing model construction [22–26].

We sorted 35 symptoms from the electronic medical records data and collected the related targets from the SymMap (version 2.0) database [18] to establish the symptom phenotypes database. In addition, the related targets of coronary artery plaque were collected from the MalaCards database [22].

Protein–protein interactions

The complexity of the physiological and pathological states of the human body originates from the functional and regulatory interactions between proteins, new protein interactions are constantly being discovered, and information is still dispersed in different database resources and experimental papers. The STRING [27] database systematically collects and integrates protein-protein interactions, including physical interactions and functional associations, all of which undergo strict scoring selection and can be used for the analysis of disease pathogenesis and the efficacy analysis of targeted drugs [28–31].

To obtain information about the interaction between symptom phenotypes and coronary plaque target proteins, we used *Homo sapiens* protein network data in the STRING (version 11.5) database to establish a protein–protein interaction network (PPIN). The PPIN contains 12371 nodes and 2283976 edges with a combined score > = 400.

Set up network graphs

By collating symptom information and coronary artery plaque information from 1498 patients' electronic medical records, a clinical feature network was established. A network was constructed with 35 symptoms and 3 plaque properties as nodes. If two nodes appear in the same patient, they are connected as edges, with the number of patients in which this connection appears serving as the weight of the edge. This network comprises 608 edges.

Knowledge fusion

Network topology characteristics

To comprehensively analyze the correlation information between the molecular network and clinical data, after merging the clinical feature network with the PPIN, we used the Dijkstra algorithm [32] to calculate the shortest distance between nodes. Combined with the Euclidean distance, we quantify the spatial characteristics of network nodes from point-to-point and point-network perspectives.

Correlation quantization

First, in the clinical-PPIN composite network, the Apriori algorithm [33] is used to calculate the lift value of coronary artery plaque and symptom phenotypes, which is used to quantify the correlation value between each item as the value of the plaque-symptom matrix. Each patient has a corresponding matrix, which contains all symptoms and plaque types in the dataset, while the data in this matrix include only the symptoms present in that patient. Because symptoms tend to aggregate specifically with different plaques, the local receptive fields of CNN models can sensitively capture local correlations in the matrix. The translation invariance of CNNs can also identify common features among matrices of different patients. The hierarchical structure filters out edge symptom phenotypes at lower levels and completes the identification of plaque feature symptom phenotypes at higher levels. Therefore, we used the actual symptom phenotypes and plaque properties of the patients in the medical records as the standard and used the CNN model to train the correlation matrix to establish the model. Then, the quantitative correlation value is determined by using the training model.

Biological characteristics

There are complex biological interactions in the molecular network. Nodes in PPINs often have different weights in biological processes. Gene Ontology (GO) biological process semantic similarity (GoSemSim) [34, 35] can return an association value between two genes after inputting them. Therefore, we used GoSemSim to quantify the similarity of different protein sets involved in biological processes and calculated the GoSemSim values of symptoms and plaques.

Knowledge graph embedding

The above analysis revealed clinical and molecular features across multiple dimensions. The basic assumption of the KNN model is that similar samples are close to each other in the feature space, effectively adapting to data distributions without requiring complex data transformations, thus capturing correlations between features well. Through appropriate feature selection and parameter tuning, KNN plays a role in handling multidimensional data, achieving classification filtering of data points. Additionally, predictions based on local neighborhoods can effectively capture the local structure of plaque and symptom data. Therefore, we use the KNN-A model for model training. The value of the RWR algorithm is taken as the judgment. We set the network topology feature value, correlation quantization value and biometric value as input information and trained with 10 cross-verifications [36] with $K = 6$.

To verify the necessity of model input information, two comparative models were constructed in this study. KNN-B takes the network topology eigenvalues and correlation quantization values as input information, and KNN-C takes biometric values of biological process similarity as input information. Both models used the KNN with the same set.

In addition, to further verify the wide applicability of the KNN model under this setting, first, we utilize gradient boosting decision tree (GBDT) and Bayesian network (BN) models to handle the association data between symptoms and plaques and train them with 10-fold cross-validation to compare their effectiveness with that of the KNN model. Additionally, we collected the electronic medical records of 2055 patients with hypertensive nephropathy using the same data processing method with the KNN model to analyze the relationships between hypertensive nephropathy and different symptoms and medications.

Correlations between symptom phenotypes and clinical features

Using the KNN model, we screened the symptom phenotypes and related genes associated with three types of plaques. To further analyze the inherent associations between chronic diseases and blood parameters, we utilized weighted correlation network analysis (WGCNA) [37] to identify highly collaborative hub genes associated with phenotypes. WGCNA first calculates the weighted correlation coefficients between any two genes, that is, the N power of the gene correlation coefficient, ensuring that the connections between genes in the network follow scale-free properties. Then, gene selection is achieved through a threshold to obtain hub genes. Next, a hierarchical clustering tree was constructed based on the correlation coefficients between genes to identify modules where genes with similar patterns were grouped into different branches. Furthermore, the degree of association between genes within modules and phenotypes was measured using Pearson correlation coefficients. WGCNA was implemented using the *WGCNA* R package (version 1.72-5).

Pathway enrichment analysis

To determine the biological processes involved in each gene set, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis and GO term enrichment analysis

were used. A hypergeometric distribution was used, and a P value < 0.05 was considered to indicate a significant difference.

Validation of the hub genes

The GSE28829 [38], GSE97210 [39], GSE104140 and GSE109048 [40] datasets were used to validate the expression of the hub genes. These four datasets contain RNA-seq data from plaque patients. The samples for these four studies were all derived from human subjects, including patients with early intimal thickening of atherosclerotic plaques and those with late-stage fibrous cap formation, as well as from healthy individuals, totaling 124 samples. The datasets underwent rigorous quality control and processing. Within each dataset, the expression data of the samples were balanced. Differential expression analysis was performed using the GEO2R tool to obtain differentially expressed genes in the four sample groups compared to their respective control groups ($P < 0.05$). All analyses were performed using the default settings of GEO2R. Gene expression was measured using the logFC value, and the effectiveness of the hub genes divided by WGCNA was evaluated using receiver operating characteristic (ROC) curves.

Evaluation of immune cell infiltration

Pathway enrichment analysis revealed that the generation and growth of plaques are closely related to inflammation. To further confirm these findings and explore the immune cells involved in this pathological process, we used the GSE120521 [41] dataset and conducted analyses using the *xCell* (version 1.1.0) R package [42] and Cibersortx algorithms [43]. GSE120521 data were derived from RNA-seq of plaque tissue from patients. The permutation setting for the CIBERSORTx algorithm was set to 100. xCell analysis was performed using the preset function xCellAnalysis to calculate the enrichment scores of genes for each cell type, which were then converted into linear proportions while reducing dependencies between highly correlated cell types. These analyses allowed us to investigate the characteristics of the gene expression profiles of different immune cells. Additionally, we evaluated the correlation of genes related to immune cells in the internal environment using the immune score, stroma score, and microenvironment score.

Results

Symptom phenotypes and plaque association network graph

To obtain clinically relevant symptom information, we manually converted the symptoms recorded in patients' electronic medical records into standard MeSH descriptions, resulting in a total of 35 symptom types. Among these, chest pain, dyspnea, and fatigue are common cardiovascular symptoms (Fig. 2A). We collected data from 424 patients with no coronary artery plaque and 1074 patients with coronary artery plaque, including patients with calcified plaque, noncalcified plaque, and mixed plaque (Fig. 2B). Using these data, we established a network linking symptoms and coronary artery plaques. We also collected information on symptom phenotypes and coronary plaque-related genes from the SymMap and Malacards databases and used these data to establish a protein–protein interaction network (PPIN) using the STRING database.

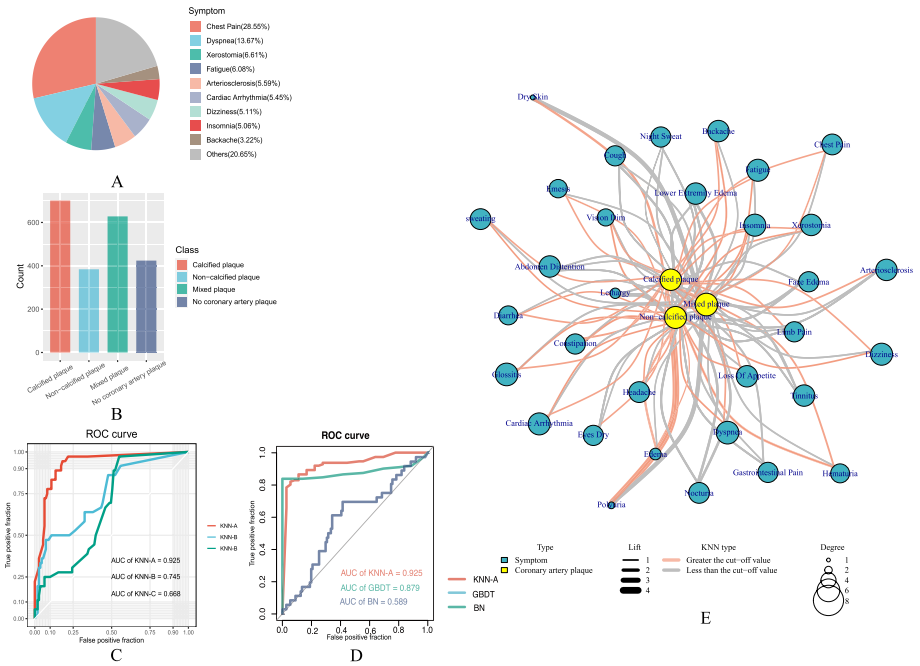

Fig. 2 The association network between symptoms and plaques: (**A**) the symptom distribution of patients, (**B**) the number of different kinds of plaques in patients, (**C**) the ROC curve trained by the KNN model, (**D**) the ROC curve trained by the KNN, GBDT and BN models, and (**E**) the network diagram of the relationship between symptoms and plaques. The color of the points indicates the symptoms or plaques, the thickness of the edges represents the lift value in the Aprori algorithm, the color of the edges represents the results of the KNN model, and the size of the points indicates the degree of nodes

Model analysis of network

To analyze the information from the PPIN and the clinical characteristics of patients, we utilized the KNN model to combine and correct the biological and clinical features obtained from algorithms such as CNN, Dijkstra, and GoSemSim. The resulting AUC value of KNN-A was 92.5%, which was greater than the AUC values obtained from KNN-B and KNN-C (74.5% and 66.8%, respectively), as shown in Fig. 2C, indicating that the model was effective. Additionally, as shown in Fig. 2D, the efficacy of the KNN-A model surpassed that of GBDT (87.9%) and BN (58.9%). We further validated the model by analyzing data from patients with hypertensive nephropathy, resulting in an AUC of 91.3%, which was consistent with related studies [44–47]. This demonstrates the wide applicability of the training data and the KNN model used in this study.

The KNN model was used to comprehensively evaluate the correlation between symptoms and plaques, resulting in the identification of 3 plaque properties, 23 kinds of symptoms, 41 association rules, and 61 hub genes. Fig. 2E shows that common symptoms such as chest pain, dizziness, and backache are associated with all 3 plaque properties. Calcified plaque was found to be associated with most symptoms (20 in total), including common symptoms of long-term chronic diseases such as dry eyes, limb pain, and arteriosclerosis. Symptoms related to fluid metabolism, such as edema and hematuria, were found to have higher lift values with noncalcified plaque, indicating that these symptoms are not commonly associated with other plaque types.

Mining the diversity of clinical features

In addition to analyzing the relationship between symptoms and plaques, we collected information on patients' comorbidities and blood laboratory test results to gain a comprehensive understanding of the clinical characteristics of coronary artery plaques. Using the 61 hub genes identified by the KNN model, we evaluated their correlation with the actual symptoms and plaque properties of patients and conducted WGCNA to explore the relationships between these genes and the clinical conditions of the patients.

Based on the coefficients of the KNN model for each item, an association matrix between patients and proteins was established. After clustering and screening patients, the soft threshold of the network recognition module was determined to be 9 through network topology analysis (Fig. 3A, B). The genes were divided into four modules using hierarchical clustering and coexpression similarity (Fig. 3C, D). By calculating the correlation coefficient between gene modules and clinical characteristics, the genes screened by KNN were associated with the related indexes of lipid metabolism in patients. The genes in the blue group were concentrated in the serum lipid content, the genes in the turquoise group were concentrated in the serum apolipoprotein content, and the genes in the gray group were also associated with complicated cerebrovascular diseases (Fig. 3E).

Molecular network mechanism of clinical symptoms

Functional enrichment analysis was performed for each group of genes categorized by WGCNA, and a functional annotation table was obtained. According to the Gene

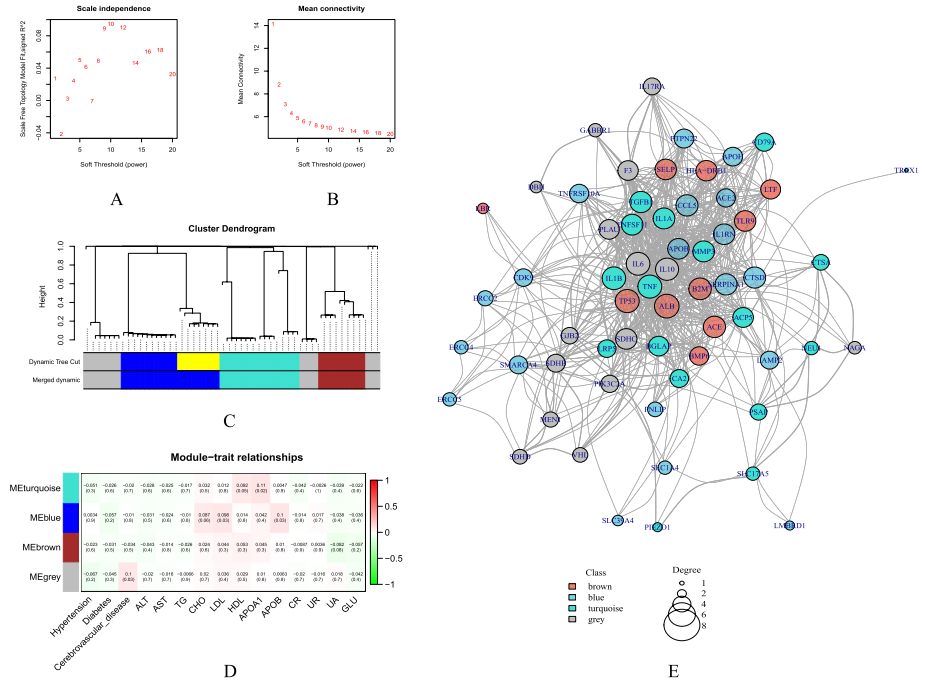

Fig. 3 WGCNA results: (**A**) and (**B**) determine the soft threshold of the network recognition module; (**C**) the module division of the gene; (**D**) the relationship between the gene module and the clinical characteristics; (**E**) the network map of the gene module division. The color of the dot indicates the gene grouping, and the size of the point indicates the degree of the node

Ontology (GO) biological process category, each gene group was primarily involved in the response to lipopolysaccharide, regulation of cell–cell adhesion, cytokine activity, and cytokine receptor binding (Fig. 4A). These genes are closely related to the inflammatory response and energy metabolism. According to the KEGG analysis, 20 signaling pathways, including cytokine–cytokine receptor interaction, lipid and atherosclerosis, the Toll-like receptor signaling pathway, and the TNF signaling pathway, which are associated with multiple gene groups and have the most extensive symptoms, were significantly enriched (Fig. 4B). The most significant KEGG pathway was cytokine–cytokine receptor interaction (Fig. 4C).

Expression characteristics of the hub genes

According to the WGCNA of the clinical phenotype (Fig. 3E), the hub genes were divided into 4 subgroups: blue, brown, gray, and turquoise. As shown in Fig. 5 A-D, all four subgroups obtained higher AUC values in the RNA-seq dataset and showed abnormal expression compared to the control group (Fig. 5E).

Immune cell infiltration

According to the results of the CIBERSORTx algorithm (uploaded to the supporting files), CD8 T cells, follicular helper T cells, M0 macrophages, M1 macrophages, and M2 macrophages were strongly correlated with plaque formation. There were significant differences in the immune score and microenvironment score compared with those of the control group, but there was no difference in the stroma score (Fig. 6A-C), suggesting that the percentage of immune cells in the plaque environment was greater and positively correlated with the expression of the hub genes.

Discussion

In the early stage of plaque formation or stable plaque, the patient may have latent coronary heart disease, and myocardial ischemia may not occur until the cardiac load increases. As the plaque changes, typical symptoms of angina pectoris begin to appear, but the signs of angina pectoris are complex and have low specificity. Patients may experience anxiety, pale skin, chills or sweating, slight increases or decreases in blood pressure, an increased heart rate, and a systolic murmur in the apical area. The purpose of analyzing symptom characteristics is to improve doctors' ability to diagnose, predict, and treat diseases, deepen the understanding of disease mechanisms, and ultimately reduce the risk of adverse events. However, most symptoms are nonspecific.

Phenotyping can be performed based on clinical and molecular characteristics. Clinical phenotyping focuses on demographic information and laboratory tests, while molecular phenotyping emphasizes DNA, mRNA, proteins, and metabolites based on molecular characteristics. To comprehensively analyze this information and elucidate the relationship between plaque features and symptom phenotypes from multiple perspectives, we used the KNN model for knowledge graph embedding.

In this study, we utilized network analysis as the data source to establish a prescription feature network based on electronic medical records and a biological network based on protein–protein interactions. We quantified the network information from multiple levels, including analyzing the biological similarity of different symptoms using GoSemSim,

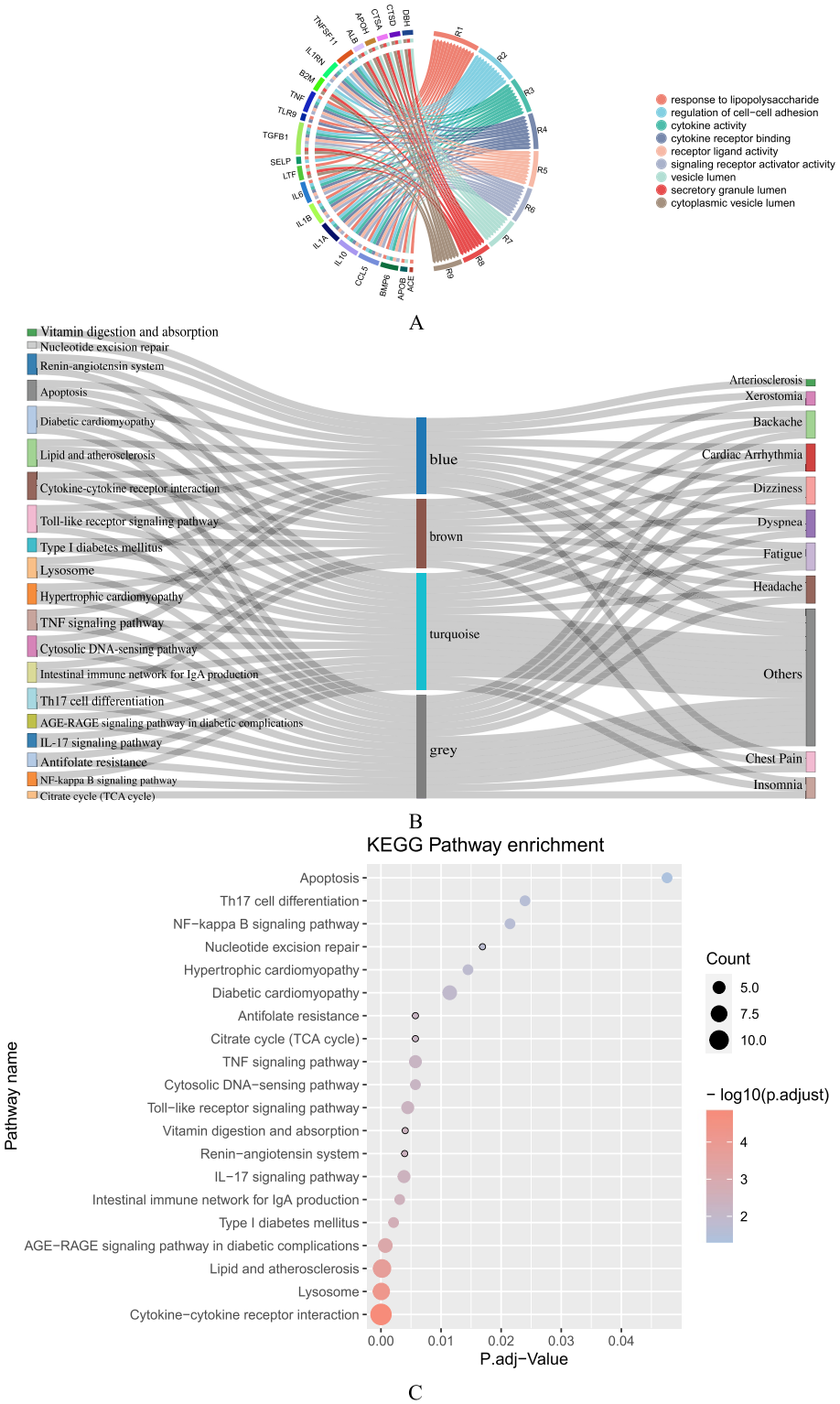

Fig. 4 Pathway enrichment results: (**A**) GO enrichment results; (**B**) KEGG pathway, gene grouping and associated distribution of symptoms, and the colors in the middle column correspond to the grouping in WGCNA; (**C**) KEGG enrichment results

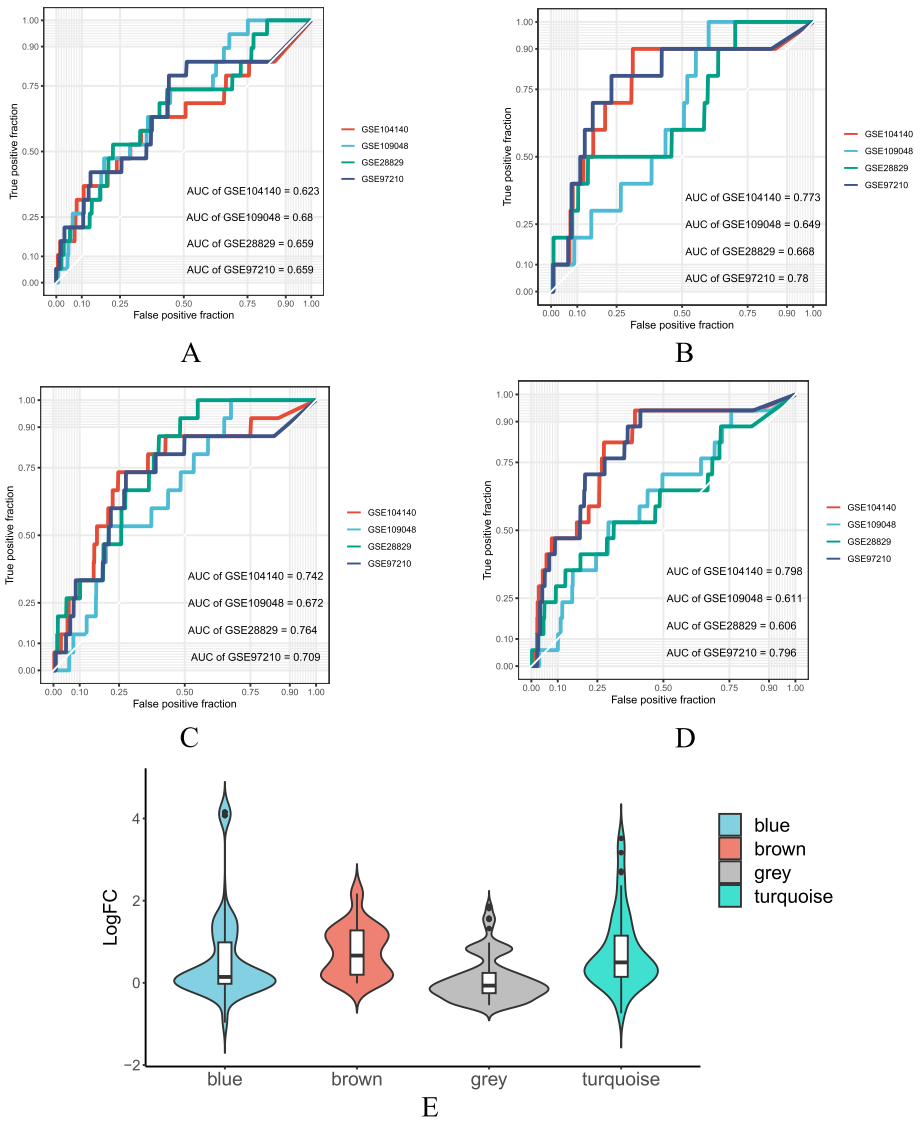

Fig. 5 The expression characteristics of Hub genes: (**A-D**): The figures show ROC curves for the expression of four hub gene subgroups (blue, brown, gray, and turquoise) in different RNA-seq datasets (GSE28829, GSE97210, GSE104140, and GSE109048). The area under the curve (AUC) was used to evaluate the specificity of the hub genes. **E**: The distribution of logFC values of the hub genes in the RNA-seq dataset

discussing the feature weights of symptom nodes and plaque nodes in the composite network using the Dijkstra algorithm and Apriori algorithm, and initially correcting the biological information and clinical information using the CNN model. Finally, we used the KNN model to train the knowledge graph model with multilevel information as input data to quantify the final association degree and determine the clinical symptoms and specific genes corresponding to the three plaques.

To study the mechanism of the symptom phenotype, we used WGCNA to construct a gene network, which formed a topological matrix and was divided into modules. Combined with pathway enrichment analysis, we found that among the three different plaque types, calcified plaques corresponded to the most extensive symptoms and were widely related to inflammation and lipid metabolism pathways. Moreover, lipid abnormalities

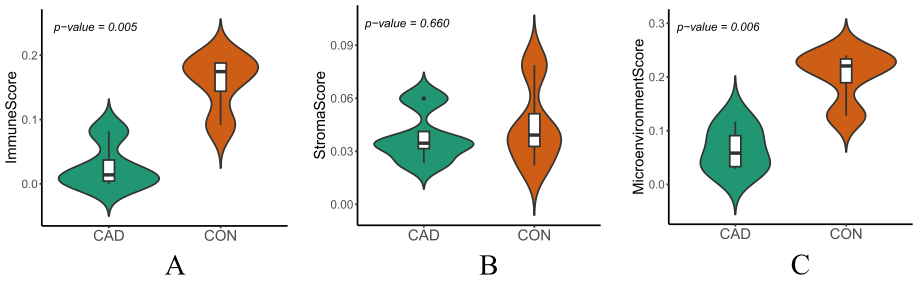

Fig. 6 Immune cell infiltration: The results of the *xCell* package analysis, which is used to measure the correlation between genes and the plaque microenvironment. **A** Immune score, which measures overall immune activity in the plaque microenvironment. **B** Stroma score, which measures the degree of stromal cell infiltration in the plaque microenvironment. **C** Microenvironment score, which measures the overall level of interaction between immune and stromal cells in the plaque microenvironment

in patients and other laboratory test results were associated with hub genes, while low-density lipoprotein levels were positively correlated with coronary artery calcification [48–50]. Immune cell infiltration analysis and existing studies have shown that coronary artery calcification mainly occurs in the intimal layer of blood vessels, and this process involves a variety of cells, such as macrophages, intimal cells, media smooth muscle cells, and fibroblasts [51]. Induced by a variety of calcium-stimulating factors, including interleukin enhancer-binding factor 3 (ILF3), smooth muscle cells inhibit alpha smooth muscle actin (α-SMA) and increase osteomodulin while migrating to plaques, which leads to phenotypic transformation of smooth muscle cells through their association with SMAD3 and TGFB1 signals and their interaction with BMP2 in vascular tissues [52–55]. On the other hand, macrophages exert their proinflammatory effects to further induce inflammatory cytokines such as IL-12 and IL-6, oxidize low-density lipoprotein in the arterial wall, and accelerate the decomposition of oxidized lipids by macrophages, resulting in increased plaque fragility [56, 57].

In the clinical context, inflammation has long been considered a primary driving factor in the formation of atherosclerosis and a key element in the development of vulnerable plaques [58]. The landmark Canakinumab Anti-Inflammatory Thrombosis Outcomes Study (CANTOS) [59] demonstrated, for the first time, a reduction in cardiovascular event recurrence with the use of Canakinumab, a monoclonal antibody targeting IL-1β. Studies on colchicine treatment for acute myocardial infarction have also underscored the role of inflammation in the pathogenesis of the disease [60]. Thus, accurate detection of vascular inflammation could help better stratify cardiovascular risk. Understanding the associations between plaque characteristics and symptom phenotypes can help clinicians understand the relationships between different symptoms and coronary artery plaques, thereby guiding clinical diagnosis and treatment decisions. This study makes full use of clinical data and molecular networks, integrating them through model construction, with the potential to provide more comprehensive information for disease diagnosis and treatment. This integrative research approach facilitates the connection between symptom manifestations, diseases, and molecules, offering new insights and methods for medical research.

The framework of this biomedical knowledge graph aims to establish a multilevel relationship between symptom phenotype and disease. By considering multiple

symptoms rather than a single symptom, the patient population can be classified more specifically based on their symptom groups. Moreover, this study explored the molecular mechanism of the symptom phenotype, and the results were preliminarily verified at the RNA, biological pathway, and cellular levels, which provides a method for further investigation of the potential network mechanism of symptoms.

At the same time, there are also many shortcomings in this study. First, for the arrangement of clinical symptoms, the standard terms of symptoms in our biomedical database are quite different from the writing habits of Chinese medical records. We often encounter that there are no corresponding standard terms for the common descriptions in Chinese medical records, which are all excluded from this study. Second, in different types of plaques, calcified plaques tend to have a longer course of disease, which increases the likelihood that patients will experience abnormal blood pressure, blood sugar, or other target organ damage, which is the focus of this study. It is also a risk factor for potential bias. Third, the relevant evidence of hub genes comes from the algorithm simulation of clinical symptom data, and only a preliminary verification is carried out. In the future, we hope to further verify the hub genes at the cellular, animal, and clinical levels. The algorithm model will be further improved, and this model will be used to screen effective drug targets and better achieve fine individualized treatment services.

Conclusion

This study presented a biomedical knowledge graph to comprehensively analyze disease characteristics and symptom phenotypes. This graph was used to investigate the characteristics and molecular mechanisms of coronary artery plaque features and symptoms, ultimately identifying the corresponding symptoms for three types of plaques. The findings of this study revealed that patients with calcified plaques exhibited more combined symptoms and complex biological processes. The underlying mechanism of the symptomatic phenotype was linked to the inflammatory response and biological processes related to lipid metabolism.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00365-1.

Supplementary Material 1.

Acknowledgments
The authors thank Shandong Engineering Laboratory of Traditional Chinese Medicine Precision Therapy for Cardiovascular Diseases for its support of this study.

Authors' contributions
JM.H wrote the manuscript. JM.H, SJ.W, Y.L, SJ.Z analysed data. Y.L, SJ.Z, XJ,W, YL,H involved in data collection. YL,L is responsible for quality control.

Funding
This work was supported by the National Natural Science Foundation of China [81974566] and the 20 Policies in Colleges and Universities Project of Jinan [2020GXRC017].

Availability of data and materials
The original electronic medical records data used in this study were uploaded to the supplementary materials. CNN model training was implemented in Python 3.7.2, and other data analysis was implemented in R 4.2.2.

Declarations

Ethics approval and consent to participate
This study was reviewed and approved by the Ethics Committee of the Affiliated Hospital of Shandong University of Traditional Chinese Medicine and was supervised by the Ethics Committee during the test. This work was conducted in accordance with the Helsinki Declaration. No personal information was requested, as all forms were anonymous, making the identification of participants impossible.
This study was retrospective; therefore, informed consent was waived. Informed consent was waived by the Ethics Committee of the Affiliated Hospital of Shandong University of Traditional Chinese Medicine.

Consent for publication
Not applicable

Competing interests
The authors declare no competing interests.

Received: 28 June 2023 Accepted: 17 May 2024
Published online: 21 May 2024

References

1. Mortensen MB, Dzaye O, Steffensen FH, Bøtker HE, Jensen JM, Rønnow Sand NP, Kragholm KH, Sørensen HT, Leipsic J, Mæng M, Blaha MJ, Nørgaard BL. Impact of Plaque Burden Versus Stenosis on Ischemic Events in Patients With Coronary Atherosclerosis. J Am Coll Cardiol. 2020;76(24):2803–13.
2. Villines TC, Rodriguez Lozano P. Transitioning From Stenosis to Plaque Burden in the Cardiac CT Era: The Changing Risk Paradigm. J Am Coll Cardiol. 2020;76(24):2814–6.
3. Ferraro R, Latina JM, Alfaddagh A, Michos ED, Blaha MJ, Jones SR, Sharma G, Trost JC, Boden WE, Weintraub WS, Lima JAC, Blumenthal RS, Fuster V, Arbab-Zadeh A. Evaluation and Management of Patients With Stable Angina: Beyond the Ischemia Paradigm: JACC State-of-the-Art Review. J Am Coll Cardiol. 2020;76(19):2252–66.
4. Sanchis-Gomar F, Perez-Quilis C, Leischik R, Lucia A. Epidemiology of coronary heart disease and acute coronary syndrome. Ann Transl Med. 2016;4(13):256.
5. Stone PH, Libby P, Boden WE. Fundamental Pathobiology of Coronary Atherosclerosis and Clinical Implications for Chronic Ischemic Heart Disease Management-The Plaque Hypothesis: A Narrative Review. JAMA Cardiol. 2023;8(2):192–201.
6. Brush JE Jr, Sherbino J, Norman GR. How Expert Clinicians Intuitively Recognize a Medical Diagnosis. Am J Med. 2017;130(6):629–34.
7. Brush JE Jr, Hajduk AM, Greene EJ, Dreyer RP, Krumholz HM, Chaudhry SI. Sex Differences in Symptom Phenotypes Among Older Patients with Acute Myocardial Infarction. Am J Med. 2022;135(3):342–9.
8. Valdeolivas A, Tichit L, Navarro C, Perrin S, Odelin G, Levy N, Cau P, Remy E, Baudot A. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics. 2019;35(3):497–505.
9. Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010;26(9):1219–24.
10. Cheng B, Zhang J, Liu H, Cai M, Wang Y. Research on Medical Knowledge Graph for Stroke. J Healthc Eng. 2021;24(2021):5531327.
11. Mohamed SK, Nounu A, Nováček V. Biological applications of knowledge graph embedding models. Brief Bioinform. 2021;22(2):1679–93.
12. Yang K, Lu K, Wu Y, Yu J, Liu B, Zhao Y, Chen J, Zhou X. A network-based machine-learning framework to identify both functional modules and disease genes. Hum Genet. 2021;140(6):897–913.
13. Shu Z, Wang J, Sun H, Xu N, Lu C, Zhang R, Li X, Liu B, Zhou X. Diversity and molecular network patterns of symptom phenotypes. NPJ Syst Biol Appl. 2021;7(1):41.
14. Yang J, Tian S, Zhao J, Zhang W. Exploring the mechanism of TCM formulae in the treatment of different types of coronary heart disease by network pharmacology and machining learning. Pharmacol Res. 2020;159:105034.
15. Fensel D, Şimşek U, Angele K, et al. Introduction: what is a knowledge graph?. Knowledge graphs: Methodology, tools and selected use cases, 2020:1–10.
16. Shortliffe EH, Shortliffe EH, Cimino JJ, et al. Biomedical informatics: computer applications in health care and biomedicine. Springer; 2014.
17. Musen MA. The protégé project: a look back and a look forward. AI matters. 2015;1(4):4–12.
18. Wu Y, Zhang F, Yang K, Fang S, Bu D, Li H, Sun L, Hu H, Gao K, Wang W, Zhou X, Zhao Y, Chen J. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Res. 2019;47(D1):D1110–7.
19. Hong M, Wu Y, Zhang H, Gu J, Chen J, Guan Y, Qin X, Li Y, Cao J. Network pharmacology and experimental analysis to reveal the mechanism of Dan-Shen-Yin against endothelial to mesenchymal transition in atherosclerosis. Front Pharmacol. 2022;24(13):946193.
20. Ye XW, Wang HL, Cheng SQ, Xia LJ, Xu XF, Li XR. Network Pharmacology-Based Strategy to Investigate the Pharmacologic Mechanisms of Coptidis Rhizoma for the Treatment of Alzheimer's Disease. Front Aging Neurosci. 2022;21(14):890046.
21. Tian G, Wu C, Li J, Liang B, Zhang F, Fan X, Li Z, Wang Y, Li Z, Liu D, Lai-Han Leung E, Chen J. Network pharmacology based investigation into the effect and mechanism of Modified Sijunzi Decoction against the subtypes of chronic atrophic gastritis. Pharmacol Res. 2019;144:158–66.

22. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M, Lancet D. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017;45(D1):D877–87.
23. Wilkerson BA, Zebroski HL, Finkbeiner CR, Chitsazan AD, Beach KE, Sen N, Zhang RC, Bermingham-McDonogh O. Novel cell types and developmental lineages revealed by single-cell RNA-seq analysis of the mouse crista ampullaris. Elife. 2021;18(10):e60108.
24. Robertson MJ, Kent K, Tharp N, Nozawa K, Dean L, Mathew M, Grimm SL, Yu Z, Légaré C, Fujihara Y, Ikawa M, Sullivan R, Coarfa C, Matzuk MM, Garcia TX. Large-scale discovery of male reproductive tract-specific genes through analysis of RNA-seq datasets. BMC Biol. 2020;18(1):103.
25. Yang M, Wu G, Zhao Q, Li Y, Wang J. Computational drug repositioning based on multisimilarities bilinear matrix factorization. Brief Bioinform. 2021;4:bbaa267.
26. Luo H, Li M, Yang M, Wu FX, Li Y, Wang J. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform. 2021;22(2):1604–19.
27. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1):D605–12.
28. Váraljai R, Zimmer L, Al-Matary Y, Kaptein P, Albrecht LJ, Shannan B, Brase JC, Gusenleitner D, Amaral T, Wyss N, Utikal J, Flatz L, Rambow F, Reinhardt HC, Dick J, Engel DR, Horn S, Ugurel S, Sondermann W, Livingstone E, Sucker A, Paschen A, Zhao F, Placke JM, Klose JM, Fendler WP, Thommen DS, Helfrich I, Schadendorf D, Roesch A. Interleukin 17 signaling supports clinical benefit of dual CTLA-4 and PD-1 checkpoint inhibition in melanoma. Nat Cancer. 2023;4(9):1292–308. https://doi.org/10.1038/s43018-023-00610-2. Epub 2023 Jul 31. Erratum in: Nat Cancer. 2023 Aug 14.
29. Liu X, Qi X, Han R, Mao T, Tian Z. Gut microbiota causally affects cholelithiasis: a two-sample Mendelian randomization study. Front Cell Infect Microbiol. 2023;9(13):1253447.
30. Wang HL, Li JN, Kan WJ, Xu GY, Luo GH, Song N, Wu WB, Feng B, Fu JF, Tu YT, Liu MM, Xu R, Zhou YB, Wei G, Li J. Chloroquine enhances the efficacy of chemotherapy drugs against acute myeloid leukemia by inactivating the autophagy pathway. Acta Pharmacol Sin. 2023;44(11):2296–306.
31. Zhang G, Ji P, Xia P, Song H, Guo Z, Hu X, Guo Y, Yuan X, Song Y, Shen R, Wang D. Identification and targeting of cancer-associated fibroblast signature genes for prognosis and therapy in Cutaneous melanoma. Comput Biol Med. 2023;167:107597.
32. Noto M, Sato H. A method for the shortest path search by extended Dijkstra algorithm. Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0. IEEE, 2000, 3: 2316-2320.
33. Srikant R, Vu Q, Agrawal R. Mining association rules with item constraints. InKdd. 1997;97:67–73.
34. Yu G. Gene Ontology Semantic Similarity Analysis Using GOSemSim. Methods Mol Biol. 2020;2117:207–15.
35. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010;26(7):976–8.
36. Wang N, Li P, Hu X, Yang K, Peng Y, Zhu Q, Zhang R, Gao Z, Xu H, Liu B, Chen J, Zhou X. Herb Target Prediction Based on Representation Learning of Symptom related Heterogeneous Network. Comput Struct Biotechnol J. 2019;8(17):282–90.
37. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;29(9):559.
38. Döring Y, Manthey HD, Drechsler M, Lievens D, Megens RT, Soehnlein O, Busch M, Manca M, Koenen RR, Pelisek J, Daemen MJ, Lutgens E, Zenke M, Binder CJ, Weber C, Zernecke A. Autoantigenic protein-DNA complexes stimulate plasmacytoid dendritic cells to promote atherosclerosis. Circulation. 2012;125(13):1673–83.
39. Bai HL, Lu ZF, Zhao JJ, Ma X, Li XH, Xu H, Wu SG, Kang CM, Lu JB, Xu YJ, Xiao L, Wu Q, Ye S, Wang Q, Zheng L, Hu YW. Microarray profiling analysis and validation of novel long noncoding RNAs and mRNAs as potential biomarkers and their functions in atherosclerosis. Physiol Genomics. 2019;51(12):644–56.
40. Gobbi G, Carubbi C, Tagliazucchi GM, Masselli E, Mirandola P, Pigazzani F, Crocamo A, Notarangelo MF, Suma S, Paraboschi E, Maglietta G, Nagalla S, Pozzi G, Galli D, Vaccarezza M, Fortina P, Addya S, Ertel A, Bray P, Duga S, Berzuini C, Vitale M, Ardissino D. Sighting acute myocardial infarction through platelet gene expression. Sci Rep. 2019;9(1):19574.
41. Mahmoud AD, Ballantyne MD, Miscianinov V, Pinel K, Hung J, Scanlon JP, Iyinikkel J, Kaczynski J, Tavares AS, Bradshaw AC, Mills NL, Newby DE, Caporali A, Gould GW, George SJ, Ulitsky I, Sluimer JC, Rodor J, Baker AH. The Human-Specific and Smooth Muscle Cell-Enriched LncRNA SMILR Promotes Proliferation by Regulating Mitotic CENPF mRNA and Drives Cell-Cycle Progression Which Can Be Targeted to Limit Vascular Remodeling. Circ Res. 2019;125(5):535–51.
42. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220.
43. Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. Methods Mol Biol. 2020;2117:135–57.
44. Yan Dong,Yue Bowen,Qian Muyan,Zhao Lili,Zhang Zihan,Qian Hui,Yan Shihai,Qian Yuliang, Fang Zhuyuan. JYYS Granule Mitigates Renal Injury in Clinic and in Spontaneously Hypertensive Rats by Inhibiting NF- κ B Signaling-Mediated Microinflammation. Evidence-based complementary and alternative medicine : eCAM,2018,2018.
45. Owoicho Orgah J, Wang M, Yang X, Wang Z, Wang D, Zhang Q, Fan G, Han J, Qin G, Gao X, Zhu Y. Danhong Injection Protects Against Hypertension-Induced Renal Injury Via Down-Regulation of Myoglobin Expression in Spontaneously Hypertensive Rats. Kidney Blood Press Res. 2018;43(1):12–24.
46. Wu L, Liu M, Fang Z. Combined Therapy of Hypertensive Nephropathy with Breviscapine Injection and Antihypertensive Drugs: A Systematic Review and a Meta-Analysis. Evid Based Complement Alternat Med. 2018;2018:2958717.
47. Li Y, Yan S, Qian L, Wu L, Zheng Y, Fang Z. Danhong Injection for the Treatment of Hypertensive Nephropathy: A Systematic Review and Meta-Analysis. Front Pharmacol. 2020;11:909.

48. Ceponiene I, Li D, El Khoudary SR, Nakanishi R, Stein JH, Wong ND, Nezarat N, Kanisawa M, Rahmani S, Osawa K, Tattersall MC, Budoff MJ. Association of Coronary Calcium, Carotid Wall Thickness, and Carotid Plaque Progression With Low-Density Lipoprotein and High-Density Lipoprotein Particle Concentration Measured by Ion Mobility (From Multiethnic Study of Atherosclerosis [MESA]). Am J Cardiol. 2021;1(142):52–8.
49. Hsu JJ, Tintut Y, Demer LL. Lipids and cardiovascular calcification: contributions to plaque vulnerability. Curr Opin Lipidol. 2021;32(5):308–14.
50. Kristanto W, van Ooijen PM, Greuter MJ, Groen JM, Vliegenthart R, Oudkerk M. Noncalcified coronary atherosclerotic plaque visualization on CT: effects of contrast-enhancement and lipid-content fractions. Int J Cardiovasc Imaging. 2013;29(5):1137–48.
51. Bardeesi ASA, Gao J, Zhang K, Yu S, Wei M, Liu P, Huang H. A novel role of cellular interactions in vascular calcification. J Transl Med. 2017;15(1):95.
52. Skenteris NT, Seime T, Witasp A, Karlöf E, Wasilewski GB, Heuschkel MA, Jaminon AMG, Oduor L, Dzhanaev R, Kronqvist M, Lengquist M, Peeters FECM, Söderberg M, Hultgren R, Roy J, Maegdefessel L, Arnardottir H, Bengtsson E, Goncalves I, Quertermous T, Goettsch C, Stenvinkel P, Schurgers LJ, Matic L. Osteomodulin attenuates smooth muscle cell osteogenic transition in vascular calcification. Clin Transl Med. 2022;12(2):e682.
53. Durham AL, Speer MY, Scatena M, Giachelli CM, Shanahan CM. Role of smooth muscle cells in vascular calcification: implications in atherosclerosis and arterial stiffness. Cardiovasc Res. 2018;114(4):590–600.
54. Furmanik M, van Gorp R, Whitehead M, Ahmad S, Bordoloi J, Kapustin A, Schurgers LJ, Shanahan CM. Endoplasmic Reticulum Stress Mediates Vascular Smooth Muscle Cell Calcification via Increased Release of Grp78 (Glucose-Regulated Protein, 78 kDa)-Loaded Extracellular Vesicles. Arterioscler Thromb Vasc Biol. 2021;41(2):898–914.
55. Chistiakov DA, Myasoedova VA, Melnichenko AA, Grechko AV, Orekhov AN. Calcifying Matrix Vesicles and Atherosclerosis. Biomed Res Int. 2017;2017:7463590.
56. Henein MY, Vancheri S, Longo G, Vancheri F. The Role of Inflammation in Cardiovascular Disease. Int J Mol Sci. 2022;23(21):12906.
57. Otsuka F, Yasuda S, Noguchi T, Ishibashi-Ueda H. Pathology of coronary atherosclerosis and thrombosis. Cardiovasc Diagn Ther. 2016;6(4):396–408.
58. Antonopoulos AS, Sanna F, Sabharwal N, Thomas S, Oikonomou EK, Herdman L, Margaritis M, Shirodaria C, Kampoli AM, Akoumianakis I, Petrou M, Sayeed R, Krasopoulos G, Psarros C, Ciccone P, Brophy CM, Digby J, Kelion A, Uberoi R, Anthony S, Alexopoulos N, Tousoulis D, Achenbach S, Neubauer S, Channon KM, Antoniades C. Detecting human coronary inflammation by imaging perivascular fat. Sci Transl Med. 2017;9(398):eaal2658.
59. Ridker PM, Everett BM, Thuren T, MacFadyen JG, Chang WH, Ballantyne C, Fonseca F, Nicolau J, Koenig W, Anker SD, Kastelein JJP, Cornel JH, Pais P, Pella D, Genest J, Cifkova R, Lorenzatti A, Forster T, Kobalava Z, Vida-Simiti L, Flather M, Shimokawa H, Ogawa H, Dellborg M, Rossi PRF, Troquay RPT, Libby P, Glynn RJ. CANTOS Trial Group. Antiinflammatory Therapy with Canakinumab for Atherosclerotic Disease. N Engl J Med. 2017;377(12):1119–31.
60. Tardif JC, Kouz S, Waters DD, Bertrand OF, Diaz R, Maggioni AP, Pinto FJ, Ibrahim R, Gamra H, Kiwan GS, Berry C, López-Sendón J, Ostadal P, Koenig W, Angoulvant D, Grégoire JC, Lavoie MA, Dubé MP, Rhainds D, Provencher M, Blondeau L, Orfanos A, L'Allier PL, Guertin MC, Roubille F. Efficacy and Safety of Low-Dose Colchicine after Myocardial Infarction. N Engl J Med. 2019;381(26):2497–505.

Publisher's Note