Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

Integrative analysis of bone-formation associated genes and immune cell infiltration in osteoporosis, and the prediction of active ingredients in targeted traditional Chinese medicine

WANG Kai[a†], DONG Ping[b†], GUO Hongzhang[a*]

a. Department of Orthopedics, Gansu Provincial Hospital, Lanzhou, Gansu 730000, China

b. College of Traditional Chinese Medicine, Inner Mongolia Medical University, Hohhot, Inner Mongolia 010110, China

ARTICLE INFO

Article history
Received 08 April 2024
Accepted 05 June 2024
Available online 25 June 2024

Keywords
Osteoporosis
Bone formation
Differentially expressed genes
Biological information
Traditional Chinese medicine (TCM)
Active ingredients
Molecular mechanism

ABSTRACT

Objective To explore the differential expression and mechanisms of bone formation-related genes in osteoporosis (OP) leveraging bioinformatics and machine learning methodologies, and to predict the active ingredients of targeted traditional Chinese medicine (TCM) herbs.

Methods The Gene Expression Omnibus (GEO) and GeneCards databases were employed to conduct a comprehensive screening of genes and disease-associated loci pertinent to the pathogenesis of OP. The R package was utilized as the analytical tool for the identification of differentially expressed genes. Least absolute shrinkage and selection operator (LASSO) logistic regression analysis and support vector machine-recursive feature elimination (SVM-RFE) algorithm were employed in defining the genetic signature specific to OP. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses for the selected pivotal genes were conducted. The cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm was leveraged to examine the infiltration patterns of immune cells, with Spearman's rank correlation analysis utilized to assess the relationship between the expression levels of the genes and the presence of immune cells. Coremine Medical Database was used to screen out potential TCM herbs for the treatment of OP. Comparative Toxicogenomics Database (CTD) was employed for forecasting the TCM active ingredients targeting the key genes. AutoDock Vina 1.2.2 and GROMACS 2020 softwares were employed to conclude analysis results, facilitating the exploration of binding affinity and conformational dynamics between the TCM active ingredients and their biological targets.

Results Ten genes were identified by intersecting the results from the GEO and GeneCards databases. Through the application of LASSO regression and SVM-RFE algorithm, four pivotal genes were selected: coat protein (*CP*), kallikrein 3 (*KLK3*), polymerase γ (*POLG*), and transient receptor potential vanilloid 4 (*TRPV4*). GO and KEGG pathway enrichment analyses revealed that these trait genes were predominantly engaged in the regulation of defense response activation, maintenance of cellular metal ion balance, and the production of chemokine ligand 5. These genes were notably associated with signaling pathways such as ferroptosis, porphyrin metabolism, and base excision repair. Immune infiltration analysis showed that key genes were highly correlated with immune cells. Macrophage M0, M1, M2,

†The authors contributed equally.

*Corresponding author: GUO Hongzhang, E-mail: hongzhangguo2022@126.com.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2024.09.007

Citation: WANG K, DONG P, GUO HZ. Integrative analysis of bone-formation associated genes and immune cell infiltration in osteoporosis, and the prediction of active ingredients in targeted traditional Chinese medicine. Digital Chinese Medicine, 2024, 7(2): 160-170.

and resting dendritic cell were significantly different between groups, and there were significant differences between different groups ($P < 0.05$). The interaction counts of resveratrol, curcumin, and quercetin with *KLK3* were 7, 3, and 2, respectively. It shows that the interactions of resveratrol, curcumin, and quercetin with *KLK3* were substantial. Molecular docking and molecular dynamics simulations further confirmed the robust binding affinity of these bioactive compounds to the target genes.

Conclusion Pivotal genes including *CP*, *KLK3*, *POLG*, and *TRPV4*, exhibited commendable significant prognostic value, and played a crucial role in the diagnostic assessment of OP. Resveratrol, curcumin, and quercetin, natural compounds found in TCM, showed promise in their potential to effectively modulate the bone-forming gene *KLK3*. This study provides a scientific basis for the interpretation of the pathogenesis of OP and the development of clinical drugs.

1 Introduction

Osteoporosis (OP) is a systemic metabolic bone disorder characterized by a decrease in bone mass and the degradation of bone microarchitecture [1]. The etiology of OP is multifaceted, with disorders of bone metabolism ensuing from a multitude of factors, where the predominance of bone resorption over bone formation constitutes the pathological underpinnings of the disease [2]. Clinical management of OP revolves around the suppression of bone resorption, the augmentation of bone formation, and the mitigation of osteoporotic fracture incidence [3]. The immune system maintains a multifaceted and complex association with the physiology and pathology of bone. Empirical evidence has substantiated the pivotal role of the immune system in the orchestration of bone metabolism, with immune cells exerting their influence on the process of bone formation through direct cell-to-cell interactions or via paracrine signaling mechanisms [4].

During the pathogenesis of OP, the immune system is mildly activated, with tumor necrosis factor (TNF)-α expressed by specific subtypes of T lymphocytes. The receptor activator of nuclear factor-κB (RANK) ligand (RANKL) produced by B cells accelerates the apoptosis of osteoblasts and indirectly stimulates the activity of osteoclasts, thereby causing bone mass loss [5]. The release of interleukin (IL)-17 by Th17 cells promotes the differentiation of mesenchymal stem cells into osteogenic lineages, and also indirectly increases the differentiation of osteoclasts. When estrogen levels are low, B cells primarily function to inhibit bone resorption and serve as the main source of osteoprotegerin (OPG). OPG competitively binds to RANKL, preventing RANK from engaging with it, thus inhibiting the formation of osteoclasts. Additionally, OPG promotes osteoclast apoptosis by stimulating the secretion of transforming growth factor (TGT)-β [6].

Traditional Chinese medicine (TCM) and its bioactive constituents have proven to be effective and safe in the prophylaxis and management of OP, showcasing considerable therapeutic potential and promising applications [7]. The prognostication and selection of TCM and its bioactive ingredients targeting key genes are anticipated to furnish novel insights and avenues for both fundamental research and clinical pharmacotherapy in the context of OP. Therefore, in this study, the core genes related to bone formation were deeply mined using bioinformatics technology and machine learning, and the infiltration of immune cells was analyzed to further reveal the pathological mechanism of OP. The prediction and screening of TCM active ingredients were carried out for the key genes to provide new ideas and directions for the basic research and clinical medication of OP.

2 Data and methods

2.1 Data collection

Data were searched with “osteoporosis” as the keyword and downloaded from the Gene Expression Omnibus (GEO) Datasets (http://www.ncbi.nlm.nih.gov/gds/) public database on October 21, 2023, specifically the datasets GSE7429 and GSE56815. The corresponding data file was noted as GPL96. The expression profile data from GSE7429 encompassed 20 sample groups, consisting of 10 with normal bone mineral density (BMD) and 10 with low BMD. Additionally, the expression profile analysis incorporated a comprehensive dataset of 80 sample groups from GSE56815, evenly distributed across 40 cases with normal BMD and 40 with low BMD. The T value of BMD was compared between the BMD measured by the BMD measuring instrument and the BMD of normal young people. The T value $\leqslant -1$ indicates a decrease in BMD, and T value > -1 indicates a normal BMD. The normal BMD samples were categorized as control group and the low BMD samples as disease group. The ComBat algorithm was employed to normalize the microarrays. The “osseous formation” “bone formation” and “osteoporosis” were used to search genes pertinent to OP and bone formation from GeneCards (https://www.genecards.org/).

2.2 Analysis of differentially expressed genes

The R package was invoked to elucidate the molecular mechanisms underlying the differences within the GSE7429 and GSE56815 datasets, respectively [8]. This involved identifying the genes that exhibited differential expression between the normal and low BMD samples. The screening criteria for differentially expressed genes were established with a *P* value less than 0.05. Subsequently, volcano plots and heatmaps were generated to visualize the differentially expressed genes.

2.3 Least absolute shrinkage and selection operator (LASSO) logistic regression and support vector machine-recursive feature elimination (SVM-RFE) algorithms for screening the features of the genes

LASSO logistic regression and SVM-RFE algorithms, though based on different principles, can efficiently handle large datasets and identify key features that significantly influence predictive outcomes. This capability enables the construction of robust and reliable prediction models. The two methodologies were employed in this study to identify the pivotal genes. LASSO logistic regression applies L1 regularization, effectively filtering out features with significant influences on the target variable. This approach simplifies the model and enhances its interpretability [9]. The SVM-RFE model was constructed utilizing the "e1071" software package V1.7-10, which served to further delineate the featured genes associated with OP. LASSO regression was performed with the parameters: family = binomial, alpha = 1, and nfold = 5. The SVM-RFE model was built with the parameters: nfold = 5, k = 10, and halve.above = 100. The top 10 variables were selected for the construction of the SVM model.

2.4 Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses

The R package cluster Profiler V3.6.0 was utilized for functional annotation and graphical representation, conducting enrichment analyses of the GO and KEGG databases for the identified key genes [10]. GO enrichment analysis is mainly based on biological processes (BP) and molecular functions (MF). $P < 0.05$ was considered statistically significant. The results of enrichment analysis were presented in a bar chart.

2.5 Analysis of immune cell infiltration

The top 2 000 genes screened from the GeneCards database were intersected with GSE7429 and GSE56815 differential genes, and 10 intersecting genes were obtained. Using LASSO regression and SVM-RFE, these 10 intersection genes were used to screen out the characteristic genes coat protein (*CP*), kallikrein 3 (*KLK3*), polymerase γ (*POLG*), and transient receptor potential vanilloid 4 (*TRPV4*) in OP. A comprehensive analysis was conducted to explore the interplay between the pivotal genes and immune cell infiltration within the OP dataset. The cell-type identification by estimating relative subsets of RNA transcripts (CIBERSORT) algorithm was selected to analyze the data of patients, enabling the estimation of 22 distinct types of immune cell infiltration. Additionally, Spearman's rank correlation analysis was conducted to assess the relationship between the expression levels of the genes and the composition of immune cells , and $P < 0.05$ was considered statistically significant [11].

2.6 Receiver operating characteristic (ROC) curve for assessing specificity and sensitivity of the key genes in disease diagnosis

To assess the diagnostic potential of the key genes (*CP*, *KLK3*, *POLG*, and *TRPV4*), the ROC curve was employed to evaluate the predictive capabilities of these genes in the OP dataset. The GSE7429 dataset was utilized to determine the expression levels of the four key genes, and their variations were observed.

2.7 Targeted TCM active ingredients screening

The pivotal genes were uploaded to the Coremine Medical database to screen for potentially effective TCM ingredients for the treatment of OP. Comparative Toxicogenomics Database (CTD) offers an extensive array of data on interactions among chemical substances, genes, functional phenotypes, and diseases, facilitating the exploration of disease-associated environmental factors and the potential pharmacological mechanisms of action. CTD encompasses over 46.64 million interactions, including more than 2.3 million chemicals, 46 689 genes, 4 340 phenotypes, and 7 212 disease-related entries [12]. Therefore, we utilized CTD to validate the interactions between these key genes and the pathophysiology of OP. Interactions $\geqslant 2$ was used as the screening criteria to ensure that the screened active ingredients of TCM have shown targeting effects on at least two different genes in the database.

2.8 Molecular docking

Only the active ingredients of resveratrol, curcumin, and quercetin acting on *KLK3* were screened from CTD. To assess the binding affinity of resveratrol, curcumin, and quercetin with *KLK3*, the binding affinity of resveratrol, curcumin, and quercetin with *KLK3*, a semi-flexible molecular docking procedure was conducted utilizing AutoDock Vina software V1.1.2. A docking score below

– 7.0 kcal/mol denotes a strong binding affinity between the target and the compound; a score ranging from – 5.0 to – 7.0 kcal/mol suggests a favorable binding affinity; while a score between – 5.0 and – 4.25 kcal/mol indicates a moderate binding capacity with the macromolecular target [13]. From the PubChem compound database (https://pubchem.ncbi.nlm.nih.gov/), the molecular structures of the active ingredients were retrieved, and the protein three-dimensional (3D) structures were downloaded from the Protein Data Bank (PDB) (http://www.rcsb.org/). The protein and ligand files were converted into PDBQT format, deleting all water molecules, introducing polar hydrogen atoms, and adjusting the grid frame to envelop each protein domain and allow for unencumbered molecular movement. Molecular docking was performed utilizing AutoDock Vina V1.2.2 software, repeated three times to ensure the reliability of the results, and the subsequent visual processing was conducted with the use of PyMOL V2.2.0 software.

2.9 Molecular dynamics simulations

To further investigate the molecular mechanism of curcumin, quercetin, and resveratrol binding to *KLK3*, we performed molecular dynamics simulations using GROMACS software 2020. *KLK3* was parameterized using the AMBER99SB-ILDN force field parameters, the active ingredients were parameterized using the gaff generic force field parameters, and the Sobtop program dev 3.1 was used to construct a small molecule topology and restrained electrostatic potential (RESP) for charge fitting. The transferable intermolecular potential 3 point (TIP3P) dominant water model was selected with atoms at a minimum distance of 1.0 nm from the edge of the water box, and sodium ions were leveraged to neutralize the system charge based on docking results. The workflow of the molecular dynamics simulations included energy minimization, heating, equilibration, and production kinetics simulations. Protein and small molecule heavy atoms were subjected to 10 000 steps of energy minimization, which included 5 000 steps of the steepest descent method and 5 000 steps of the conjugate gradient method, focusing on water molecules. Following this, all atoms in the whole system can move freely, which makes the simulation closer to the molecular behavior in the natural environment, and thus more realistically simulates the dynamic behavior and interaction of protein-ligand complexes after constraints were lifted. The entire system underwent 10 000 steps of energy minimization. The system was gradually heated to 300 K over 50 ps. Post-heating, it was equilibrated for an additional 50 ps under the isothermal-isobaric ensemble, number of particles, pressure, and temperature (NPT) ensemble. Subsequently, a 100 ns molecular dynamics simulation was conducted under the NPT conditions, using the trjconv module for correlation analysis. The binding free energy of ligands and proteins was calculated using the gmx_MMPBSA method in GROMACS 2020. Root mean square deviation (RMSD) is the sum of all atomic position deviations between calculated conformation at a certain time and the initial conformation. The wider the distribution range of RMSD value, the richer change of molecular conformation. Root mean square fluctuation (RMSF) is used to characterize the conformational changes of each amino acid on the protein chain during simulation. The larger the RMSF value, the greater conformational fluctuation of the amino acid, and the more flexible the movement of the amino acid residues.

3 Results

3.1 Analysis of differentially expressed genes

The GSE7429 and GSE56815 datasets comprised expression profiles from 100 patient groups, evenly divided between 50 with normal BMD and 50 with low BMD. In the GSE7429 dataset, a total of 1 033 differentially expressed genes was identified, with 497 genes up-regulated and 536 genes down-regulated (Figure 1A and 1C). The GSE56815 dataset revealed a total of 2 498 differentially expressed genes, comprising 1 391 genes up-regulated and 1 107 genes down-regulated (Figure 1B and 1D).

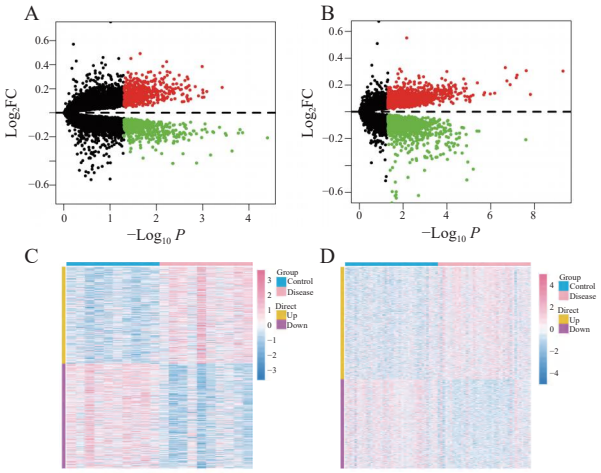

Figure 1 Volcano plots and heatmaps of the differentially expressed genes in the GSE7429 and GSE56815 datasets

A and B, volcano plots of the differentially expressed genes from the GSE7429 and GSE56815 datasets, respectively. There was no significant difference in genes represented by black color. Genes represented by red color were up-regulated and green down-regulated. C and D, heatmaps of the differentially expressed genes from the GSE7429 and GSE56815 datasets in the two groups, respectively. The yellow square in the grid represented the up-regulation of the genes, and the purple square represented the down-regulation of the genes.

3.2 Key genes screening

The top 2 000 genes were selected and intersected with the differentially expressed genes in the GSE7429 and GSE56815 datasets, resulting in a final set of 10 common genes (Figure 2A). The selection of top 2 000 genes not only ensures accuracy but also controls the number of genes in a reasonable range. LASSO logistic regression analysis identified a consensus of seven genes as the signature genes of OP (Figure 2B and 2C). Furthermore, SVM-RFE algorithm was leveraged to appraise the featured genes in the context of OP, resulting in the identification of the top 7 genes with the highest predictive accuracy (Figure 2D). The genes *CP*, *KLK3*, *POLG*, and *TRPV4*, pinpointed by the LASSO logistic regression analysis and SVM-RFE algorithm, were recognized as key genes for subsequent investigation (Figure 2E).

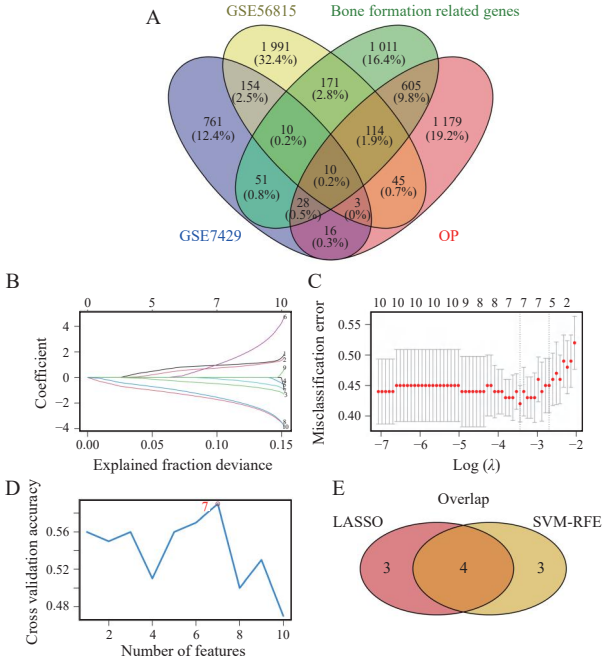

Figure 2 Screening and analysis of key genes associated with OP

A, the intersection of differentially expressed genes concerning OP in the GSE7429 and GSE56815 datasets. B, fraction deviance explained by LASSO logistic regression analysis. C, misclassification error by LASSO logistic regression analysis. D, feature selection accuracy using SVM-RFE algorithm. E, Venn diagram of genes identified by LASSO and SVM-RFE.

3.3 GO and KEGG analyses of key genes

GO analysis revealed that *CP*, *KLK3*, *POLG*, and *TRPV4* were predominantly associated with the positive regulation of defense responses, the maintenance of cellular metal ion homeostasis and the production of antimicrobial peptides (Figure 3A). KEGG pathway analysis indicated that these genes were significantly enriched in three distinct signaling pathways: ferroptosis, porphyrin metabolism, and base excision repair (P = 0.019, P = 0.014, and P = 0.020, respectively, Figure 3B).

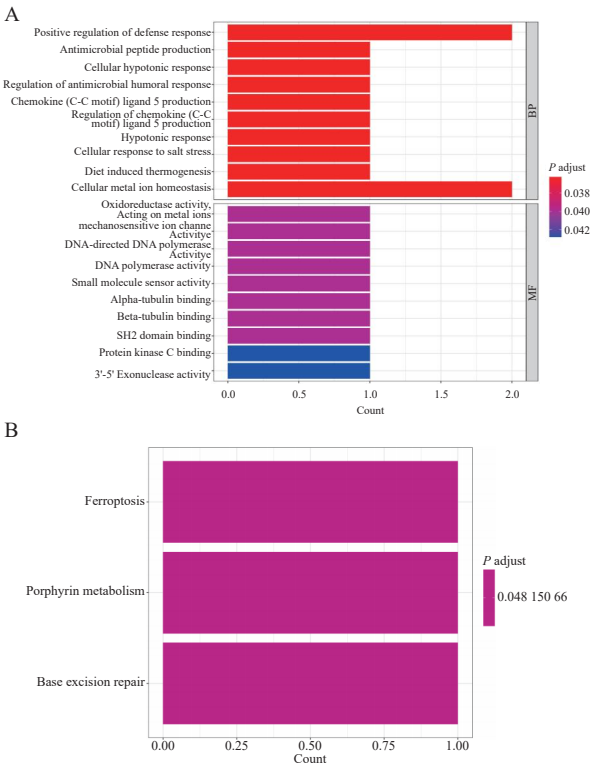

Figure 3 GO and KEGG enrichment analyses of key genes

A, the top 10 items of GO enrichment. B, the top 3 pathways of KEGG enrichment.

3.4 Immune cell infiltration analysis

Each patient was found to have 22 distinct types of immune cells, including natural killer (NK) cells, regulatory T cells (Tregs), macrophages (M0, M1, and M2), naïve and memory B cells, and $CD4^+$ and $CD8^+$ T cells (Figure 4A). Macrophages (M0, M1, and M2) and resting dendritic cells demonstrated a significant correlation with one another (P = 0.021, P = 0.007, P = 0.026, and P = 0.018, respectively, Figure 4B and 4C). *CP* exhibited a positive correlation with resting dendritic cells (P = 0.037) and activated NK cells (P = 0.124), while it exhibited a negative correlation with resting NK cells (P = 0.168) and monocytes (P = 0.415, Figure 4D). *KLK3* was found to have a positive correlation with Tregs (P = 0.032) and naïve B cells (P = 0.102), and a negative correlation with eosinophils (P = 0.179) and resting mast cells (P = 0.227) (Figure 4E). *POLG* demonstrated a positive correlation with Tregs (P < 0.001) and naïve B cells (P < 0.001), and a negative correlation with resting dendritic cells (P < 0.001) and CD4 memory T cells (P < 0.001, Figure 4F). *TRPV4* was positively correlated with Tregs (P < 0.001) and plasma cells (P = 0.040), and negatively correlated

with resting dendritic cells (P = 0.001) and M2 macrophages (P = 0.004, Figure 4G).

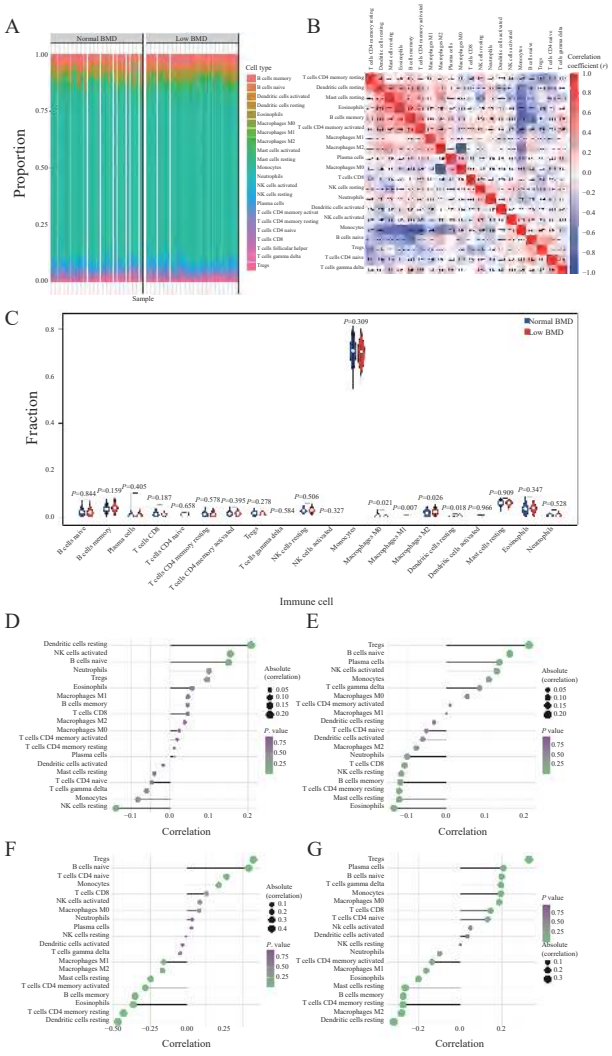

Figure 4 Comprehensive analysis of immune cell infiltration and gene expression correlations
A, immune cell infiltration histogram. B, the correlation between 22 kinds of immune cells. Blue indicates negative correlation, and red indicates positive correlation. C, differences in immune cell content between the two groups of patients. Blue indicates negative correlation, and red indicates positive correlation. D – G, correlation between CIBERSORT and the expressions of key gene *CP*, *KLK*, *POLG*, and *TRPV4*, respectively.

3.5 The prediction of specificity and sensitivity of key genes in disease diagnosis with ROC curves

The area under the ROC curve (AUC) is a performance index to measure the quality of a binary classification model. The value range of AUC is between 0.5 and 1. A higher AUC denotes superior predictive performance for each gene. Based on the GSE7429 dataset, the expression differences of key genes were displayed (Figure 5A). The findings revealed that the AUC for *CP*, *KLK3*, *POLG*, and *TRPV4* were 0.770, 0.810, 0.810, and 0.800, respectively (Figure 5B – 5E).

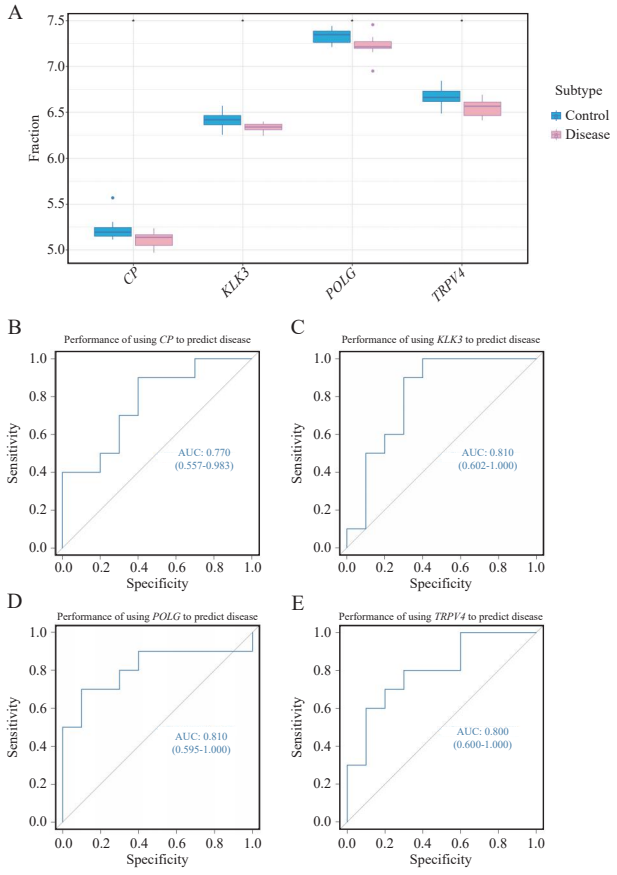

Figure 5 Specificity and sensitivity of key genes in OP
A, the expression levels of key genes from the GSE7429 database. B – E, the ROC curve of *CP*, *KLK3*, *POLG*, and *TRPV4* genes for predicting disease occurrence, respectively.

3.6 Screening of TCM active ingredients targeting the key genes

The findings indicated that resveratrol, curcumin, and quercetin exhibited substantial interaction counts with *KLK3*, with values of 7, 3, and 2, respectively. However, no TCM active ingredients were found to have interactions with *CP*, *POLG*, or *TRPV4* (Table 1).

Table 1 Interactions between TCM active ingredients and *KLK3*

Gene	Active ingredient	ID number	Interaction count
KLK3	Resveratrol	D000077185	7
	curcumin	D003474	3
	quercetin	D011794	2

3.7 Molecular docking

The binding energies for KLK3 with curcumin, quercetin and resveratrol were determined to be – 6.9, – 8.7, and – 7.2 kcal/mol, respectively, indicating good, robust, and strong binding affinities (Figure 6).

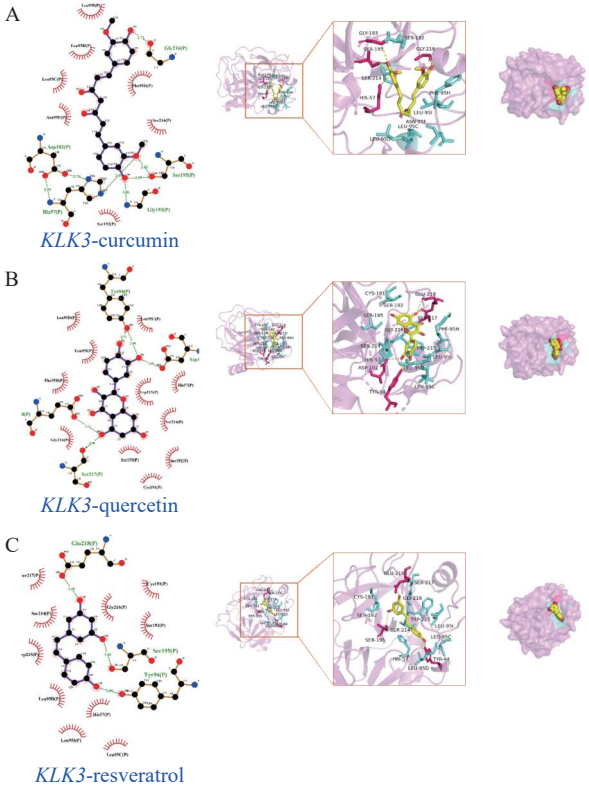

Figure 6 Molecular docking of curcumin, quercetin, and resveratrol with *KLK3*

A – C, molecular docking of quercetin, resveratrol, and curcumin with *KLK3*, respectively.

3.8 Molecular dynamics simulation analysis

RMSD calculations showed that the average RMSD values for all complexes were all less than 2 Å, indicating that they essentially achieved dynamic equilibrium around 20 ns (Figure 7A). Notably, the *KLK3*-curcumin complex exhibited significantly less RMSD fluctuation compared with *KLK3*-quercetin and *KLK3*-resveratrol. Additionally, a salt bridge formed between GLU-218 of *KLK3* and ARG-95G contributed significantly to the local stability of the complex (Figure 7B). Curcumin formed multiple hydrogen bonds with the protein site residues SER-217, TYR-94, and SER-192, and its benzene ring engaged in π-π conjugation with HIS-57. These interactions suggest that the small molecule can still form stable complexes with proteins after undergoing kinetic adjustments (Figure 7C – 7E). RMSF results showed that a small portion of amino acids in the *KLK3*-ligand complex experienced significant conformational changes, particularly near residues 90 – 111 (Figure 7F). The majority of the amino acid conformations remained within an acceptable range. Additionally, radius of gyration (Rg) measurements indicated a reduction in *KLK3*'s stability by prompting more hydrophobic contacts and fostering more effective internal interactions (Figure 7G).

From the energy composition analysis, van der Waals

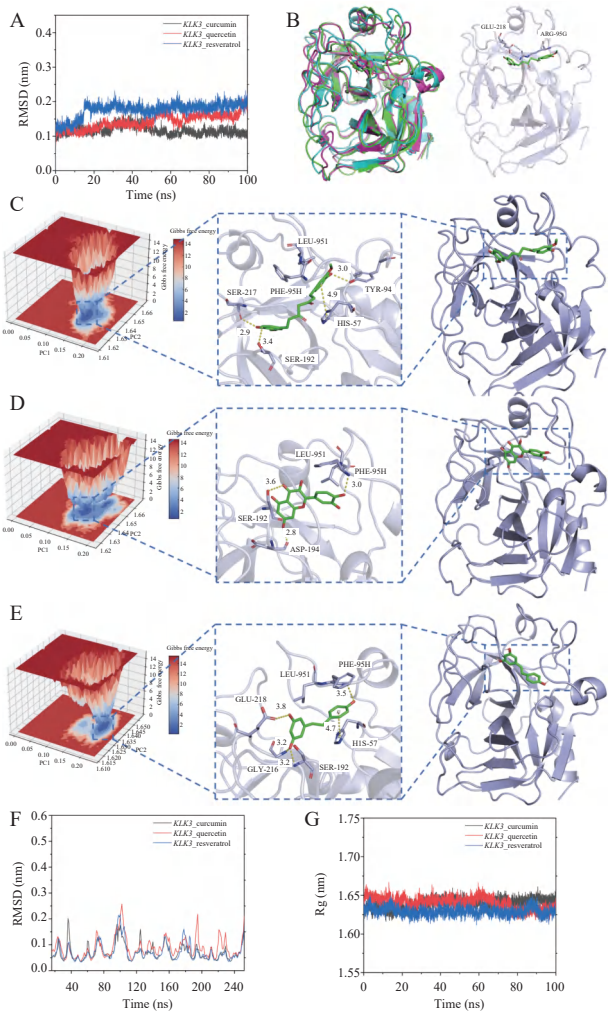

Figure 7 Molecular dynamics simulations and binding analysis of curcumin, quercetin, and resveratrol with *KLK3*

A, the RMSD of curcumin, quercetin, and resveratrol with *KLK3*. B, the 3D structure of complex after molecular dynamics simulation. C – E, the binding mode of curcumin, quercetin, and resveratrol proteins with *KLK3* after 100 ns molecular dynamics simulation, respectively. F, the RMSD of curcumin, quercetin, and resveratrol with *KLK3*. G, the Rg of curcumin, quercetin, and resveratrol with *KLK3*.

force energy (E_{VDW}) plays a major role, which indicates that the compound can be stably maintained in the pocket of the protein site. In addition, due to the effective hydrogen bond interaction between the compound and the protein pocket, the electrostatic (E_{ELE}) interaction also makes an important contribution in stabilizing small molecules. Because the compound itself has certain hydrophobicity, the non-polar contribution to solvation energy (E_{SA}) also contributes to the binding of the protein to the compound. The negative value of $G_{binding\ energy}$ highlights the stability of the system, while the positive value shows instability. The generalized born energy (E_{GB}) reflects the solvation effect of polar molecules in solvents. The positive value indicates that the solvation energy of molecules in solvents increases, which may indicate that the molecules are not stable in solvents (Table 2).

Table 2 Binding energy of *KLK3* with curcumin, quercetin, and resveratrol by MMGBSA method (KJ/mol)

Type	*KLK3*-curcumin	*KLK3*-quercetin	*KLK3*-resveratrol
E_{VDW}	– 217.32 ± 10.74	– 182.66 ± 13.02	– 183.19 ± 16.06
E_{ELE}	– 82.33 ± 15.59	– 62.34 ± 10.57	– 79.55 ± 19.21
E_{GB}	192.17 ± 13.45	149.68 ± 15.80	153.28 ± 21.13
E_{SA}	– 21.18 ± 0.91	– 14.01 ± 0.81	– 14.23 ± 0.88
$G_{binding\ energy}$	– 128.66 ± 15.42	– 109.37 ± 8.67	– 123.69 ± 14.36

E_{VDW}, van der Waals energy. E_{ELE}, electrostatic energy. E_{GB}, polar contribution to solvation. E_{SA}, non-polar contribution to solvation.

4 Discussion

4.1 Identification of key genes

Excessive bone resorption over bone formation mediated by ostoclasts is the primary cause of bone loss and a major pathogenic factor in OP [14]. In pursuit of biomarkers with diagnostic and therapeutic significance for OP, this investigation employed bioinformatics techniques and machine learning algorithms to identify genes differentially expressed in relation to bone formation and to elucidate their potential regulatory mechanisms, aiming to furnish novel insights into the clinical prophylaxis and treatment of OP. *CP*, *KLK3*, *POLG*, and *TRPV4* were ultimately selected as genes of interest. The GO and KEGG pathway enrichment analyses revealed that these genes were predominantly involved in the positive regulation of defense responses through ferroptosis, porphyrin metabolism, and base excision repair (BER) signaling pathways, cellular metal ion homeostasis, and the regulation of chemokine ligand 5 production, among other processes, where they exerted their biological effects. Ferroptosis is an iron-dependent mode of cell death that significantly impacts bone metabolism. Inhibiting ferroptosis in osteoblasts and promoting it in osteoclasts represent a novel strategy for the prevention and treatment of OP [15]. BER is a repair mechanism that emerges in response to cellular oxidative stress and exogenous damage, safeguarding deoxyribonucleic acid (DNA) integrity. By rectifying DNA damage and chain breaks caused by endogenous and exogenous mutations, BER provides a protective effect against the DNA damage associated with aging, thus preventing related pathologies [16]. The expression levels of the aforementioned four pivotal genes were ascertained through GSE7429, revealing a down-regulation trend for *CP*, *KLK3*, *POLG*, and *TRPV4*. The efficacy prediction scores, ranked by the AUC, were led by *KLK3*, followed by *POLG*, *TRPV4*, and *CP*.

4.2 Analysis of the current research status of *KLK3*, *POLG*, *TRPV4*, and *CP*

The *KLK3* gene may be just one of many genes that influence OP risk. However, the *KLK3* gene itself is not a major risk factor for OP, but as part of the polygenic score, it may have some impact on an individual's overall risk. Prostate cancer is associated with decreased bone mineral density and an increased risk of fractures, and the loss of bone mass and muscle mass is further exacerbated during androgen deprivation therapy [17]. The potential of *KLK3* as a target gene in OP treatment warrants further exploration. The *POLG* gene encodes the catalytic subunit of mitochondrial DNA polymerase γ (*Polγ*), responsible for mitochondrial DNA replication and a frequent cause of multiple mitochondrial deletions [18]. Mitochondria serve as the primary energy source for cells and play a critical role in balancing osteoblasts and osteoclasts activity. Mitochondrial dysfunction can lead to metabolic disorders, reduced adenosine triphosphate (ATP) production, increased reactive oxygen species (ROS) production, enhanced apoptosis, and ultimately induce OP [19]. The question of whether *POLG* gene regulation of mitochondrial function can influence bone metabolism is a novel research avenue for the future. *TRPV4,* a non-selective cation channel with moderate permeability and high Ca^{2+} ion expression, plays a significant role in the transmission of mechanical signals and substantially impacts bone metabolism [20]. Exposure to a hypotonic solution activates the *TRPV4* channel, allowing calcium ions to enter the cell and triggering intracellular signaling transduction pathways. *TRPV4* exerts a pivotal role in bone health by regulating calcium reabsorption and transport, and by responding to steroid hormones and agonists [21]. The *CP* gene typically encodes the coat protein, a key component of the viral particle that protects the viral nucleic acid and aids in viral fusion with host cells, playing a critical role in the infection process [22]. Currently, no empirical scientific evidence substantiates a causal relationship between the *CP* gene and the onset of OP. The *CP* gene is predominantly associated with the encoding of viral coat proteins, whereas OP, a bone health condition, is influenced by a complex interplay of genetic and environmental factors. Consequently, our postulated hypothetical mechanism is predicated upon conjecture regarding possible interactions among viral infection, immune response, and bone metabolism.

4.3 Immune cell infiltration analysis

Research has substantiated the pivotal role of the immune system in regulating bone metabolism, with

immune cells exerting influence over the process of bone formation through direct cell-to-cell contact or paracrine secretion mechanisms [23]. Immune cells, including macrophages, dendritic cells and monocytes, are the principal effectors of inflammation and constitute the central conduit leading to OP by producing proinflammatory cytokines that foster chronic age-related inflammation [21]. Immune cell infiltration analysis employing the CIBERSORT algorithm has revealed a strong correlation between key genes and immune cells, with *CP* positively associated with resting dendritic cells, and *KLK3*, *POLG*, and *TRPV4* positively associated with Tregs. Treg cells, a subset of $CD4^+$ T cells, upon activation, can mitigate inflammatory responses, inhibit the function and generation of osteoclasts, and represent efficacious targets for diminishing bone loss and enhancing BMD [24]. Notably, significant disparities in the distribution of macrophages M0, M1, M2, and resting dendritic cells were observed between the groups. Macrophages can polarize in different directions due to varying microenvironments and stimuli, which may promote osteoblast differentiation and contribute to bone formation [25]. Therefore, strategies such as modifying the infiltration status of immune cells, enhancing the immune microenvironment, and regulating macrophage polarization could effectively promote bone formation and prevent OP.

4.4 Research progress of resveratrol, curcumin, and quercetin in bone metabolism

In TCM, OP is classified under the categories of “bone impotence” and “bone wither”. The primary etiological factors and mechanisms of the disease are considered to be deficiencies in the liver, spleen, and kidney, along with imbalances in Qi and blood. The treatment of OP by TCM has significant therapeutic efficacy, high safety, good patient compliance, and many other advantages. Studies have confirmed that kidney tonic Chinese medicines such as eucommia, bone marrow, and horny goat weed, as well as the flavonoids of scutellaria and astragalus, can exert estrogen-like effects and effectively maintain the positive balance of bone metabolism [26]. The potential for developing new anti-OP TCM formula and natural active ingredients is obvious, highlighting the importance of these ingredients in the creation of innovative treatments [27]. Regrettably, the Coremine Medical database failed to present ideal TCM formulas for targeted gene therapy against OP. Luckily, resveratrol, curcumin, and quercetin-active components in TCM with action on *KLK3* were identified from CTD. These three compounds regulate bone metabolism by promoting bone formation and inhibiting bone resorption, providing a range of benefits as a complementary therapy [28]. Studies have demonstrated that resveratrol exerts a bone-protective effect in postmenopausal, aged, and disused OP rat models, promoting the osteogenic differentiation of mesenchymal stem cells, stimulating osteoblast proliferation, and inhibiting osteoclast activity through multiple signaling pathways and protein molecules [29]. It was shown that regular supplementation with 75 mg of resveratrol twice a day was effective in slowing down bone mass loss in the lumbar spine and femoral neck in postmenopausal women, and in reducing the incidence of fractures [30]. Curcumin can promote bone formation and inhibit bone resorption by regulating various proteins, enzymes, and signaling pathways, thereby effectively preventing and controlling the occurrence of OP [31]. A study showed that curcumin attenuated oxidative stress-induced osteoblast dysfunction through the glycogen synthase kinase-3β (GSK-3β)/nuclear factor erythroid 2-related factor 2 (Nrf2) signaling pathway, thereby promoting the proliferative activation of osteoblasts [32]. Quercetin can inhibit RANKL-mediated osteoclast formation, osteoblast apoptosis, oxidative stress, and inflammatory response [33]. It was reported that quercetin and vitamin E effectively attenuated ovariectomy-induced osteoporosis in rats, possibly promoting bone formation and inhibiting bone resorption by regulating cellular autophagy and apoptosis [34]. The results of molecular docking further indicated that resveratrol, curcumin, and quercetin exhibited a robust binding affinity with the target gene *KLK3*. Molecular dynamics simulation verified that curcumin, quercetin, and resveratrol had strong affinity with *KLK3*, and these affinities can promote the formation of the TCM active ingredients into stable complexes with *KLK3* to exert biological effects. However, there was no direct evidence suggesting that resveratrol, curcumin, and quercetin exerted biological effects by acting on *KLK3*. Subsequent experimental verification is still needed to clarify the regulatory mechanisms of these three components in the bone-immune system. Admittedly, this study has its limitations, such as the lack of empirical validation of the predicted outcomes through subsequent experimental follow-up. Still, the study represents a critical area for future research endeavors.

5 Conclusion

In this study, we employed bioinformatics technologies and machine learning algorithms to identify *CP*, *KLK3*, *POLG*, and *TRPV4* as pivotal genes associated with OP. These genes demonstrated a significant correlation with immune cell infiltration and exhibited high predictive accuracy for the diagnosis of OP. Resveratrol, curcumin, and quercetin were pinpointed as prospective TCM active ingredients with potential to target *KLK3* gene, thereby facilitating bone formation and effectively modulating OP. These newer techniques provide a more comprehensive view of transcriptome changes and allow a deeper

understanding of the pathogenesis of OP, which is significant for revealing the pathological mechanism and expanding the clinical prevention and treatment strategies. For future research, the combination of multi-omics methods such as genomics, proteomics, and metabolomics with transcriptomics will enhance our understanding of the potential molecular mechanisms of OP. In addition, more advanced machine learning models and artificial intelligence-driven algorithms can be applied to improve the gene selection process and improve prediction accuracy.

Fundings

National Natural Science Foundation of China (81960877).

Competing interests

The authors declare no conflict of interest.

References

[1] LEBOFF MS, GREENSPAN SL, INSOGNA KL, et al. The clinician's guide to prevention and treatment of osteoporosis. Osteoporosis International, 2022, 33(10): 2049–2102.

[2] ADEJUYIGBE B, KALLINI J, CHIOU D, et al. Osteoporosis: molecular pathology, diagnostics, and therapeutics. International Journal of Molecular Sciences, 2023, 24(19): 14583.

[3] MUNIYASAMY R, MANJUBALA I. Insights into the mechanism of osteoporosis and the available treatment options. Current Pharmaceutical Biotechnology, 2024, 25(12): 1538–1551.

[4] SRIVASTAVA RK, SAPRA L. The rising era of "immunoporosis": role of immune system in the pathophysiology of osteoporosis. Journal of Inflammation Research, 2022, 15: 1667–1698.

[5] TONI R, DI CONZA G, BARBARO F, et al. Microtopography of immune cells in osteoporosis and bone lesions by endocrine disruptors. Frontiers in Immunology, 2020, 11: 1737.

[6] UDAGAWA N, KOIDE M, NAKAMURA M, et al. Osteoclast differentiation by RANKL and OPG signaling pathways. Journal of Bone and Mineral Metabolism, 2021, 39(1): 19–26.

[7] AN J, YANG H, ZHANG Q, et al. Natural products for treatment of osteoporosis: the effects and mechanisms on promoting osteoblast-mediated bone formation. Life Sciences, 2016, 147: 46–58.

[8] LIU SY, WANG ZT, ZHU RH, et al. Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. Journal of Visualized Experiments, 2021(175): e62528.

[9] GARCIA-CARRETERO R, VIGIL-MEDINA L, BARQUERO-PEREZ O, et al. Logistic LASSO and elastic net to characterize vitamin D deficiency in a hypertensive obese population. Metabolic Syndrome and Related Disorders, 2020, 18(2): 79–85.

[10] WEI HU, SHU W, YU HB, et al. Bioinformatic analysis on hepatic metabolic disorders induced by triphenyl phosphate. Journal of Environmental and Occupational Medicine, 2020, 37(2): 111–120.

[11] XU D, CHU MM, CHEN YY, et al. Identification and verification of ferroptosis-related genes in the pathology of epilepsy: insights from CIBERSORT algorithm analysis. Frontiers in Neurology, 2023, 14: 1275606.

[12] DAVIS AP, WIEGERS TC, JOHNSON RJ, et al. Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Research, 2023, 51(D1): D1257–D1262.

[13] CHEN XJ, WANG JY, TANG L, et al. The therapeutic effect of Fufang Zhenshu Tiaozhi (FTZ) on osteoclastogenesis and ovariectomized-induced bone loss: evidence from network pharmacology, molecular docking and experimental validation. Aging, 2022, 14(14): 5727–5748.

[14] KAUR M, NAGPAL M, SINGH M. Osteoblast-n-osteoclast: making headway to osteoporosis treatment. Current Drug Targets, 2020, 21(16): 1640–1651.

[15] GAO L, HUA WZ, TIAN LX, et al. Molecular mechanism of ferroptosis in orthopedic diseases. Cells, 2022, 11(19): 2979.

[16] DIANOV GL, HÜBSCHER U. Mammalian base excision repair: the forgotten archangel. Nucleic Acids Research, 2013, 41(6): 3483–3490.

[17] TUCK SP, HANUSCH B, WALKER J, et al. Prostate cancer and osteoporosis. Current Osteoporosis Reports, 2013, 11(1): 11–20.

[18] RAHMAN S, COPELAND WC. POLG-related disorders and their neurological manifestations. Nature Reviews Neurology, 2019, 15: 40–52.

[19] GAO XY, YU X, ZHANG C, et al. Telomeres and mitochondrial metabolism: implications for cellular senescence and age-related diseases. Stem Cell Reviews and Reports, 2022, 18(7): 2315–2327.

[20] LIU N, LU WW, DAI XL, et al. The role of TRPV channels in osteoporosis. Molecular Biology Reports, 2022, 49(1): 577–585.

[21] FISCHER V, HAFFNER-LUNTZER M. Interaction between bone and immune cells: implications for postmenopausal osteoporosis. Seminars in Cell & Developmental Biology, 2022, 123: 14–21.

[22] CHIANG CH, WANG JJ, JAN FJ, et al. Comparative reactions of recombinant papaya ringspot viruses with chimeric coat protein (*CP*) genes and wild-type viruses on CP-transgenic papaya. The Journal of General Virology, 2001, 82(Pt 11): 2827–2836.

[23] SAXENA Y, ROUTH S, MUKHOPADHAYA A. Immunoporosis: role of innate immune cells in osteoporosis. Frontiers in Immunology, 2021, 12: 687037.

[24] MUÑOZ J, AKHAVAN NS, MULLINS AP, et al. Macrophage polarization and osteoporosis: a review. Nutrients, 2020, 12(10): 2999.

[25] TOITA R, KANG JH, TSUCHIYA A. Phosphatidylserine liposome multilayers mediate the M1-to-M2 macrophage polarization to enhance bone tissue regeneration. Acta Biomaterialia, 2022, 154: 583–596.

[26] YUAN YL, YANG Z, DING W, et al. Effect and mechanism of Chinese medicine in treatment of osteoporosis. Chinese Journal of Experimental Traditional Medical Formulae, 2024, 30(4): 290–298.

[27] ZHANG ND, HAN T, HUANG BK, et al. Traditional Chinese medicine formulas for the treatment of osteoporosis: implication for antiosteoporotic drug discovery. Journal of Ethnopharmacology, 2016, 189: 61–80.

[28] INCHINGOLO AD, INCHINGOLO AM, MALCANGI G, et al. Effects of resveratrol, curcumin and quercetin supplementation

on bone metabolism — a systematic review. Nutrients, 2022, 14(17): 3519.
[29] AHMAD HAIRI H, JAYUSMAN PA, SHUID AN. Revisiting resveratrol as an osteoprotective agent: molecular evidence from *in vivo* and *in vitro* studies. Biomedicines, 2023, 11(5): 1453.
[30] WONG RH, THAUNG ZAW JJ, XIAN CJ, et al. Regular supplementation with resveratrol improves bone mineral density in postmenopausal women: a randomized, placebo-controlled trial. Journal of Bone and Mineral Research, 2020, 35(11): 2121–2131.
[31] WU HY, YU HT, KANG B, et al. Curcumin nanoparticles and the therapeutic potential of curcumin for musculoskeletal disorders. European Review for Medical and Pharmacological Sciences, 2023, 27(20): 9680–9702.
[32] LI XM, CHEN Y, MAO YX, et al. Curcumin protects osteoblasts from oxidative stress-induced dysfunction via GSK3β-Nrf2 signaling pathway. Frontiers in Bioengineering and Biotechnology, 2020, 8: 625.
[33] WONG SK, CHIN KY, IMA-NIRWANA S. Quercetin as an agent for protecting the bone: a review of the current evidence. International Journal of Molecular Sciences, 2020, 21(17): 6448.
[34] VAKILI S, ZAL F, MOSTAFAVI-POUR Z, et al. Quercetin and vitamin E alleviate ovariectomy-induced osteoporosis by modulating autophagy and apoptosis in rat bone cells. Journal of Cellular Physiology, 2021, 236(5): 3495–3509.

骨质疏松症骨形成相关基因与免疫细胞浸润的分析及靶向中药活性成分预测

王凯[a†], 董平[b†], 郭洪章[a*]

a. 甘肃省人民医院骨科, 甘肃 兰州 730000, 中国
b. 内蒙古医科大学中医学院, 内蒙古 呼和浩特 010110, 中国

【摘要】目的 基于生物信息学技术与机器学习探讨骨形成相关基因在骨质疏松症（OP）中的差异表达及作用机制，并预测靶向中药活性成分。**方法** 从高通量基因表达 (GEO) 与 GeneCards 数据库获取疾病靶点和骨形成相关基因，使用 R 语言包筛选差异表基因。最小绝对收缩和选择算子（LASSO）逻辑回归分析和支持向量机-递归特征消除（SVM-RFE）算法筛选 OP 的特征基因，选择关键基因进行基因本体（GO）与京都基因与基因组百科全书（KEGG）通路富集分析。使用估算 RNA 转录物相对子集的细胞类型识别（CIBERSORT）算法进行免疫细胞浸润分析，并对基因表达量以及免疫细胞含量进行 Spearman 相关性分析。Coremine Medical 数据库获取治疗 OP 的潜在有效中药，比较毒理学数据库（CTD）预测作用于关键基因的靶向中药活性成分，AutoDock Vina 1.2.2 与 GROMACS 2020 软件分析活性成分与目标靶标之间的结合能及相作模式。**结果** 从 GEO 和 GeneCards 数据库获得交集基因 10 个，LASSO 逻辑回归分析和 SVM-RFE 算法筛选出 4 个关键基因：外壳蛋白（*CP*）、激肽释放酶 3（*KLK3*）、聚合酶 γ（*POLG*）和瞬时受体电位香草酸受体 4（*TRPV4*）。GO 功能与 KEGG 通路富集分析显示，靶基因主要通过铁死亡、叶啉代谢和碱基切除修复信号通路参与调控防御反应的正调控、细胞金属离子稳态、调控趋化因子配体 5 的生成等过程而发挥生物学效应。免疫浸润分析显示关键基因与免疫细胞高度相关，巨噬细胞 M0、M1、M2 和静息树突状细胞在组间具有显著差异，不同组间均表现出显著差异（$P < 0.05$）。白藜芦醇、姜黄素和槲皮素与 *KLK3* 的相互作用数分别为 7、3 和 2，表明白藜芦醇、姜黄素和槲皮素与 *KLK3* 有明显的相互作用。分子对接与动力学模拟显示活性成分与靶基因之间具有良好的结合能力。**结论** 关键基因 *CP*、*KLK3*、*POLG* 和 *TRPV4* 在 OP 诊断中具有良好的预测效能，白藜芦醇、姜黄素以及槲皮素可能是靶向骨形成基因 *KLK3* 干预 OP 的潜在中药活性成分。这项研究为 OP 发病机制的阐释和临床药物的研发提供了科学依据。

【关键词】骨质疏松症；骨形成；差异表达基因；生物信息；中药；活性成分；分子机制