Bioinformatics approach for potential genes associated with osteoarthritis

Yi Chai
Nanjing University of Chinese Medicine
Cong-Ying Mai
Nanjing University of Chinese Medicine
Ming-Yue Zhang
Nanjing University of Chinese Medicine
Dao-Ming Xu
Affiliated hospital of nanjing university of traditional chinene medicine
Feng Tan
Nanjing University of Chinese Medicine
Xing Li
Nanjing University of Chinese Medicine
Qiaoling Fan (tcmscience@163.com)
Nanjing University of Chinese Medicine

Research article

Keywords: osteoarthritis, bioinformatics analysis, microarray, genes

Posted Date: January 24th, 2020

DOI: https://doi.org/10.21203/rs.2.21759/v1

Abstract

Background: Osteoarthritis is the main cause of disability and pain. Due to limited understanding of the disease mechanism, it is generally difficult to diagnose early and eliminate sequelae. The aim of this study was to identify novel biomarkers for osteoarthritis (OA) using a bioinformatics approach. Method: The gene expression dataset GSE82107 were obtained from the Gene Expression Omnibus (GEO) database. Matrix quality assessment was performed using corrplot packages and principal component analysis. The differentially expressed genes (DEGs) picked out using GEO2R tool. Enrichment analyses were performed using The Database for Annotation, Visualization and Integrated Discovery and Gene Set Enrichment Analysis (GSEA). Weighted correlation network analysis (WGCNA) was used to find gene modules highly associated with OA. Cytoscape with Molecular Complex Detection (MCODE) plug-in was utilized to analyze protein-protein interaction of these DEGs. Receiver operator characteristic curve analysis was used to evaluate the diagnostic effectiveness of genes. Results: Samples with different conditions (HC and OA) are obviously distinguished, and the distance between biological replicates is relatively close. In total, 1676 DEGs were identified. Enrichment analysis showed that there were some gene sets related to OA pathology, such as chondrocyte development. The results of WGCNA analysis showed that 298 genes were most positive associated with OA. 10 common genes obtained were selected as candidate core genes. ROC results showed that 5 of these genes had the greatest diagnostic value. Conclusion: This is the first study to identify biomarkers related to OA by combining multiple algorithms such as GSEA, WGCNA, MCODE and ROC. We suggest that GLG1, PAPSS2, CTSK, TIMP1 and SDC1 could serve as valuable biomarkers. Further studies are needed to examine the precise role and mechanism of these genes in OA.

Background

Osteoarthritis (OA) is considered to be the most major chronic joint disease, which is the main cause of pain and disability [1, 2]. It is estimated that the percentage of the aging population over the age of 65 will more than double in the next 20 years in Asia and the incidence of OA is rising due to the increase in the ageing population [3]. Clinicians recognize that OA is usually diagnosed later in the process of the disease, and the incidence of serious adverse events caused by treatment such as total knee arthroplasty is higher than non-surgical treatment [4]. Therefore, despite efforts to develop disease markers in the past, it is still necessary to identify effective biochemical markers to reliably describe OA pathological processes, or to make an early diagnosis of OA, or to follow the course and therapeutic effects of OA.

Bioinformatics has become an important part of scientific research to analyze, identify and interpret biomarkers and therapeutic targets [5]. Many bioinformatic studies on OA identified uncover differentially expressed genes (DEGs) directly, but quite a few bioinformatics studies on OA have not previously conducted quality assessments on gene expression matrix, which may lead to reduce the reliability of results [6]. Though DEGs usually have a significantly different expression, enrichment analysis only for DEGs may also miss important genes with insignificant expression in OA. In addition, the interesting gene sets enriched by Gene Set Enrichment Analysis (GSEA) are often considered as candidate for hub genes.

But these genes may be in a non-concentrated interaction region of the whole biological network leading to the poor linkage between genes. Weighted correlation network analysis (WGCNA) uses gene expression datasets to construct weighted gene co-expression network mining synergistically expressed gene modules and exploring the relationship between gene modules and biological phenotypes [7]. However, few studies have used WGNCA to identify the co-expression gene modules of OA to find core genes. The bioinformatic studies using this method on OA only included DEGs for WGCNA analysis, which may also lead to bias of the results. Besides, gene modules associated with biological phenotypes and statistical significance analyzed by WGCNA are performed in GSEA to identify core genes. These designs focus on preset subsets of the gene expression dataset may cause some implicit information to be ignored in all gene sets.

In the present study, the gene expression dataset GSE82107 was selected from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo) [8]. Data quality assessment was performed using correlation analysis and principal component analysis (PCA) firstly, then the GEO2R online analysis software (https://www.ncbi.nlm.nih.gov/geo/geo2r/) was used to uncover DEGs. In addition, the functions of the DEGs and all genes were analyzed respectively, using the Database for Annotation, Visualization and Integrated Discovery (DAVID) and GSEA [9–11]. The study also used WGCNA to construct gene co-expression network and explore the correlation within gene modules and biological phenotypes of OA screening important gene modules [7]. Using these loci, a protein-protein interaction (PPI) network of DEGs was obtained and network analysis and Molecular Complex Detection (MCODE) were performed in order to get degrees of connectivity for each gene and interactive modules [12]. We detected common genes in each gene set obtained based on the above methods and identified the relationship between genes and musculoskeletal diseases using comparative toxicogenomics database (CTD) [13]. Receiver operator characteristic (ROC) analysis was used to evaluate the diagnostic effectiveness of genes providing a basis for further studies.

Materials And Methods

Data

The gene expression dataset GSE82107 was obtained from the publicly accessible GEO database (http://www.ncbi.nlm.nih.gov/geo/). The study was based on the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array; Affymetrix; Thermo Fisher Scientifc, Inc., Waltham, MA, USA). The series matrix file of GSE82107 was downloaded from the GEO database. The dataset includes synovium samples from 7 individuals without a joint disease and 10 patients suffering from end-stage OA [14].

Matrix quality assessment

After the gene expression matrix is obtained, it is important to test and analyze the relationship between biological samples and experimental design. Usually, the methods of sample clustering and PCA are used to calculate and visualize the matrix. In this study, generate the numerical correlation matrix using cor, and we use the corrplot packages to graph the correlation matrix [15]. At the same time, p-values were

calculated, which was concidered to be statistically significant at $P < 0.01$. In addition, PCA online tool was used to show the distribution of sample points on the two-dimensional plane, and the distribution of sample points indicates whether samples in the same group are united, or samples in different groups are obviously separated.

Data processing for DEG identification

GEO2R was used to identify DEGs among experimental samples. GEO2R offers a convenient interface enabling sophisticated R-based analyses of GEO data and is useful to identify and evaluate DEGs [8]. $P < 0.05$ and $|logFC| \geq 1.5$ were set as the thresholds for DEGs detection.

Enrichment analysis of DEGs

Gene Ontology (GO) analyses used to annotate genes or gene products and to determine biological characteristics of high-throughput genome or transcriptome data and Kyoto Encyclopedia of Genes and Genomes (KEGG) that is a group of databases for various biological data, including genomes and biological pathways were carried out using DAVID v6.8 (https://david.ncifcrf.gov/) [16]. The Functional Annotation Tool of DAVID was used and then the upregulated and downregulated DEGs were inserted into the tool for GO and KEGG analysis, respectively. The data was downloaded and $P < 0.05$ was used to indicate a statistically signifcant difference. Besides, GSEA (http://www.broadinstitute.org/gsea/index.jsp) was performed to explore whether the identified sets of genes showed statistical differences. | Normalized enrichment score (NES) | > 1 and NOM P value < 0.05 were used to determine the statistical significance.

Construction of the gene co-expression network

Because a sequencing data set may have tens of thousands of probes, many of these genes may not be expressed very differently in each sample. In order to reduce the computational cost, the top 25% of variance genes were screened and weighted co-expression modules were constructed using the WGCNA package [7]. Firstly, we clustered the gene sets and removed them if there were significant outliers. Next, an appropriate soft threshold based on the standard scale-free network was selected and network and module detection were constructed by one-step method. Subsequently, gene modules were correlated with the biological phenotype, and 1000 genes were randomly selected to visualize the correlation of the genes in the modules. In addition, we performed cluster analysis based on gene expression and obtained correlations between each module.

PPI network and module analysis

The Search Tool for the Retrieval of Interacting Genes (STRING) v11 (https://string-db.org/) is an online tool designed to evaluate PPI information and this was used to detect potential associations among the DEGs [17]. The results were input into Cytoscape v3.7.1 (http://www.cytoscape.org/) to map the associations among the DEGs [18]. A confidence score > 0.4 was set as criteria. MCODE v1.5.1 (Bader Lab, University of Toronto, Toronto, Ontario, Canada) was used to evaluate the interactive clusters in the

PPI network in Cytoscape. The study combined degree, GSEA functional enrichment modules, gene co-expression network and DEGs interactive clusters to identify the common genes. The CTD (http://ctdbase.org/) was used to find integrated chemical-gene, chemical-disease, and gene-disease interactions to generate expanded networks and predict novel associations. We used these data to analyze relationships between gene products and musculoskeletal diseases. Thus, relationships between genes and OA and an implied association or association were identified. Receiver operator characteristic (ROC) curve analysis was used to evaluate the diagnostic effectiveness of genes using expression profiling datasets in GraphPad Prism 7.0 (GraphPad Software, Inc., La Jolla, CA) and $P < 0.05$ was considered as a statistical alteration.

Results

1. Access of expression patterns and identification of DEGs.

The study calculated the correlation based on the gene expression matrix in the chip and performed cluster analysis in samples from GSM2183532 to GSM2183548 to explore the composition of the dataset, the potential classification and the internal relationship using the corrplot package to draw the correlation matrix. The results show that 17 samples can be basically distinguished in hierarchical clustering, and samples with the same phenotype are mainly clustered together (Fig. 1A). Meanwhile, dimensionality reduction is performed by PCA, and the distribution of sample points is displayed on two-dimensional plane. The study shows that the samples with different conditions (HC and OA) are obviously distinguished, and the distance between biological replicates is relatively close, indicating that the consistency of biological replicates and the difference of groups are comparatively obvious (Fig. 1B). Subsequently, the gene expression profiles of 17 samples in the gene expression dataset GSE82107, including 7 control and 10 OA samples, were analyzed. A total of 1676 DEGs between the control and OA samples were identified, including 218 upregulated and 1458 downregulated genes (Fig. 1C).

2. GO function and KEGG pathway enrichment analysis.

GO and KEGG pathway enrichment analyses were performed using DAVID and GSEA in order to gain a comprehensive understanding of the functions of the genes. DAVID focuses on the gene sets contained upregulated or downregulated genes, while GSEA can observe the consistency of gene expression in entire dataset with specific functional gene sets to interpret the biological information. The results of the GO analysis from DAVID indicated that upregulated and downregulated genes were enriched for various BP terms, which are showed in Supplemental Fig. 1 and Supplemental Table 1 (top 5). For upregulated DEGs these included ‘response to organic substance’, ‘protein modification process’, ‘cellular protein modification process’, ‘cellular response to chemical stimulus’ and ‘cellular response to organic substance’, and for downregulated DEGs they included ‘regulation of signaling’, ‘regulation of cell communication’, ‘positive regulation of metabolic process’, ‘positive regulation of cellular metabolic process’, ‘regulation of signal transduction’. In the molecular functions category, the upregulated genes were enriched for ‘receptor binding’, ‘protein complex binding’, ‘glycosaminoglycan binding’, ‘cytokine

activity' and 'anion binding', while the downregulated genes were enriched for 'adenyl ribonucleotide binding', 'adenyl nucleotide binding', 'ATP binding', 'molecular function regulator' and 'cytoskeletal protein binding'. A cellular components analysis further demonstrated that the upregulated genes were enriched for 'extracellular region', 'extracellular region part', 'membrane-bounded vesicle', 'extracellular exosome' and 'extracellular vesicle', while the downregulated genes were enriched for 'intrinsic component of plasma membrane', 'integral component of plasma membrane', 'cell junction', 'neuron part' and 'plasma membrane region'. In addition, 10 KEGG pathways were identified, as listed in Supplemental Fig. 2 and Supplemental Table 2 (top 5), involving the 'cytokine-cytokine receptor interaction', 'chemokine signaling pathway', 'ECM-receptor interaction', 'TNF signaling pathway' and 'osteoclast differentiation' for upregulated DEGs, and 'neuroactive ligand-receptor interaction', 'calcium signaling pathway', 'cAMP signaling pathway', 'oxytocin signaling pathway' and 'dopaminergic synapse' for downregulated DEGs.

GSEA analysis revealed that 8 GO terms, including 'bone development' and 'bone morphogenesis', 'bone remodeling', 'cartilage development involved in endochondral bone morphogenesis', 'chondrocyte development', 'chondrocyte differentiation', as well as 'positive regulation of osteoblast differentiation', and 'regulation of bone development', shown significantly differential enrichment in OA phenotype based on NES and NOM P value (Fig. 2, Table 1).

Table 1

Gene Set Enrichment Analysis of genes associated with osteoarthritis.

Term	Size	NES	NOM P value	FDR q value	Leading edge
GO_BONE_DEVELOPMENT	151	-1.63	0.000	0.115	tags = 36%, list = 19%, signal = 44%
GO_BONE_MORPHOGENESIS	77	-1.95	0.000	0.038	tags = 36%, list = 15%, signal = 43%
GO_BONE_REMODELING	35	-1.65	0.008	0.114	tags = 40%, list = 14%, signal = 47%
GO_CARTILAGE_DEVELOPMENT_ INVOLVED_IN_ENDOCHONDRAL_ BONE_MORPHOGENESIS	19	-1.73	0.010	0.089	tags = 53%, list = 15%, signal = 62%
GO_CHONDROCYTE_DEVELOPMENT	21	-2.39	0.000	0.003	tags = 57%, list = 13%, signal = 66%
GO_CHONDROCYTE_DIFFERENTIATION	57	-1.57	0.006	0.135	tags = 33%, list = 13%, signal = 38%
GO_POSITIVE_REGULATION_ OF_OSTEOBLAST_DIFFERENTIATION	58	-1.91	0.000	0.047	tags = 38%, list = 15%, signal = 45%
GO_REGULATION_OF_BONE_DEVELOPMENT	17	-2.23	0.000	0.010	tags = 65%, list = 16%, signal = 77%

NES: normalized enrichment score. Size represents the total number of genes in the gene set. NOM P value indicates the credibility of enrichment results. FDR q value represents the corrected P value of multiple hypothesis testing. Tags represents the proportion of core genes in the total number of genes in the gene set, while list represents the proportion of core genes in the total number of genes. For a gene set, when the number of core genes is the same as the total number of genes under the gene set, the signal value is the largest.

3. Construction of weighted gene correlation network analysis.

In this study, WGCNA was used to construct the gene correlation module associated with the sample trait. In total, 5879 genes were included for analysis, and the soft threshold β was calculated before construction the weighted co-expression network. We set the correlation coefficient to 0.9 as screening criteria and calculated the value of soft threshold β as 6 (Fig. 3A). Then a total of 13 gene modules with different colors were recognized by hierarchical clustering (Fig. 3B). Each module was assigned a unique color: black, blue, brown, green, greenyellow, magenta, pink, purple, red, salmon, tan, turquoise and yellow represented 231, 835, 472, 298, 84, 132, 134, 96, 257, 48, 72, 2829 and 326 genes, respectively. Of all genes, 65 were not assigned. We clustered modules based on gene expression and obtained a correlation heatmap between modules (Fig. 3C), and then transformed them into a topological overlap matrix (TOM) and visualized the system clustering tree of gene (Fig. 3D). Eigengenes was correlated with external traits

in order to search for the significant associated modules. It was clear that the MEgreen (298 genes) was most positive associated with OA (Fig. 3E).

4. Sub-modules in the PPI network and identification of core genes.

We detected densely linked regions in large PPI networks evaluated by STRING online tool that may represent molecular complexes using an MCODE analysis, and several significant modules were identified. The 3 modules with the highest score were selected (Fig. 4A ~ Fig. 4C). The study also calculated degree using NetworkAnalyzer, and degree > 30 was set as the thresholds. So far, the gene sets obtained by the analysis of MCODE, Degree, GSEA and WGCNA were put together and the common genes were detected (Fig. 4D). The results showed that golgi glycoprotein 1 (GLG1), secreted frizzled related protein 2 (SFRP2), secreted protein acidic and cysteine rich (SPARC), 3'-phosphoadenosine 5'-phosphosulfate synthase 2 (PAPSS2), vitamin K epoxide reductase complex subunit 1 (VKORC1) and cathepsin K (CTSK) were coexisted in GSEA and WGCNA gene sets, as well as TIMP metallopeptidase inhibitor 1 (TIMP1) and syndecan 1 (SDC1) were commonly contained in Degree and WGNCA gene sets, and C-C motif chemokine receptor 5 (CCR5) was covered by Degree, MCODE, and WGCNA gene sets. The CTD database showed that these genes targeted several musculoskeletal diseases and these data appear in Fig. 4E. ROC analysis showed that 5 genes such as GLG1, PAPSS2, CTSK, TIMP1 and SDC1 could serve as valuable biomarkers for distinguishing patients with OA from healthy controls (Fig. 4F, Table 2).

Table 2
Receiver operator characteristic analysis of 9 candidate genes.

Genes	AUC	Std. Error	P value
GLG1	0.8143	0.1116	0.0318
SFRP2	0.7429	0.1251	0.0971
SPARC	0.6286	0.1385	0.3798
PAPSS2	0.8571	0.09201	0.0147
VKORC1	0.5143	0.1495	0.9223
CTSK	0.8286	0.1018	0.0248
TIMP1	0.8429	0.09748	0.0192
SDC1	0.9571	0.04869	0.0018
CCR5	0.7857	0.1192	0.0510
AUC: area under curve. AUC value is the area covered by the ROC curve			

Discussion

OA is characterized by cartilage degeneration, subchondral bone changes and bone marrow lesions [19]. Although non-pharmacological interventions have made some breakthroughs in treatment of OA, improved treatment of non-invasively is lacking in the early or mid-stage OA phase [20]. One of the reasons is complexity of OA pathology and difficulty of early diagnosis. Bioinformatics analysis reveals the biological mechanisms form large and complex biological data through a combination of biology, computer science, and information technology. In the present study, the gene expression dataset GSE82107 was obtained from the GEO database to explore core genes.

We analyzed the gene expression matrix to observe the relationship between samples. If the quality assessment results show that the biological repeatability of the samples is very poor or there are many interlaced regions between samples and groups, this may lead to a decrease in the reliability of the bioinformatics analysis results, even if the analysis results are finally obtained. However, quite a few bioinformatics studies on OA have not performed quality assessments. Our results indicated that the repeatability of the samples in the same group is relatively stable, and there are also obvious biological differences between different groups, thus laying relatively credible reliability for subsequent analysis. Afterward, the study applied different calculation methods to deal with the gene expression matrix. By comparing the difference of gene expression between two groups and setting a threshold, 1676 DEGs were filtered out, including 218 upregulated genes and 1458 downregulated genes. Gene functional annotation enrichment analysis is a high-throughput research strategy that increases the likelihood for researchers to pinpoint the biological processes most relevant to them [21]. However, the differences in expression of most genes may not be significant, and focusing only on these genes with significant changes may leave out or ignore a lot of information. Therefore, we also applied GSEA to analyze the enrichment properties of the entire gene expression data in specific functional gene sets, and the results showed that some genes were enriched in bone development, bone remodeling, and cartilage development and differentiation. OA is associated with bone loss due to bone remodeling, and the slowing of bone turnover in the later stage of OA leads to the densification of the subchondral plate, which further hinders the signal exchange between cartilage and bone and inhibits the repair of damaged cartilage [22]. These genes included in the functional annotation sets may be involved in the pathological process of OA. In addition, PPI networks based on DEGs are important in most OA biological functions and processes, and most proteins appear to activate their function through their interactions [23], and WGCNA algorithm represents a new biological approach for detecting key genes associated with sample traits in gene co-expression networks among all datasets [24]. Besides, MCODE can also obtain interactive clusters by calculating the PPI network based on DEGs. In this study, the common genes in the gene set obtained by the above algorithm were searched as candidate core genes, including GLG1, SFRP2, SPARC, PAPSS2, VKORC1, CTSK, TIMP1, SDC1 and CCR5. Further, we explored the relationship between these genes and musculoskeletal diseases and found that all genes were involved in different musculoskeletal diseases such as osteoporosis, OA and arthritis. Moreover, SPARC, CTSK and PAPSS2 may be biomarkers of some diseases or play a role in the etiology of the diseases. However, when we carried out ROC analysis on these candidate genes, we found that only GLG1, PAPSS2, CTSK, TIMP1 and SDC1 could be served as potential biomarkers to distinguish patients with OA from healthy people.

GLG1 located in golgi apparatus belongs to the cysteine-rich fibroblast growth factor receptor family [25, 26]. GLG1 is expressed widely, mainly in skeletal muscle, placenta, bone marrow, ovary and testis [27]. It may have a chaperone function that participates in the processing and targeting of growth factors in cells [28]. Bone morphogenetic proteins 4 (BMP4) promotes the healing of fractures by stimulating the synthesis of extracellular matrix in chondrocytes and BMP4 is reduced in patients with OA and rheumatoid arthritis [29]. Studies have shown that silencing GLG1 induces hormone transcription, suggesting that GLG1 may be a negative regulator of pituitary hormone transcription. And GLG1 has a negative impact on the expression of BMP4 [30]. Sulfation is a ubiquitous modification of exogenous and endogenous compounds. In mammals, PAPS is the only source of sulfate, which is produced by ATP and inorganic sulfate. PAPSS2 encodes PAPS synthase. It is specifically expressed in cartilage and adrenal gland and may play an important role in bone development during postnatal growth [31, 32]. The researchers found that the lack of PAPSS2 activity lead to degenerative knee joint disease in mice, while people who lacked normal PAPSS2 activity showed shortening and bending of long bones, as well as degenerative arthropathy [33]. In addition, PAPSS2 may be a candidate gene transactivated by sex-determining region Y-box containing gene 9 (SOX9), and thus it may act as a regulator of SOX9 expression in cartilage [34]. The protein encoded by CTSK is lysosomal cysteine protease, which is a member of the peptidase C1 protein family and is mainly expressed in osteoclasts. CTSK is closely involved in bone remodeling and bone resorption [35, 36]. Researchers performed a study on OA for four phenotypes, including hip OA, knee OA, knee and/or hip OA, and any OA. A genome-wide association study (GWAS) found that CTSK is a possible OA effector gene, which has mechanisms to support the evaluation of the efficacy of OA [37]. And this is similar to our findings. Additionally, experimental studies have confirmed that $CTSK^{-/-}$ mice exhibit a reduction in the remodeling of subchondral bone and calcified cartilage in the destabilization of the medial meniscus induced OA [38]. TIMP1 belongs to TIMP gene family, which encodes a protein that is a natural inhibitor of matrix metalloproteinase (MMP) involved in the pathology of OA [39]. The inhibitor works by forming a one-to-one complex with target metalloproteinases, such as collagenase [40]. By comparing the gene expression of human OA synovial fibroblasts stimulated by transforming growth factor β, experimental OA mouse synovium and end-stage OA human synovium, the results showed that the expression of TIMP1 was up-regulated under all these conditions [41]. Besides, clinical studies have found that TIMP1 level is associated with at least one radiographic grading in the evaluation of generalized OA [42]. SDC1 encodes a cell surface proteoglycan that is a member of the syndecan proteoglycan family. Syndecans mediate cytoskeletal organization, cell signaling and cell binding [43]. There is evidence that SDC1 expression changes in the meniscus in the early stages of OA, possibly because the pathological changes of the meniscus precede the lesions of cartilage [44]. Although there is relatively rare research on SDC1 in OA, it is related to the cytoskeleton that affects the shape of cells. According to the role of SDC1 in decentralized mechanical transmission [45], we speculate that it may be involved in the physiological or pathological process of OA by regulating the bone mechanostat by affecting the tensegrity of the cytoskeleton.

Conclusions

The hub-genes of GLG1, PAPSS2, CTSK, TIMP1 and SDC1 may be potentially valuable biomarkers or therapeutic targets for OA. However, in order to determine the precise role of these genes, further experimental verification is needed.

Declarations

Acknowledgements

Not applicable.

Funding

The National Natural Science Foundation of China (grant no. 81573874 and grant no. 81873229) provided financial support for the conduct of the study.

Availability of data and materials

The datasets generated and analyzed during the current study are available in the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/) [8].

Authors' contributions

YC (chinachaiyi@163.com), CYM (43399581@qq.com) and QLF (tcmscience@163.com) designed the study. YC, MYZ (2138300980@qq.com), DMX (3569609016@qq.com), FT (528518817@qq.com) and XL (707303248@qq.com) analyzed and interpreted the data. YC and QLF were major contributors in the writing of the manuscript. All authors read and approved the submitted manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Footnote

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Bijlsma JW, Berenbaum F, Lafeber FP: Osteoarthritis: an update with relevance for clinical practice. Lancet. 2011, 377(9783):2115-2126.

2. Glyn-Jones S, Palmer A, Agricola R, Price A, Vincent T, Weinans H, Carr A: Osteoarthritis. Lancet. 2015, 386(9991):376-387.
3. Fransen M, Bridgett L, March L, Hoy D, Penserga E, Brooks P: The epidemiology of osteoarthritis in Asia. Int J Rheum Dis. 2011, 14(2):113-121.
4. ST S, EM R, MB L, MS R, L A-N, O S, medicine RSJTNEjo: A Randomized, Controlled Trial of Total Knee Replacement. N Engl J Med. 2015, 373(17):1597-1606.
5. Oliver GR, Hart SN, Klee EW, J Clinical Chemistry: Bioinformatics for clinical next generation sequencing. Clin Chem. 2015, 61(1):124-135.
6. J W, V B, S N, G R, G R, R G, T M, R B, F R, research GBJNa: Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. Nucleic Acids Res. 2018, 46(null):W11-W16.
7. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008, 9:559.
8. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M: NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012, 41(D1):D991-D995.
9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: Gene Ontology: tool for the unification of biology. Nat Genet. 2000, 25(1):25-29.
10. Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000, 28(1):27-30.
11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA. 2005, 102(43):15545-15550.
12. Bader G, Hogue C: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003, 4:2.
13. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ: The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res. 2017, 45(D1):D972-d978.
14. Broeren MG, de Vries M, Bennink MB, van Lent PL, van der Kraan PM, Koenders MI, Thurlings RM, van de Loo FA: Functional Tissue Analysis Reveals Successful Cryopreservation of Human Osteoarthritic Synovium. PLoS ONE. 2016, 11(11):e0167076.
15. G KM: A new measure of rank correlation. Biometrika. 1938, 30(1/2):81-93.
16. Huang DW, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009, 4(1):44-57.
17. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP: STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2014, 43(D1):D447-D452.

18. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003, 13(11):2498-2504.
19. Valdes AM, Spector TD: The genetic epidemiology of osteoarthritis. Curr Opin Rheumatol. 2010, 22(2):139-143.
20. Hügle T, Geurts J: What drives osteoarthritis?-synovial versus subchondral bone pathology. Rheumatology. 2017, 56(9):1461-1471.
21. Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009, 37(1):1-13.
22. Burr DB, Gallant MA: Bone remodelling in osteoarthritis. Nat Rev Rheumatol. 2012, 8(11):665.
23. Athanasios A, Charalampos V, Vasileios T, Ashraf GM: Protein-Protein Interaction (PPI) Network: Recent Advances in Drug Discovery. Curr Drug Metab. 2017, 18(1):5-10.
24. Giulietti M, Occhipinti G, Principato G, Piva F: Identification of candidate miRNA biomarkers for pancreatic ductal adenocarcinoma by weighted gene co-expression network analysis. Cell Oncol. 2017, 40(2):181-192.
25. Ahn J, Febbraio M, Silverstein RL: A novel isoform of human Golgi complex-localized glycoprotein-1 (also known as E-selectin ligand-1, MG-160 and cysteine-rich fibroblast growth factor receptor) targets differential subcellular localization. J Cell Sci. 2005, 118(Pt 8):1725-1731.
26. Croul S, Mezitis SG, Stieber A, Chen YJ, Gonatas JO, Goud B, Gonatas NK: Immunocytochemical visualization of the Golgi apparatus in several species, including human, and tissues with an antiserum against MG-160, a sialoglycoprotein of rat Golgi apparatus. J Histochem Cytochem. 1990, 38(7):957-963.
27. Olofsson A, Hellman U, Ten Dijke P, Grimsby S, Ichijo H, Morén A, Miyazono K, Heldin CH: Latent transforming growth factor-beta complex in Chinese hamster ovary cells contains the multifunctional cysteine-rich fibroblast growth factor receptor, also termed E-selectin-ligand or MG-160. Biochem J. 1997, 324 (Pt 2):427-434.
28. Köhl R, Antoine M, Olwin BB, Dickson C, Kiefer P: Cysteine-rich fibroblast growth factor receptor alters secretion and intracellular routing of fibroblast growth factor 3. J Biol Chem. 2000, 275(21):15741-15748.
29. Bramlage CP, Häupl T, Kaps C, Ungethüm U, Krenn V, Pruss A, Müller GA, Strutz F, Burmester GR: Decrease in expression of bone morphogenetic proteins 4 and 5 in synovial tissue of patients with osteoarthritis and rheumatoid arthritis. Arthritis Res Ther. 2006, 8(3):R58.
30. Ahmed K, LaPierre MP, Gasser E, Denzler R, Yang Y, Rülicke T, Kero J, Latreille M, Stoffel M: Loss of microRNA-7a2 induces hypogonadotropic hypogonadism and infertility. J Clin Invest. 2017, 127(3):1061-1074.
31. Noordam C, Dhir V, McNelis JC, Schlereth F, Hanley NA, Krone N, Smeitink JA, Smeets R, Sweep FC, Claahsen-van der Grinten HL et al: Inactivating PAPSS2 mutations in a patient with premature pubarche. N Engl J Med. 2009, 360(22):2310-2318.

32. Strott CA: Sulfonation and molecular action. Endocr Rev. 2002, 23(5):703-732.
33. Ford-Hutchinson AF, Ali Z, Seerattan RA, Cooper DM, Hallgrímsson B, Salo PT, Jirik FR: Degenerative knee joint disease in mice lacking 3'-phosphoadenosine 5'-phosphosulfate synthetase 2 (Papss2) activity: a putative model of human PAPSS2 deficiency-associated arthrosis. Osteoarthr Cartil. 2005, 13(5):418-425.
34. C S, A B, P H, B Z, UH D: Expression profile of Papss2 (3'-phosphoadenosine 5'-phosphosulfate synthase 2) during cartilage formation and skeletal development in the mouse embryo. Dev Dyn. 2007, 236(5):1313-1318.
35. An J, Hao D, Zhang Q, Chen B, Zhang R, Wang Y, Yang H: Natural products for treatment of bone erosive diseases: The effects and mechanisms on inhibiting osteoclastogenesis and bone resorption. Int Immunopharmacol. 2016, 36:118-131.
36. Gelb BD, Shi GP, Chapman HA, Desnick RJ: Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. Science. 1996, 273(5279):1236-1238.
37. Tachmazidou I, Hatzikotoulas K, Southam L, Esparza-Gordillo J, Haberland V, Zheng J, Johnson T, Koprulu M, Zengini E, Steinberg J et al: Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. Nat Genet. 2019, 51(2):230-236.
38. Soki FN, Yoshida R, Paglia DN, Duong LT, Hansen MF, Drissi H: Articular cartilage protection in Ctsk mice is associated with cellular and molecular changes in subchondral bone and cartilage matrix. J Cell Physiol. 2018, 233(11):8666-8676.
39. Zeng GQ, Chen AB, Li W, Song JH, Gao CY: High MMP-1, MMP-2, and MMP-9 protein levels in osteoarthritis. Genet Mol Res. 2015, 14(4):14811-14822.
40. O'Shea M, Willenbrock F, Williamson RA, Cockett MI, Freedman RB, Reynolds JJ, Docherty AJ, Murphy G: Site-directed mutations that alter the inhibitory activity of the tissue inhibitor of metalloproteinases-1: importance of the N-terminal region between cysteine 3 and cysteine 13. Biochemistry. 1992, 31(42):10146-10152.
41. Remst DF, Blom AB, Vitters EL, Bank RA, van den Berg WB, Blaney Davidson EN, van der Kraan PM: Gene expression analysis of murine and human osteoarthritis synovium reveals elevation of transforming growth factor β-responsive genes in osteoarthritis-related fibrosis. Arthritis Rheumatol. 2014, 66(3):647-656.
42. Takahashi M, Naito K, Abe M, Sawada T, Nagano A: Relationship between radiographic grading of osteoarthritis and the biochemical markers for arthritis in knee osteoarthritis. Arthritis Res Ther. 2004, 6(3):R208-212.
43. Baietti MF, Zhang Z, Mortier E, Melchior A, Degeest G, Geeraerts A, Ivarsson Y, Depoortere F, Coomans C, Vermeiren E et al: Syndecan-syntenin-ALIX regulates the biogenesis of exosomes. Nat Cell Biol. 2012, 14(7):677-685.
44. Endo J, Sasho T, Akagi R, Muramatsu Y, Watanabe A, Akatsu Y, Fukawa T, Tahara M, Yamaguchi S: Comparative Analysis of Gene Expression between Cartilage and Menisci in Early-Phase Osteoarthritis of the Knee-An Animal Model Study. J Knee Surg. 2018, 31(7):664-669.

45. Ebong EE, Lopez-Quintero SV, Rizzo V, Spray DC, Tarbell JM: Shear-induced endothelial NOS activation and remodeling via heparan sulfate, glypican-1, and syndecan-1. Integr Biol. 2014, 6(3):338-347.

Supplemental Information

Supplemental Figure 1. Gene Ontology analysis of differentially expressed genes associated with OA. The top 5 terms were selected according to P-value. (A), (B) and (C) are the enrichment results for up-regulated genes. (D), (E) and (F) are the enrichment results for down-regulated genes. The GO biological process terms for biological process, cellular component and molecular function are represented in (A) and (D), (B) and (E), (C) and (F) respectively. Set size means the number of genes contained in each term, and intersection size represents the number of genes shared in each term.

Supplemental Figure 2. Kyoto Encyclopedia of Genes and Genomes pathway analysis of differentially expressed genes associated with OA. The top 5 terms were selected according to P-value. (A) is the enrichment results for up-regulated genes, and (B) for down-regulated genes. Set size means the number of genes contained in each term, and intersection size represents the number of genes shared in each term.

Supplemental Table 1. Gene Ontology analysis of differentially expressed genes associated with osteoarthritis. The top 5 terms were selected according to *P* value. Count is the number of enriched genes for each term. GO, Gene Ontology; BP, biological process; MF, molecular function; CC, cellular component. % represents the percentage of genes contained in each term to the upregulated or downregulated genes.

Supplemental Table 2. Kyoto Encyclopedia of Genes and Genomes pathway analysis of DEGs associated with osteoarthritis. The top 5 terms were selected for upregulated and downregulated DEGs according to *P* values. Count is the number of enriched genes for each term. DEGs, differentially expressed genes. % represents the percentage of genes contained in each term to the upregulated or downregulated genes.

Figures

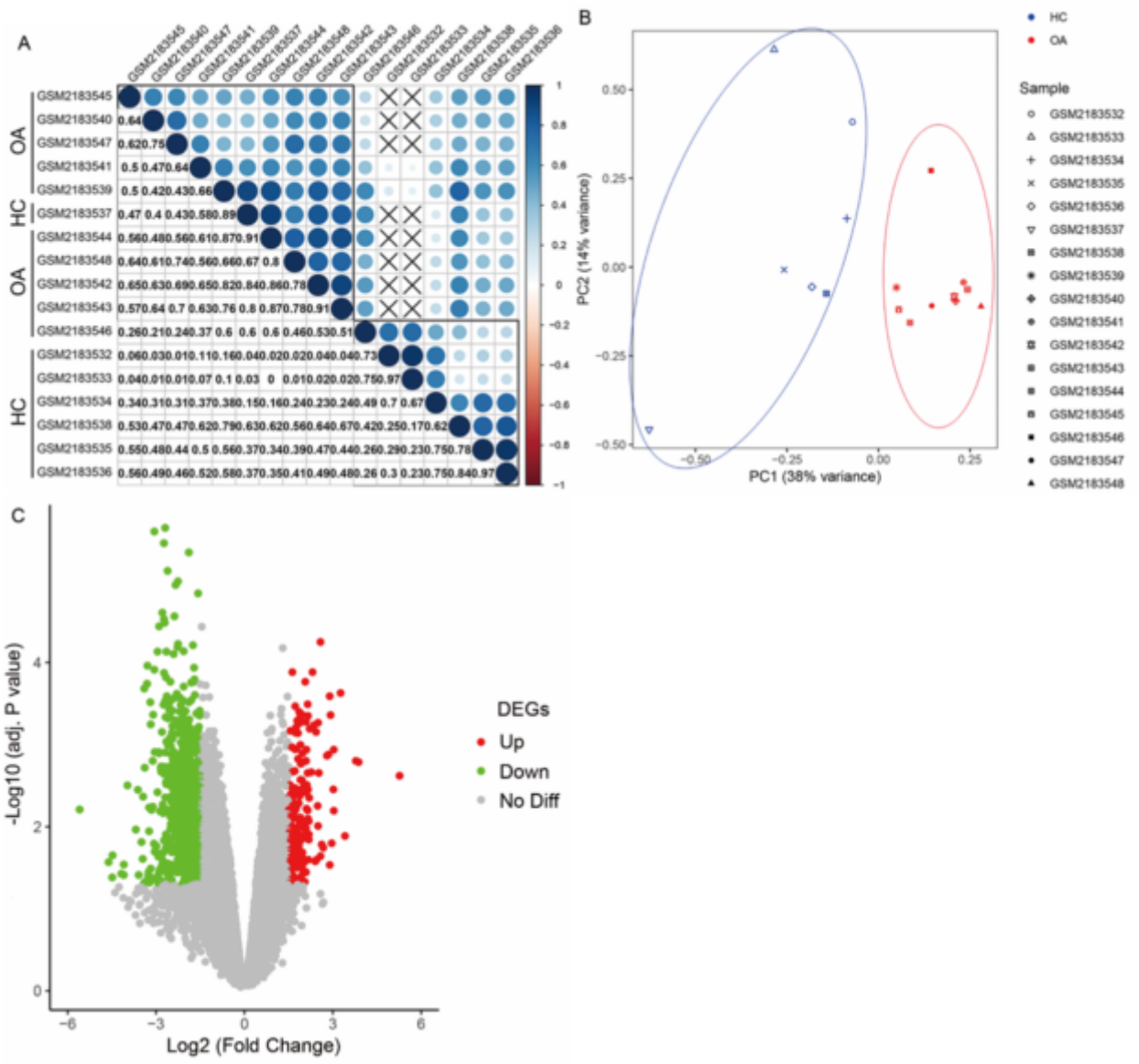

Figure 1

Quality evaluation of gene expression matrix and visualization of DEGs. A. Correlation plot of samples in OA vs. HC. The black broken line in the figure represents the classification result using hierarchical clustering calculation, and × represents the result P > 0.01. B. Principal component plot, OA vs. HC samples. C. DEGs between OA and HC samples in the dataset GSE82107. The red points represent upregulated genes and the green ones represent downregulated genes screened on the basis of |logFC| > 1.5 and P < 0.05. The gray points represent genes with no significant difference in expression.

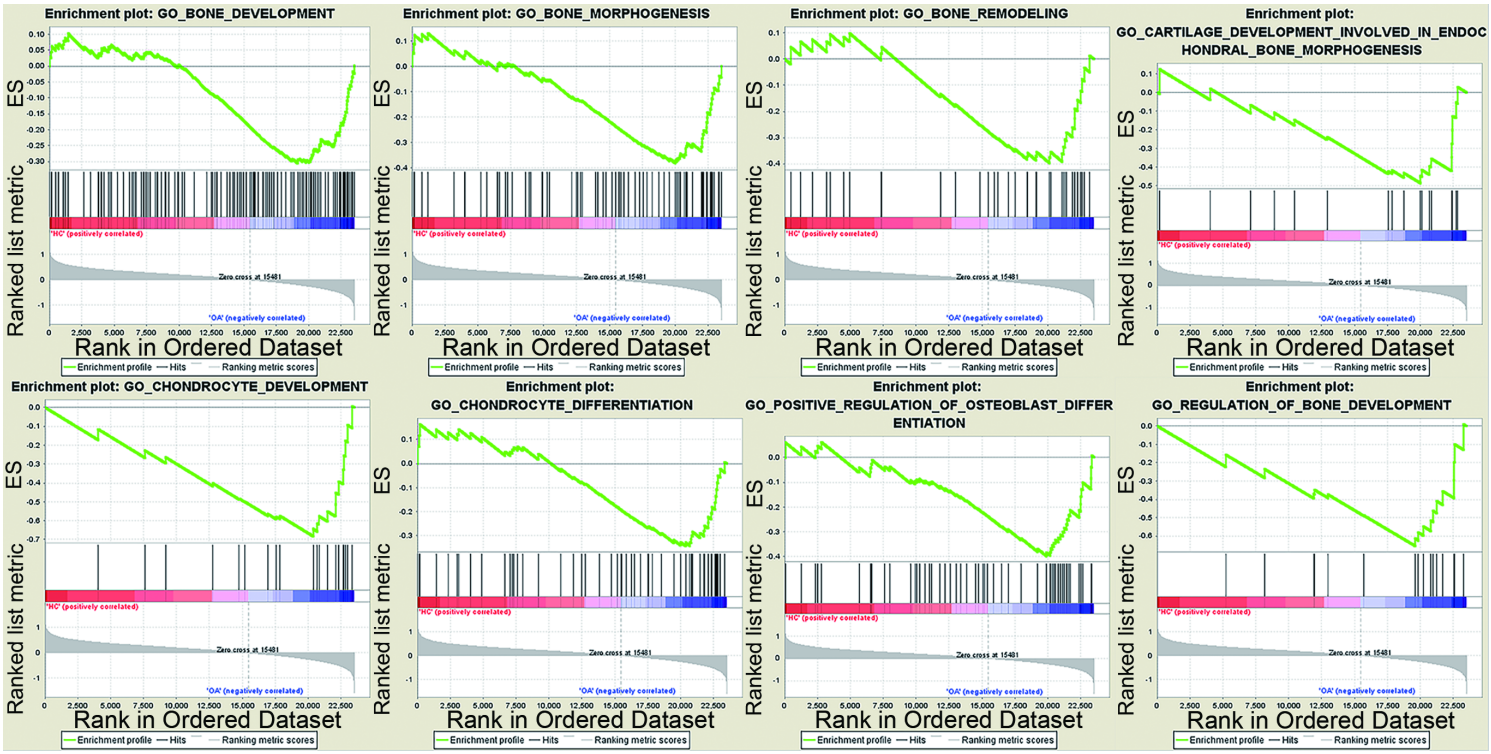

Figure 2

GSEA between OA and HC. ES: enrichment score.

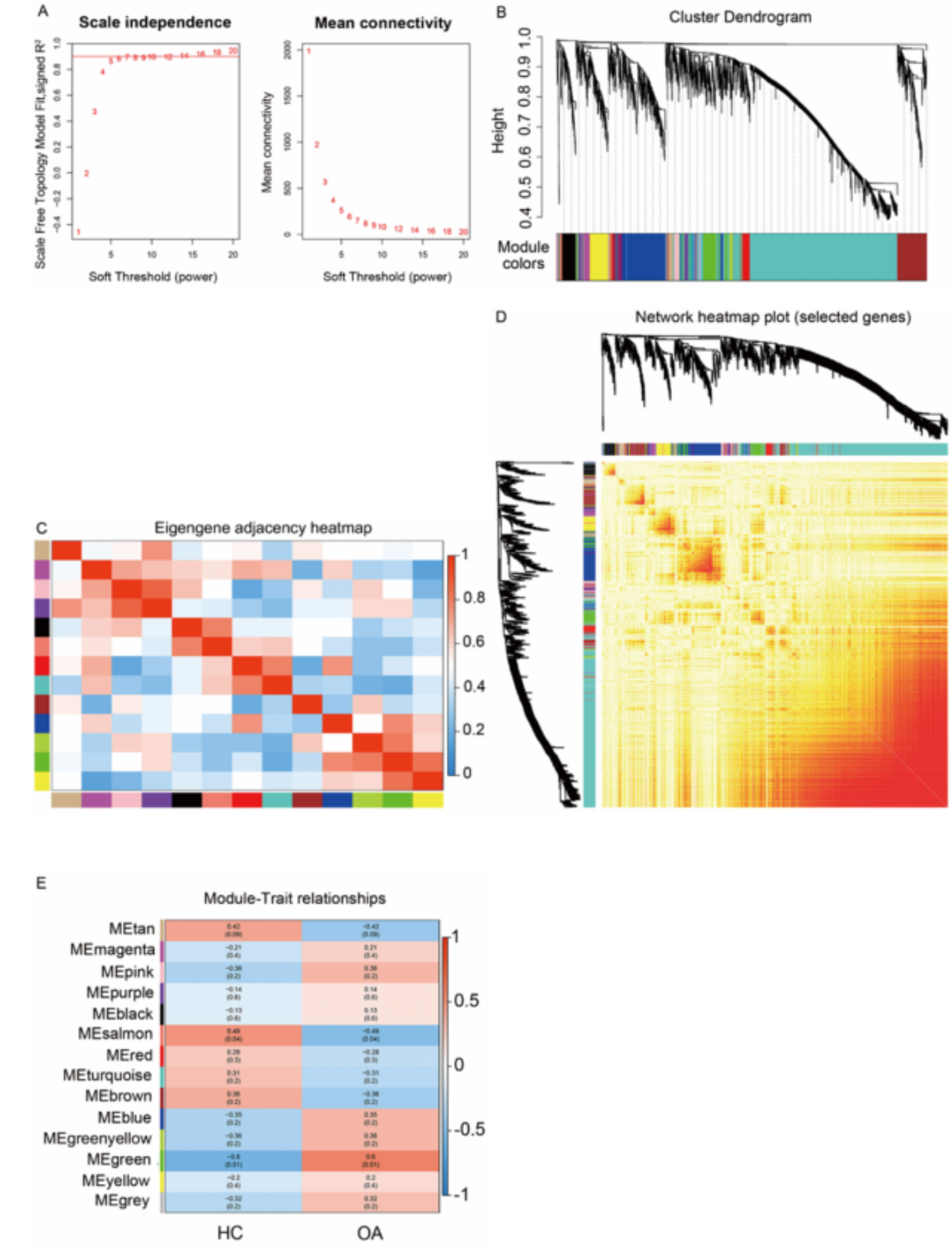

Figure 3

Weighted gene correlation network analysis. A. Determination of parameter β of the adjacency function in the WGCNA algorithm. The left panel shows the scale free topology model fit index as a function of the soft-thresholding power. The red line in the left panel indicates R2 = 0.9. There is a trade-off between maximizing the scale-free topology model fitting index and maintaining a high mean number of connections. Thus, we set β = 6. The right panel displays the mean connectivity. B. Clustering

dendrogram of genes. Gene clustering tree (dendrogram) obtained by hierarchical clustering of adjacency-based dissimilarity. The colored row below the dendrogram indicates module membership identified by the dynamic tree cut method. C. Relationships among modules based on modules' eigengenes. Each row and column in the heatmap correspond to one module eigengene (labeled by color). In the heatmap, red represents high adjacency, while blue color represents low adjacency. D. Heatmap plot of topological overlap in the gene network. Each color represents a certain gene module. Light color denotes low topological overlap, and progressively darker red denotes higher topological overlap. Darker squares along the diagonal correspond to modules. The gene dendrogram and module assignment are shown along the left and top. E. Module-feature associations. Each row corresponds to a module Eigengene and each column to a clinical feature. Each cell contains the corresponding correlation in the first line and the P value in the second line. The table is color-coded by correlation according to the color legend.

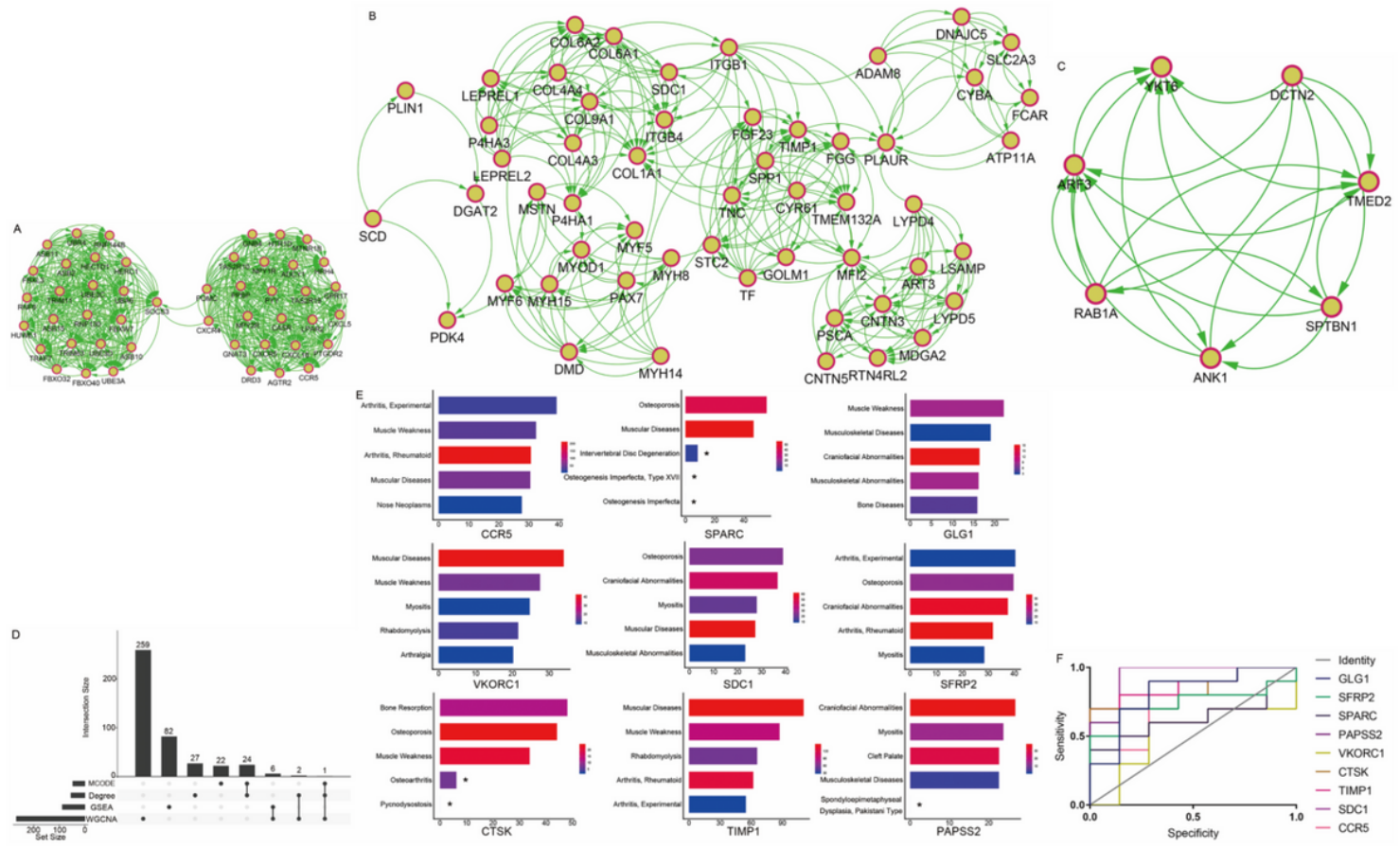

Figure 4

Top 3 modules from the PPI network and identification of core genes. A. Module-A. Module A (Score: 23.087) has 47 nodes and 531 edges. B Module-B. Module B (Score: 9.692) has 53 nodes and 252 edges C. Module-C. Module C (Score: 7) has 7 nodes and 21 edges. D. Overlap of gene sets based on the four research methods. The horizontal bars in the lower left corner shows the number of genes contained in each gene set. In the vertical bar, the first 4 represent the number of genes belonging only to the group, and the last 4 indicate the number of genes shared with each other. And black dots indicate the gene set

has been selected. E. Relationship to musculoskeletal diseases related to candidate genes based on the CTD database. *Direct evidence of marker or mechanism in this disease. F. ROC curves of 9 genes for OA diagnosis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigure1A.tif
- SupplementalFigure2A.tif
- SupplementalFigure2B.tif
- SupplementalFigure1F.tif
- SupplementalFigure1D.tif
- SupplementalFigure1E.tif
- SupplementalFigure1C.tif
- SupplementalTable2.docx
- SupplementalFigure1B.tif
- SupplementalTable1.docx