Development and Validation of Diagnostic Models for Transcriptomic Signature Genes for Multiple Tissues in Osteoarthritis

Qichang Gao¹, Yiming Ma¹, Tuo Shao¹, Xiaoxuan Tao², Xiansheng Yang¹, Song Li¹, Jiaao Gu¹, Zhange Yu¹

¹Department of Spinal Surgery, The 1st Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang Province, People's Republic of China; ²Department of Radiotherapy, The 3st Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang Province, People's Republic of China

Correspondence: Zhange Yu, Email yuzhange1967@163.com

Background: Progress in research on expression profiles in osteoarthritis (OA) has been limited to individual tissues within the joint, such as the synovium, cartilage, or meniscus. This study aimed to comprehensively analyze the common gene expression characteristics of various structures in OA and construct a diagnostic model.

Methods: Three datasets were selected: synovium, meniscus, and knee joint cartilage. Modular clustering and differential analysis of genes were used for further functional analyses and the construction of protein networks. Signature genes with the highest diagnostic potential were identified and verified using external gene datasets. The expression of these genes was validated in clinical samples by Real-time (RT)-qPCR and immunohistochemistry (IHC) staining. This study investigated the status of immune cells in OA by examining their infiltration.

Results: The merged OA dataset included 438 DEGs clustered into seven modules using WGCNA. The intersection of these DEGs with WGCNA modules identified 190 genes. Using Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest algorithms, nine signature genes were identified (CDADC1, PPFIBP1, ENO2, NOM1, SLC25A14, METTL2A, LINC01089, L3HYPDH, NPHP3), each demonstrating substantial diagnostic potential (areas under the curve from 0.701 to 0.925). Furthermore, dysregulation of various immune cells has also been observed.

Conclusion: CDADC1, PPFIBP1, ENO2, NOM1, SLC25A14, METTL2A, LINC01089, L3HYPDH, NPHP3 demonstrated significant diagnostic efficacy in OA and are involved in immune cell infiltration.

Keywords: osteoarthritis, machine learning, immune cells infiltration, diagnostic model

Introduction

Osteoarthritis (OA) is a degenerative joint disease that is characterized by inflammation and pain.¹ OA not only considerably diminishes life satisfaction for patients, but also increases the substantial financial burden on healthcare systems.² Among the joints most frequently affected by OA, the knee joint encompasses various structural components including the meniscus, synovium, and articular cartilage.³

Meniscal damage marks an early pivotal event in knee OA development, exacerbating the load on the articular cartilage and potentially initiating inflammatory responses that hasten cartilage deterioration.⁴ Inflammation and hypertrophy of the synovium play vital roles in OA, and synovitis is recognized as a major contributor to OA-associated pain and its progression.⁵ Gradual wear and degradation of articular cartilage are essential pathological manifestations of OA, diminishing the cushioning ability of the joint, which results in pain and restricted mobility.⁶ The treatment of OA includes nonsurgical approaches (health management, medication) and surgical interventions. With the advancement in the research of OA's molecular mechanisms and the maturation of pluripotent stem cell research, new minimally invasive treatment methods have gradually increased. Although the long-term benefits are still inconclusive, intra-articular therapies such as hyaluronic acid injection and articular cartilage repair are becoming more popular. However, these treatments are often associated with high costs and limited efficacy. Therefore, the development of diagnostic models for OA can provide valuable insights into the pathogenesis of the disease and facilitate the discovery of new therapeutic targets.
acid (HA), platelet-rich plasma (PRP), and mesenchymal stem cells have become effective methods for early to mid-stage OA. Moreover, with the development of big data machine learning analysis models based on disease genomics and molecular phenotypes, novel diagnostic methods, distinct from traditional demographic and clinical imaging approaches, have demonstrated promising predictive capabilities and are being utilized in clinical practice.\(^7,8\)

Recent research has increasingly illuminated the critical role of the immune system in OA development, particularly the significant impact of inflammatory responses in promoting damage to joint cartilage and accelerating disease progression.\(^9\) However, research on noninvasive biomarkers for OA diagnosis is challenging.\(^10\)

In this study, we collectively analyzed the meniscus, synovium, and articular cartilage for the first time using various bioinformatics approaches to identify the distinctive genes associated with knee osteoarthritis. These genes have shown exceptional diagnostic efficacy as confirmed using external datasets and clinical tissue samples. We also explored the signaling pathways in which these genes are involved and their role in immune cell infiltration, aiming to deepen our understanding of the interplay between meniscal damage, synovial inflammation, and cartilage degeneration, while providing references for clinical diagnosis and treatment.

**Methods and Materials**

**Clinical Samples**

We collected six osteoarthritis (OA) tissue samples from patients who underwent total knee arthroplasty at the 1st Affiliated Hospital of Harbin Medical University and six normal tissue samples from patients who underwent reconstructive surgery due to trauma.

**Microarray Data**

Figure 1 illustrates the study. In this study, we downloaded three datasets, GSE206848, GSE98918, and GSE178557 (http://www.ncbi.nlm.nih.gov/geo/).\(^11-13\) GSE206848 and GSE98918 (Including 19 healthy control groups and 19 disease groups) were designated as the training sets, whereas GSE178557 (Including 4 healthy control groups and 4 disease groups) was used as the validation dataset. The detailed characteristics of these datasets, including microarray platforms, composition of sample groups, and sample sizes (Table 1).

**Screening of DEGs**

Initially, the training datasets were merged using R software’s insilomerging package.\(^14\) Subsequently, the method proposed by Johnson et al was used to eliminate batch effects.\(^15\) The limma package of R software (version 4.1.2) was used for differential analysis.\(^16\)

**Co-Expression Network Analysis**

GSE206848 and GSE98918 datasets were processed using WGCNA.\(^17\) Briefly, the genes that exhibited the lowest 50% of the median absolute deviation values were discarded. The goodSamplesGenes method was used to exclude outlier genes and samples, thereby facilitating the construction of a scale-free network. The modules were subsequently delineated using dynamic tree cutting. WGCNA analysis was used to identify significant modules in OA.

**Exploration of Functions**

Gene profiling and visualization were conducted using the clusterProfiler package.\(^18\) Gene Ontology (GO) was used to explore the biological pathways involved in the DEGs.\(^19\) Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was used to analyze potential biological functions.\(^20\) Pathway association analysis of the characteristic genes was conducted using GSEA software. Significance was assessed based on the criterion of an adjusted p < 0.05.\(^21\)

**Protein Network**

To further investigate the interactions between DEGs from OA patients, a protein network was analyzed using STRING (https://string-db.org/). Subsequently, the network was visualized using the CentiScape plugin in Cytoscape3.10.0.
Screening OA Diagnostic Genes

The LASSO and Random Forest algorithms were used to analyze the characteristic genes associated with osteoarthritis (OA). Data integration was performed using R software’s glmnet package. Subsequently, the randomForest package of R software was used to screen hub genes. The overlapping genes identified by both the machine learning models were used as characteristic genes for further analysis.

Figure 1 Designing of this research.
Immune Infiltration Analysis
The CIBERSORT algorithm was utilized to analyze correlations among immune cells, with the goal of investigating differences in immune infiltration in osteoarthritis (OA) patients and healthy patients. Additionally, the correlation between characteristic genes and immunity was assessed using Spearman’s method.24

RT-qPCR
Total RNA was extracted by the centrifuge column method (DP451; Invitrogen). An automatic medical PCR analysis system (SLAN-96S, China) and SYBR Green qPCR Mix (G3326-01; Servicebio) were used for RT-qPCR analysis. The mRNA expression levels were statistically analyzed using the housekeeping genes GAPDH. All experiments were repeated three times and significant differences were statistically analyzed using the normalized $2^{-\Delta\Delta \text{ct}}$. Primers were designed as follows.

SLC25A14: 5′-GCCCTGCCACCTTCTACTC-3′, 5′-AACCACCCACCTTCTTCCATC-3′; NOM1: 5′-GGAGGCAGTGCTGGAGGAAGTTC-3′; 5′-ATAAATGGGCAATGACGGTGAAC-3′; METTL2A1: 5′-ACGACCTGTGATGACGGGAGAAGAGAAG-3′, 5′-AGAAGCCTGCTAGCTGTTG-3′; NPHP3: 5′-CCTCTGCCTTGTCAATGTTAGTCAC-3′, 5′-AAGCAAGCCACCATCCA-3′; ENO2: 5′-CGCACTTTCCACTTCTTTCTTC-3′, 5′-TCTCTAACACCTCAGCACACCAAC-3′; PPFIBP1: 5′-TTGACCGTACCCCTGAGGAGAAGG-3′; 5′-CAATAGACAAATCCGCTGCTGTTG-3′; GAPDH: 5′-TGACATCAAGAAGGTGAGTGAAGCAG-3′, 5′-GTGTCGCTGTTGAAGTCAAGAGGAG-3′.

Immunohistochemistry Staining
The expression levels of PPFIBP1, ENO2, NPHP3, and METTL2A (13961-1-AP, 10149-1-AP, 22026-1-AP, 16983-1-AP; Proteintech) in both OA and normal tissues were confirmed using immunohistochemical staining. Briefly, tissues were initially fixed for 24 h, followed by stepwise dehydration using alcohol gradients, paraffin embedding, and sectioning at a thickness of 4 mm. Antigen retrieval was conducted and endogenous peroxidases were removed using 3% H2O2. Sections were subsequently blocked with goat serum (AR1009; Boster). Primary and secondary antibodies were then incubated sequentially. DAB (ZLI-9018, Origene) was used for chromogenic detection, and hematoxylin (G1004, Servicebio) was used for nuclear staining.

Statistical Analysis
All experiments were performed at least three times, with data presented as mean ± SD. Data were analyzed using GraphPad Prism software version 9.4.0. Comparisons between two groups were conducted using Student’s $t$-test, while multiple group comparisons were performed using ANOVA. *$p < 0.05$ was considered statistically significant.

Results
Identifying DEGs Between OA and Control
Following the removal of batch effects, the GSE206848 and GSE98918 datasets were merged (Supplementary Figure S1). Through differential analysis, we obtained all differentially expressed genes (Supplementary Table S1), including 123 upregulated and 315 downregulated genes (Figure 2a and b).
Functional Enrichment Analysis

KEGG analysis revealed that the primary enriched pathways were “Adherens junction”, “Spliceosome and RNA degradation” (Figure 3a). GO analysis indicated that the main enriched terms in BP were “RNA splicing” and “positive regulation of DNA metabolic processes”. For CC, the primary enrichments were observed in “nuclear speckle”, “chromosomal region” and “ribonucleoprotein granule”. MF) analysis showed that “catalytic activity, acting on RNA” and “protein serine kinase activity” were the most significant features (Figure 3b).

Gene Modularity Analysis

We utilized the R software’s WGCNA package for module classification, setting the soft-thresholding power at 7 for relatively better average connectivity (Figure 4a and b). Subsequently, eight clustered modules were generated and are represented in distinct colors (Figure 4c and d). The blue module, consisting of 580 genes, was significantly correlated with OA (correlation coefficient = 0.59, p = 4.2e-52) (Figure 4e and f).
Analysis of Intersection Between WCGNA and OA DEGs and PPI Network Construction

A total of 190 genes were screened for DEGs and key modules. (Figure 5a). KEGG analysis revealed that the primary enriched pathways were “Spliceosome”, “Ribosome biogenesis in eukaryotes” (Figure 5b). GO analysis indicated that the main enriched terms in BP were “organelle organization” and “the cell cycle”. For CC, the primary enrichments were observed in “nuclear part” and “nuclear lumen”. MF analysis showed that “nucleic acid binding” was more prominent (Figure 5c–e). Subsequently, we constructed a protein network using the intersecting genes, and visualized the most active modules using the MCODE plugin (Figure 5f and g).

Identification of OA Signature Genes

The diagnostic model was constructed using two machine-learning algorithms. The LASSO regression algorithm identified 13 potential biomarkers (Figure 6a and b), whereas the random forest analysis identified 50 relatively important signature genes (Figure 6c and d). The identified signature genes are listed in Table 2. The convergence of these two algorithms identified nine signature genes: CDADC1, PPFIBP1, ENO2, NOM1, SLC25A14, METTL2A, LINC01089, L3HYPDH, and NPHP3 (Figure 6e).

Diagnostic Value Assessment

The identified genes were expressed at low levels in OA patients. (Figure 7a). The AUC values of the ROC analysis were 0.911 for CDADC1, 0.873 for ENO2, 0.925 for L3HYPDH, 0.853 for LINC01089, 0.701 for METTL2A, 0.781 for NPHP3, 0.783 for NOM1, 0.852 for SLC25A14, and 0.917 for PPFIBP1 (Figure 7b). Furthermore, the accuracy of each signature gene in predicting OA was evaluated using the external dataset GSE178557. The outcomes were in alignment with the trends observed in the GSE206848 and GSE98918 datasets (Figure 7c), demonstrating that these nine genes...
exhibited significant diagnostic efficiency in predicting OA. We extracted total RNA from three groups of clinical OA patients and healthy tissues. The RT-qPCR results were consistent with the expression predicted by the model. (Figure 7d).

The top ten pathways for each gene are shown in Figure 8. For instance, the expression of ENO2 is significantly correlated with RNA degradation, the cell cycle, and the Wnt signaling pathway. The expression of NOM1 correlated with the B-cell receptor pathway. The expression of PPFIBP1 significantly correlated with the cell cycle and olfactory transduction.

Figure 4 Module genes using WGCNA in OA.
Notes: (a and b) C To pursue the smoothness of the curve, $\beta = 7$ was set as the standard. (c) Different gene modules were indicated in different colors. (d) Heatmap of eigengene adjacency. (e and f). The blue module was most associated with OA.
Immune Infiltration Analysis in OA

The immunological characteristics of the patients were illustrated in Figure 9a. Notably, OA patients exhibited higher levels of naive B cells and lower levels of activated NK cells (Figure 9b). We also found a positive correlation between resting NK cells and eosinophils, whereas resting dendritic cells were negatively correlated with plasma cells and memory B cells were negatively correlated with M2 macrophages (Figure 9c).

Expression of PPFBP1, ENO2, NPHP3, and METTL2A in OA Tissues

Figure 10 illustrates the pathological features of synovium in the OA group. Marked proliferation and enlargement of synovial cells, neovascularization, pannus formation, and infiltration of various inflammatory cells including lymphocytes, macrophages, and neutrophils were observed. The compilation of target genes was also reduced along with the loss of normal synovial tissue. Additionally, immunohistochemical (IHC) analysis revealed that gene expression in the synovium of the OA group was consistent with the trends observed in the database.
Discussion

OA is a degenerative orthopedic disease that significantly affects the quality of life. The diagnosis of OA is predominantly based on clinical manifestations and radiological studies; however, specific diagnostic biomarkers are lacking. In recent years, several studies have conducted high-throughput sequencing on OA's synovium, cartilage, and meniscus in an attempt to identify diagnostic biomarkers. However, no study has integrated these three components to explore the mechanisms underlying OA progression from a holistic perspective. The diagnostic value of this model for OA was assessed using several methods.

CDADC1 is expressed in multiple tissues of the human body, particularly in the heart, muscles and brain. GO analysis revealed that CDADC1 primarily participates in key biological processes, including the regulation of hydrolase activity, protein dimerization, cytidine deaminase activity, and DNA cytosine deamination. These biological functions are involved in OA development and development of OA. For example, hydrolases play a crucial role in joint cartilage self-

Table 2 Analysis Results of Machine Learning

<table>
<thead>
<tr>
<th>LASSO Analysis</th>
<th>Random Forest Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDADC1, CYTOR, HOAG1, JPX, LINC01089, LINC02076, LOC102724814, NOM1, NPHP3, POLG2, PTPN6, SAYS1, SLC25A14</td>
<td>CDADC1, CCAR1, PFIBP1, INTS6L, ZC3H11A, ENO2, ANKRD36, PAN2, TPP2, CENPC, MON2, SMIM7, NOM1, RAD50, ARL6IP6, SRSF2, NSMC44A, NBEAL1, ANKRD12, KLHL20, SLC25A14, PIGU, METTL2A, TYW3, KLB, SOS1, LINC01089, L3HYPDH, PRPF40A, NPHP3, ABHD13, AHII, ARID4B, ATG5, ATRX, BBIP1, BCKDHB, BRIX1, C5orf24, CASK, CCDC112, CCDC12, CEDY, CEBFP3, CEP120, CEP83, CEP95, CFAP36, CHORDC1, CLIC2</td>
</tr>
</tbody>
</table>
repair by limiting excessive chondrocyte apoptosis and modulating inflammatory mediators to alleviate inflammatory reactions within the joints.\textsuperscript{29,30} These studies suggest that CDADC1 may be a potential target for alleviating OA.

ENO2 plays a pivotal role in the development of neural tissues.\textsuperscript{31} GO analysis indicated that ENO2 is involved in glucose metabolism and glycolysis. Recent studies have revealed the pivotal role of glycolysis in regulating the energy metabolism and inflammatory responses in OA chondrocytes.\textsuperscript{32} Glycolysis is correlated with the production of inflammatory mediators and aberrant chondrocyte differentiation under hypoxic conditions.\textsuperscript{33} Therefore, interventions targeting ENO2 may offer new therapeutic strategies for alleviating OA symptoms and delaying disease progression.

L3HYPDH regulates dietary proteins and other proteins such as collagen IV.\textsuperscript{34} GO analysis revealed that L3HYPDH primarily participates in the hydrolase activity. GO annotations related to METTL2A (Methyltransferase 2A) include methyltransferase and tRNA (cytidine) methyltransferase activities.\textsuperscript{35} NOM1 (Nucleolar protein with MIF4G domain1) is involved in the biogenesis of ribosomal small subunits, while another diagnostic gene, SLC25A14, participates in ion flux across the mitochondrial membrane and the regulation of membrane potential.\textsuperscript{36,37} The functions of these genes are

**Figure 7** The model genes expression.

**Notes:** (a) Gene expression in the training set. (b) The diagnostic efficiency of genes in the training set was evaluated. (c) The diagnostic efficacy of GSE17857 gene was evaluated. (d) Relevant mRNA expression levels were obtained by comparing Ct values. n=3, *p < 0.05, **p < 0.01, ***p < 0.001.
closely and intricately associated with the induction of inflammation in articular cartilage and degradation of the extracellular matrix. Identifying the specific connections between these regulatory factors and the OA phenotype is of paramount importance for future work.\textsuperscript{38}

\textit{NPHP3} is primarily involved in biological pathways, including organelle biogenesis, maintenance, and cargo trafficking to the periciliary membrane, and can also act as a molecular switch within the Wnt signaling pathway.\textsuperscript{39} Abnormal activation of the Wnt pathway affects changes in the subchondral tissue and synovial inflammation.

\textit{PPFIBP} is crucial for regulating focal adhesion disassembly and mediating protein-protein interactions at synapses.\textsuperscript{41} The coordination of internal and external signal transduction is a key factor in the perception of stress transmission.\textsuperscript{42} The pervasive presence of \textit{PPFIBP1} in the synovium, meniscus, and articular cartilage of OA patients highlights its potential as a diagnostic biomarker.\textsuperscript{43} This indicates that \textit{PPFIBP1} may be instrumental in understanding the pathological processes of OA and may serve as a novel target for therapeutic interventions.

\textbf{Figure 8} Analysis of the model genes in OA.

\textsuperscript{38} Gao et al. Journal of Inflammation Research 2024:17https://doi.org/10.2147/JIR.S472118

\textsuperscript{39} Gao et al. Journal of Inflammation Research 2024:17https://doi.org/10.2147/JIR.S472118

\textsuperscript{40} Gao et al. Journal of Inflammation Research 2024:17https://doi.org/10.2147/JIR.S472118

\textsuperscript{41} Gao et al. Journal of Inflammation Research 2024:17https://doi.org/10.2147/JIR.S472118

\textsuperscript{42} Gao et al. Journal of Inflammation Research 2024:17https://doi.org/10.2147/JIR.S472118

\textsuperscript{43} Gao et al. Journal of Inflammation Research 2024:17https://doi.org/10.2147/JIR.S472118
Naïve B cells, resting NK cells, and M0 macrophages play distinct roles in OA pathophysiological progression. Naïve B cells differentiate into plasma cells that produce antibodies that target self-antigens within the joint. This activity promotes inflammation and accelerates the tissue degradation. Resting NK cells play a dual role in OA. They can sometimes exacerbate local inflammation and, in special cases, promote knee joint repair through the secretion of growth factors. M1 macrophages drive inflammation and tissue damage, whereas M2 macrophages facilitate resolution of inflammation and tissue repair. The progression of OA may lead to a shift in balance, thereby influencing disease outcomes.

Our study has certain limitations. First, the data were sourced from public databases with a relatively small sample size, which might introduce selection bias and impact the diagnostic value. However, the reliability of our analysis was confirmed by using external datasets and immunohistochemistry. Secondly, larger clinical sample sizes and follow-up studies are required to validate our findings. Third, further investigation is required to elucidate the role of hub genes in immune response.
Conclusion

This study screened out Nine signature genes (CDADC1, PPFIBP1, ENO2, NOM1, SLC25A14, METTL2A, LINC01089, L3HYPDH, NPHP3) had a significant diagnostic value for OA. Additionally, we explored the characteristics of immune cell infiltration in patients with OA to provide new insights into OA treatment.
**Ethics Approval**

This study was conducted in accordance with the principles of the Declaration of Helsinki. The human-related data, surgical samples involved in this study have been approved by the Ethics Committee of the First Affiliated Hospital of Harbin Medical University.

**Informed Consent**

The study was conducted with informed consent from patients. All patients signed informed consent forms approved by the institutional review Committee.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


