Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative Mendelian randomization, single-cell, and bulk transcriptomic analyses reveal the landscape of druggable genome- and gut microbiome-associated biomarkers in osteoarthritis

Wangyang Zheng
Guizhou University of Traditional Chinese Medicine
Shuang Chen
Guizhou University of Traditional Chinese Medicine
Lang Zhao
Guizhou University of Traditional Chinese Medicine
Lingli Wang
Guizhou University of Traditional Chinese Medicine
Peng Liu
Guizhou University of Traditional Chinese Medicine
Ji Fei
drhosi@sina.com
Guizhou University of Traditional Chinese Medicine

Research Article

Keywords: Osteoarthritis, Druggable genome, Gut microbiome, Gene set enrichment analysis, Subcellular localization

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10016781/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Integrative Mendelian randomization, single-cell, and bulk transcriptomic analyses reveal the landscape of druggable genome- and gut microbiome-associated biomarkers in osteoarthritis

Wangyang Zheng[1#], Shuang Chen[1#], Lang Zhao[1],Lingli Wang[1],Peng Liu[1], Ji Fei[1,2*]

1. Guizhou University of Traditional Chinese Medicine, Guiyang 550002, China;

2. The First Affiliated Hospital of Guizhou University of Traditional Chinese Medicine, Guiyang 550002, China.

#Wangyang Zheng and Shuang Chen have contributed equally to this work and share first authorship.

***Corresponding Author: Ji Fei**

The Authors' email addresses and their academic degrees:

Wangyang Zheng(bachelor): 924574349@qq.com;

Shuang Chen(bachelor): 3492282726@qq.com;

Lingli Wang(bachelor): kcxfcll@sina.com;

Lang Zhao(bachelor): 13765973474@163.com;

Peng Liu(bachelor): 20232023@163.com;

Ji Fei(doctor): drhosi@sina.com.

ORCID:

Wangyang Zheng(bachelor): 0009-0006-5202-4858;

Shuang Chen(bachelor): 0009-0003-5373-5211;

Lingli Wang(bachelor): 0009-0008-2199-0695;

Lang Zhao(bachelor): 0009-0001-4557-1080;

Peng Liu(bachelor): 0009-0008-7267-7658;

Ji Fei(doctor): 0009-0008-6120-6965.

Authors' contributions

Zheng Wangyang was responsible for conceptualization, formal analysis, and writing the manuscript, Fei Ji was responsible for raising questions, reviewing, guiding analysis, and revising the article, Chen Shuang was responsible for data collection and organization, Zhao Lang, Wang Lingli and Liu Peng are responsible for data collation and analysis. All authors read and approved the final manuscript. The authors declare that all data were generated in-house, that no paper mill was used and that no AI tool has been used for the generation of text or figures.

Acknowledgments

We would like to express our sincere gratitude to all individuals and organizations who supported and assisted us throughout this research. Without your support, this research would not have been possible.

Abstract

Introduction: Osteoarthritis (OA) was found to be closely associated with druggable genome (DG) and gut microbiome (GM), whereas their interaction mechanisms were not clarified. Relevant target gene biomarkers were regarded as critical for OA research.

Aim: This study aimed to identify OA-related biomarkers linking DG and GM, and to explore their underlying mechanisms and key cell types.

Methods: Differentially expressed genes (DEGs) of OA were obtained via differential expression analysis. The OA-related DRGs and OA-related GMs were explored via Mendelian randomization (MR). GM-associated metabolites and GRGs were retrieved from corresponding databases. Candidate genes were derived from the intersection of DEGs, OA-related DRGs, and GRGs. Biomarkers were ultimately

identified via machine learning algorithms, expression level validation, and receiver operating characteristic (ROC) curve analysis. Comprehensive analyses, including gene set enrichment analysis (GSEA) and molecular dynamics simulations, were conducted to elucidate underlying mechanisms. Subsequently, key cells were acquired via single-cell RNA sequencing (scRNA-seq), and cell communication patterns within these cells were explored.

Results: A total of 5,029 DEGs, eight OA-related DRGs, four OA-related GMs, 22 GM-related metabolites, 461 GRGs, eight candidate genes, and two biomarkers (LDHA and GRIK1) were obtained in this study. Both biomarkers were significantly enriched in the unfolded protein response. The complex structures between four core metabolites and their corresponding biomarkers were relatively stable, such as those of dihydrocaffeic acid-GRIK1. ScRNA-seq analysis was performed, and endothelial cells were defined as key cells. Strong intercellular communication was detected between endothelial cells and myeloid_cells.

Conclusion: In this study, LDHA and GRIK1 were defined as OA biomarkers, and endothelial cells were recognized as pivotal cells. Novel targets were yielded for OA diagnosis and microbial metabolome-based intervention.

Keywords: Osteoarthritis, Druggable genome, Gut microbiome, Gene set enrichment analysis, Subcellular localization

1. Introduction

Osteoarthritis (OA) is a prevalent degenerative joint disease characterized by cartilage degradation, subchondral bone remodeling and osteophyte formation, which is driven by chondrocyte dysfunction and low-grade synovitis[1-3]. It clinically manifests as joint pain, morning stiffness, local swelling and

motor dysfunction, severely compromising patients' daily activities and quality of life[4]. Current clinical treatments including physical rehabilitation, NSAIDs administration and intra-articular injection only relieve symptoms, failing to halt progressive cartilage damage and joint function deterioration[5]. Although joint replacement is applied for end-stage OA, it is accompanied by surgical risks and complications[6]. The insufficient understanding of OA molecular mechanisms, especially the regulatory function of disease biomarkers, greatly hinders the development of novel targeted therapeutic strategies. The druggable genome (DG), consisting of genes encoding drug-binding target proteins, markedly improves the efficiency of drug discovery[7]. Accumulating evidence confirms the vital regulatory role of DG genes in OA progression[8]. The gut microbiome (GM), a functional "organ" with endocrine and immune regulatory capabilities[9], is closely linked to OA: prebiotics reverse high-fat diet-induced OA exacerbation, and GM dysbiosis, metabolites (e.g., lipopolysaccharide, LPS), and bacterial components correlate with OA severity, inflammation, and pain[10, 11]. In recent years, the molecular crosstalk between DG genes and the GM has attracted increasing attention. Studies on sepsis have revealed that GM-derived metabolites such as butyrate can bind to G protein-coupled receptors encoded by DG genes, thereby regulating inflammatory pathways, a mechanism that has been exploited for the precision intervention of sepsis. Systems biology studies, integrating multi-omics and virtual screening, have constructed an interaction map of GM metabolites with the druggable GIKome, uncovering the associations between 114 sepsis-related targets and 335 metabolites [12]. This suggests that in diseases characterized by inflammation and metabolic disorders, GM metabolites may exert their effects by targeting proteins encoded by DG genes. OA is also characterized by chronic inflammation and metabolic abnormalities. Related studies have revealed an intrinsic association between gut microbiota (GM) imbalance and tryptophan metabolic disorders in OA patients, with decreased concentrations of indole derivatives

represented by indole-3-lactic acid showing a significant positive correlation with the occurrence of symptomatic hand OA[13]. Several DG genes (e.g., the aryl hydrocarbon receptor *AhR*) have been confirmed as targets of GM metabolites, and in OA chondrocytes, activation of *AhR* by indole derivatives can alleviate inflammation[14]. These findings imply a potential DG-GM regulatory network in OA, yet its specific components, molecular mechanisms and clinical targets remain unclarified.

Mendelian randomization (MR) constructs a causal inference framework using genetic variants as instrumental variables, thereby effectively circumventing the confounding factors commonly present in conventional observational study designs[15]. MR has been successfully applied to OA research, identifying CEBPB and CX3CR1 as protective factors and validating causal associations between multiple GM taxa and OA[16]. Importantly, expression quantitative trait loci (eQTLs) within the DG gene region can serve as stable surrogates for gene expression[17]. Combined with the causal evidence of the GM-OA axis, MR provides a robust epidemiological tool for the systematic screening of GM-regulated DG-associated OA biomarkers[18]. Single-cell RNA sequencing (scRNA-seq) broke through the limitations of traditional bulk transcriptomics, which only acquired the average gene expression levels of tissues and masked cellular heterogeneity. It was applied to resolve gene expression profiles at the single-cell level and was utilized to identify disease-specific cell subpopulations, characterize cellular heterogeneity, and evaluate tissue immune infiltration as well as intercellular communication patterns.[19, 20]

Biomarkers associated with DG and GM for OA were screened from transcriptomic data via MR validation, machine learning and expression verification in this study. Biological functions of these biomarkers were investigated, and the interactive regulatory mechanisms of the DG-GM-OA axis were elucidated. Key cells linked to the screened biomarkers were acquired by scRNA-seq, and the

developmental trajectory as well as intercellular communication of these pivotal cells were explored, and novel references were provided for clinical diagnosis and targeted treatment of OA.

2. Materials and methods

2.1 Data

OA datasets were acquired from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). The GSE51588 (GPL13497) consists of the expression profiles of 40 OA subchondral bone tissue samples (OA) and 10 normal subchondral bone tissue samples control[21]. Additionally, the GSE114007 (GPL18573) consists of the expression profiles of 10 OA knee cartilage tissue samples (OA) and 10 normal knee cartilage tissue samples (control)[22] (Samples from other platforms were excluded). A total of 5,883 druggable genome (DG)-related genes (DRGs) were obtained from the literature with[23] (**Supplementary Table 1**).

The OA genetic variation dataset (ebi-a-GCST90038686) was downloaded from the IEU OpenGWAS platform (https://gwas.mrcieu.ac.uk/), with an inclusion criterion of more than 1,000 cases. This dataset comprised 484,598 participants of European ancestry (ncase = 39,515 and ncontrol = 445,083), encompassing a total of 9,587,836 SNPs.

In this study, two sets of MR analyses were conducted. One was the MR analysis between DRGs and OA, and the eQTL data of genes related to its exposure factors (i.e., the data related to DRG cis-eQTLs) were obtained from the IEU OpenGWAS (involving converting genes into Ensemble identifiers (IDs) to obtain exposure IDs). The other part involved MR analysis of GM and OA, with the GM-eQTL data for the corresponding exposure factors sourced from the MiBioGen database. (https://mibiogen.gcc.rug.nl/) (involving converting genes into Ensemble IDs to obtain exposure IDs).

The GSE114007 and GSE51588 data were downloaded on 24 July 2025. The MR data and DRGs were

downloaded on 23 July 2025. The GM-eQTL data were downloaded on 25 July 2025.

2.2 Filtering of instrumental variables (IVs)

Two MR analyses were conducted in this study.The outcome variable in both MR analyses was OA; eQTLs of DRGs served as exposure factors in the DRGs–OA analysis, whereas eQTLs of GM functioned as exposure factors in the GM–OA analysis. Analyses were executed using the TwoSampleMR package (v 0.6.4)[24], with IV selection required to simultaneously satisfy three core prerequisites: ① SNPs must be strongly associated with exposure factors (relevance assumption); ② SNPs must be independent of confounding factors (independence assumption); and ③ the effect of SNPs on the outcome must be transmitted exclusively through the exposure factor (exclusion restriction assumption). Based on the assumptions of MR, the extract_instruments function was used to find SNPs significantly associated with the exposure factor ($p < 5\times10^{-8}$). Then, the clump option was set to TRUE to remove the SNPs with linkage disequilibrium (LD) ($R^2 = 0.001$ (the MR analysis between DRGs and OA), $R^2 < 0.001$ (the MR analysis between GM and OA), kb = 10). The outcome variables were read using the extract_outcome_data function. Subsequently, based on the outcome GWAS data and the previously screened IVs, IVs significantly associated with the outcome were removed (proxies = TRUE; rsq = 0.8). During screening, at least two associated SNPs were retained per exposure factor, and loci with an F-statistic below 10 were excluded. The F-statistic was calculated according to the above formula: $F=R^2(N-K-1)/\{K(1-R)^2\}$, in the formula, R^2 represents the cumulative explained variance of selected SNPs for the exposure variable, while K and N correspond to the number of SNPs ultimately included and the sample size of the corresponding GWAS, respectively. Consistency correction of effect alleles and effect sizes was completed via the harmonise_data function to ensure the accuracy of downstream analyses.

155 **2.3 MR analysis and sensitivity analysis**

156 MR analysis integrated five methods via the mr function: MR Egger[25], weighted median[26], inverse
157 variance weighted (IVW)[27], simple model[24], and weighted model[28], with IVW serving as the primary
158 method for effect assessment ($p < 0.05$). Genes/GMs with OR > 1 were defined as risk factors for OA,
159 while those with OR < 1 were defined as protective factors. For visualization, mr_scatter_plot,
160 mr_forest_plot, and mr_funnel_plot functions were called successively to generate scatter plots
161 (exposure–outcome correlation), forest plots (diagnostic validity of exposure factors), and funnel plots
162 (assessment of adherence to Mendel's second law). To assess the reliability of MR results,
163 multidimensional sensitivity analyses were implemented: Cochran's Q test was completed via the
164 mr_heterogeneity function, with mr_ivw_fe activated when $p > 0.05$ and mr_ivw_mre applied otherwise;
165 horizontal pleiotropy was evaluated using the mr_pleiotropy_test function ($p > 0.05$). LOO analysis was
166 completed via mr_leaveoneout, with results output as forest plots using mr_leaveoneout_plot.
167 Subsequently, the steiger_filtering function in the TwoSampleMR package (v 0.6.4) was called to verify
168 the causal directionality between exposure and outcome, with the criterion that the causal direction was
169 TRUE and $p < 0.05$. All final results were visualized as forest plots using the forestploter package (v
170 1.1.3)[29]. Finally, the genes that passed the MR analysis were named as OA-related DRGs. The GMs that
171 passed the MR analysis were named as OA-related GMs. OA-related GMs at the genus level were
172 selected for subsequent acquisition of GM-related metabolite target genes (GRGs).

173 **2.4 Identification of GM-related metabolites, GRGs, and differentially expressed genes (DEGs)**

174 Metabolite information of OA-related GM at the genus level was sourced from the gutMGene v1.0
175 database (http://bio-computing.hrbmu.edu.cn/gutmgene/#/home), from which GM-related metabolites
176 were screened. Target prediction for these metabolites was conducted based on the SEA

(https://sea.bkslab.org/) and STP (http://www.swisstargetprediction.ch/) platforms, from which GRGs were subsequently determined.

To investigate expression differences between OA and control groups in the GSE51588 dataset, PCA was first performed and principal component analysis (PCA) plots were generated using the FactoMineR package (v 2.11)[30].Differential expression analysis was completed with the limma package (v 3.58.1))[31], with DEGs filtered at thresholds of $|\log_2 FC| > 0.5$ and $p < 0.05$. For visualization, the ggplot2 package (v 3.5.1)[32]. was used to produce differential expression plots, and the circlize package (v 0.4.15)[33] was employed to construct heatmaps highlighting the top 10 up-regulated and top 10 down-regulated genes with the highest $|\log_2 FC|$ values.

2.5 Identification and function analyses of candidate genes

Candidate genes were identified by taking the intersection of DEGs, OA-related DRGs, and GRGs using the ggvenn package (v 1.2.2)[34]. A genus-level GM-GM-related metabolite-candidate gene-OA (GMCO) network was constructed with the networkD3 package (v 0.4.1)[35],, from which candidate genes and GM-related metabolites corresponding to each genus-level GM were extracted. Metabolic pathway enrichment analysis was executed via MetaboAnalyst 6.0 and visualized as bubble plots. Gene Ontology (GO) enrichment analysis ($p < 0.05$) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis ($p < 0.05$) were further performed on the candidate genes using the clusterProfiler package (v 4.8.3)[36] to systematically dissect their biological functions and pathway interaction mechanisms in OA. The parameter pAdjustMethod was set to "Benjamini-Hochberg (BH) correction method". GO analysis mainly involved three parts: biological process (BP), molecular function (MF), and cellular component (CC). Based on the standard of the largest gene numbers, the top five pathway enrichment results of GO analysis and the top 10 pathway enrichment results of KEGG analysis were

visualized. A genus-level GM-GM-related metabolite-candidate gene-KEGG pathway (GMCK) network was established using the networkD3 package (v 0.4.1); the protein-protein interaction (PPI) network (Confidence = 0.15) for candidate genes was constructed via the STRING database (https://string-db.org/) with the species set to Homo sapiens, and ultimately visualized using the circlize package (v 0.4.15)[33]. Isolated gene nodes without PPI with other genes were removed to draw the network diagram.

2.6 Identification of biomarkers

In the GSE51588 dataset, the xgboost package (v 2.0.3.1)[37] and the randomForest package (v 4.7.1.1)[38] were respectively invoked to perform XGBoost and RF analyses on candidate genes using 10-fold cross-validation, with the top five candidates ranked by importance selected as RF genes.. In the XGBoost analysis, the search range was set to 0.001, 0.01, 0.1, and 0.3, while the remaining parameters were kept fixed. The fixed parameters for the XGBoost model were configured as: nrounds = 30, colsample_bytree = 1, min_child_weight = 1, gamma = 0, subsample = 0.5, and max_depth = 6; after the optimal eta was determined, feature importance scores were output by the xgb.importance function. The results of both machine learning algorithms were intersected using the ggvenn package (v 1.2.2) to obtain key genes. Across the GSE51588 and GSE114007 datasets, differential expression of key genes was systematically assessed using the Wilcoxon test ($p < 0.05$); key genes with consistent and directionally concordant expression differences across both datasets were incorporated as candidate biomarkers and displayed using the ggplot2 package (v 3.5.1); the pROC package (v 1.18.5)[39] was used for ROC curve analysis and AUC calculation, with $1 > AUC > 0.7$ as the criterion for final biomarker confirmation.

2.7 Construction and verification of back propagation (BP) neural network

To evaluate the overall predictive performance of the biomarkers, in GSE51588, based on the biomarkers obtained from the above analysis, the neuralnet package (v 1.44.2)[40] was used. A BP neural network

was constructed with biomarkers as input variables and its architecture was visualized; ROC curves were plotted in both the GSE51588 and GSE114007 datasets using the pROC package (v 1.18.5) to evaluate the predictive performance of the BP network (criterion: $1 > AUC > 0.7$).

2.8 Immune-related analysis

Using GSE51588 as the basis, the GSVA package (v 1.42.0)[41]was applied to quantitatively assess the infiltration proportions of 28 immune cell types to reveal immune infiltration differences between OA and control groups; the Wilcoxon test ($p < 0.05$) was used for differential cell screening, with infiltration abundances and differential immune cell distributions visualized as heatmaps via the pheatmap package (v 1.0.12)[42]and box plots via the ggplot2 package (v 3.5.1), respectively. Spearman correlation analysis was further performed using the psych package (v 2.4.3)[43], with |correlation coefficient (cor)| > 0.3 and $p < 0.05$ as the screening thresholds to assess correlations between biomarkers and differential immune cells, with results visualized using the pheatmap package (v 1.0.12)[44].

2.9 Gene set enrichment analysis (GSEA)

For OA and control group samples in GSE51588, Spearman correlation coefficients between each biomarker and all remaining genes were calculated using the psych package (v 2.4.3), with cor values ranked from high to low as input for GSEA. GSEA was completed using the clusterProfiler package (v 4.8.3), with h.all.v2024.1.Hs.symbols downloaded from the Molecular Signatures Database (MSigDB) (https://www.gsea-msigdb.org/gsea/msigdb/) as the reference gene set; enriched pathways were filtered at |normalized enrichment score (NES)| > 1 and $p < 0.05$, and the top five pathways with the smallest p-values were visualized. The aforementioned datasets were downloaded on 29 July 2025.

2.10 Biomarkers-related analysis

Regulatory miRNAs of biomarkers were predicted via the MicroCosm database

(https://tools4mirs.org/software/mirna_databases/microcosm-targets/), and corresponding lncRNA targets of these miRNAs were retrieved from the miRNet database (https://www.mirnet.ca/), jointly used to construct lncRNA-miRNA-mRNA interaction pairs with networks implemented through the networkD3 package (v 0.4.1). Chromosomal distribution of biomarkers was mapped using the RCircos package (v 1.2.2)[45]. ; subcellular localization was predicted via the mRNALocater database (http://bio-bigdata.cn/mRNALocater/) and displayed as stacked bar charts using the ggplot2 package (v 3.5.1). Interaction networks between biomarkers and functionally similar genes were constructed in the GeneMANIA database (http://genemania.org/, species: Homo sapiens).

2.11 Molecular docking, molecular dynamics simulation, and toxicological analysis

A genus-level GM-GM-related metabolite-biomarker-OA (GGBO) network was constructed using the networkD3 package (v 0.4.1), and the core metabolites associated with biomarkers were obtained.

To accelerate the screening of potential candidate drug molecules, the SMILES molecular formulas of core metabolites were retrieved from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/), while their drug-likeness and toxicological properties were predicted using the SwissADME (http://www.swissadme.ch/index.php) and ADMETlab (https://admetmesh.scbdd.com/) databases, respectively.

Based on the screened core metabolites, Protein Data Bank (PDB) format structure files of biomarker-encoded proteins were downloaded from the UniProt database (https://www.uniprot.org/), and three-dimensional structural data of the core metabolites were sourced from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/). Molecular docking was conducted via CB-Dock2 (https://cadd.labshare.cn/cb-dock2/php/index.php), with a binding energy < -5 kcal/mol defined as indicative of favorable binding activity; docking results meeting this criterion were visualized using

PyMOL software (v 1.1.0)[46].

Molecular dynamics simulations were performed on the GROMACS (v 2024.4)[47], platform using the AMBER99SB-ILDN force field and the TIP3P (TIP3-point) water model; the cubic simulation box maintained a 1-nanometer gap between its boundary and the protein surface, with system electrical neutrality maintained by ion addition. Following energy minimization via the steepest descent method, NVT (particle number-volume-temperature) and NPT (particle number-pressure-temperature) equilibration were performed sequentially, both controlled using the V-rescale method (reference temperature 300 K, time step 2 femtoseconds, each lasting 100 picoseconds), with the production simulation running for a total of 20 nanoseconds. Computational analyses covered complex RMSD, protein RMSF, total system energy, and hydrogen bond count.

2.12 ScRNA-seq analysis

The single-cell transcriptomic dataset was analyzed using the Seurat package (v 5.0.1) [48]. High-quality cells were selected based on the following stringent criteria: 200 < nFeature_RNA (The number of genes measured per cell) < 5,000, 200 < nCount_RNA (Gene expression counts were measured in each cell) < 20,000, and percent_mt (The proportion of mitochondrial genes measured in each cell) < 25%. Genes expressed in fewer than three cells were simultaneously removed. Cells and genes that met the filtering criteria were recorded as quality-controlled cells and quality-controlled genes. Data were normalized via the NormalizeData function with the LogNormalize method and scale.factor = 10,000 to guarantee reliable data for subsequent analyses. The top 2,000 highly variable genes with relatively large coefficients of variation among cells were extracted using the FindVariableFeatures function with the vst method, and the top 10 genes with the highest variability were labeled and visualized using the LabelPoints function. Data were centralized using the ScaleData function, and principal component

analysis (PCA) was performed for dimensionality reduction on the 2,000 highly variable genes in the samples. The number of principal components (PCs) was determined using the JackStraw function, and the Jackstraw plot was generated using the JackStrawPlot function to select PCs via re-clustering with the permutation test algorithm. A scree plot was constructed using the ElbowPlot function to identify available dimensions, and the inflection point was defined as the statistically significant PC cutoff. Data integration was conducted according to samples using the harmony package (v 1.2.3)[49] to eliminate batch effects with parameters set as theta = 0 and lambda = 1. Cell clustering analysis was performed using the FindNeighbors and FindCluster (resolution = 0.2) functions, and doublets were identified and removed using the scDblFinder package (v 1.16.0)[50]. Subsequently, overall clustering analysis was conducted on the selected principal component samples using the Uniform Manifold Approximation and Projection (UMAP) algorithm via the RunUMAP function. Marker genes were obtained based on the cell type markers annotated in classic literatures[51, 52], and a bubble plot was generated for visualization. Manual annotation was performed for different cell clustering results using the literature-derived markers, and the annotation results were visualized using UMAP dimensionality reduction plots. Cell proportion plots across distinct sample groups and the expression profiles of biomarkers among various cell types were displayed. Cells with high biomarker expression were selected as pivotal cell populations.

2.13 Single-cell subtype analysis, trajectory, communication and metabolic profiling

Secondary dimensionality reduction and clustering were performed on key cells with resolution set to 0.2, and clustered cell subsets were annotated according to canonical cell markers from published literature[51]. Pseudotime trajectory analysis was implemented using the monocle package (v 2.28.0)[53]. Differential expression analysis was conducted across subpopulations of key cells via the FindAllMarkers function with parameters only.pos = T, min.pct = 0.25 and logfc.threshold = 1 to screen

subtype-specific differentially expressed genes. Heatmaps illustrating the expression of the top 10 differentially expressed genes across cell subsets were generated. Ligand-receptor pairs and intercellular molecular interactions among distinct cell populations were analyzed with the CellChat package (v 1.6.1)[54], and the quantity and probability of cell communication networks were visualized. Metabolic activities of diverse cell subsets were evaluated by the VISION algorithm embedded in the scMetabolism package (v 0.2.1)[55].

2.14 Statistical analysis

All bioinformatics-related data processing was completed within the R software (v 4.3.3) environment, and statistical comparisons of inter-group differences were conducted using the Wilcoxon test ($p < 0.05$).

3. Results

3.1 Acquisition of eight OA-related DRGs

In the MR analysis of DRGs and OA, eQTL data for 2,519 exposure factors were obtained, and a total of 77,864 SNPs were acquired (**Supplementary Table 2**). A total of 12 genes all passed the IVW method (**Table 1**). Then, 12 genes were obtained after being screened through the horizontal pleiotropy test (**Table 2**). A total of four genes failed the LOO test (**Supplementary Figure 4**). Then, eight genes were all validated through the Steiger directionality test (**Table 3**). Finally, eight OA-related DRGs were identified (**Figure 1a**). For detailed illustration, the MR results of LDHA (OR = 0.997, 95% confidence interval (CI) = 0.996 - 0.999, p = 0.004) and GRIK1 (OR = 0.997, 95% CI = 0.996 - 0.998, $p < 0.001$) were selected (**Figure 1a**). Among them, LDHA and GRIK1 were identified as potential protective factors. Forest plot results indicated that the MR effect sizes of LDHA and GRIK1 were both negative under the IVW method (**Figures 1b-c**). The scatter plot showed that the slopes of LDHA and GRIK1 were negative (**Figures 1d-e**). The funnel plot revealed a symmetric distribution of IVs for LDHA and

GRIK1 on both sides of the IVW line, suggesting that the MR analysis satisfied the requirements of the second law of Mendelian randomization (**Figures 1f-g**). These results indicated that LDHA and GRIK1 had protective associations with OA. Heterogeneity test results showed that p-values for both LDHA and GRIK1 exceeded 0.05, indicating the absence of significant heterogeneity in either gene (**Table 4**). Moreover, no confounding factors were found through the horizontal pleiotropy test in the MR results (**Table 2**). No severely biased points were detected in the LOO analysis, providing further support for the robustness of the MR results (**Figures 1h-i**). Finally, the Steiger test revealed that the causal relationship in this MR result was correct (**Table 3**). In conclusion, eight genes, such as LDHA and GRIK1, were recorded as OA-related DRGs for further analysis, and the results of the other six genes were presented in the **Supplementary Figures 1-4**. In summary, after multi-step validation including IVW, horizontal pleiotropy, LOO, and Steiger tests, eight OA-related DRGs (including the protective factors LDHA and GRIK1) were identified, and no heterogeneity, confounding effects, or major bias supported the reliability of these MR results

3.2 Acquisition of four OA-related GMs

In the MR analysis of GMs and OA, a total of 253 SNPs were acquired (**Supplementary Table 3**). A total of five GMs all passed the IVW method (**Table 5**). Then, five GMs were obtained after being screened through the horizontal pleiotropy test (**Table 6**). A total of one GM failed the LOO test (**Supplementary Figure 8**). Then, four GMs were all validated through the Steiger directionality test (**Table 7**). Finally, four OA-related GMs associated with OA were identified (**Figure 2a**). For detailed illustration, the MR result of genus.Bifidobacterium.id.436 (OR = 1.00429, 95% CI = 1.00235 - 1.00623, $p < 0.001$) was selected (**Figure 2a**). Among them, genus.Bifidobacterium.id.436 was identified as a risk factor. Multiple result figures jointly validated the above findings: the IVW forest plot showed that the

MR effect size of genus.Bifidobacterium.id.436 was > 0 (**Figure 2b**), the scatter plot slope was positive (**Figure 2c**), and the funnel plot indicated a symmetric distribution of IVs on both sides of the IVW line, consistent with the second law of Mendelian randomization (**Figure 2d**), collectively suggesting a risk association between genus.Bifidobacterium.id.436 and OA. The heterogeneity test yielded $p > 0.05$ for this GM, ruling out significant heterogeneity (**Table 8**); no confounding factors were detected in the horizontal pleiotropy test (**Table 6**); LOO analysis revealed no severely biased points, supporting result robustness (**Figure 2e**); and the Steiger test confirmed the correct causal direction (**Table 7**). In conclusion, four GMs such as genus.Bifidobacterium.id.436 were recorded as OA-related GMs for further analysis, and the results of the other three GMs were presented in the **Supplementary Figures 5-8**. Overall, following validation using IVW, horizontal pleiotropy, LOO, and Steiger tests, four OA-related GMs were identified, with genus.Bifidobacterium.id.436 as a risk factor. No evident heterogeneity, confounding effects, or bias confirmed the robustness of the MR results.

3.3 Acquisition of 22 GM-related metabolites, 461 GRGs, and eight candidate genes

Among the four OA-related GMs, only one was a genus-level GM, which was genus. Bifidobacterium.id.436. A total of 22 metabolites associated with genus. Bifidobacterium.id.436 were acquired through the database (**Supplementary Table 4**). The targets of the 22 GM-related metabolites were predicted using the SEA platform and STP platform, and finally, 461 GRGs were obtained (**Supplementary Table 5**). Differences in expression between the OA and control groups in GSE51588 were analyzed. PCA results showed that PC1 and PC2 accounted for 18.9% and 13.5% of the total data variance, respectively, together reflecting approximately 32.4% of transcriptomic variation (**Figure 3a**). Differential expression analysis in GSE51588 identified 5,029 DEGs in total, comprising 2,973 up-regulated and 2,056 down-regulated genes (**Figures 3b-c, Supplementary Table 6**). Based on the

intersection of DEGs, GRGs, and OA-related DRGs, eight candidate genes were ultimately determined (**Figure 3d, Supplementary Table 7**). GO enrichment analysis assigned these candidate genes to 354 BPs, 24 CCs) and 67 MFs (**Figures 3e, Supplementary Table 8**). The pathways enriched by candidate genes included astrocyte differentiation (BP), postsynaptic membrane (CC), and organic acid binding (MF). KEGG enrichment analysis further mapped the candidate genes to eight related pathways, including axon guidance (**Figures 3f, Supplementary Table 9**). As shown in the GMCO network, one genus-level GM was associated with 16 GM-related metabolites, and these GM-related metabolites corresponded to eight candidate genes, including genus.Bifidobacterium.id.436-folic acid-LDHA-OA (**Figure 3g**). As displayed in the GMCK network, eight candidate genes were mapped to 16 GM-related metabolites and eight pathway processes, including genus.Bifidobacterium.id.436-folic acid-LDHA-pyruvate metabolism (**Figure 3h**). A PPI network containing seven candidate genes was constructed, with a total of nine interaction relationships identified. Among all candidate genes, one was not found to have interactions with other genes. The constructed PPI network encompassed seven candidate genes, among which a notable protein interaction was identified between LDHA and AKR1C3 (**Figure 3i**). From the results of metabolic pathway enrichment analysis, the "one carbon pool by folate" pathway was identified as the most significant metabolic pathway (-LOG10(p) = 2.29, Rawp = 0.005), and "cysteine and methionine metabolism" was found to have the highest impact value in the metabolic network (impact = 0.104) (**Figure 3j, Supplementary Table 10**). In short, this section summarizes the OA-gut microbiome analysis, including microbial and metabolite screening, target prediction, functional enrichment, network construction, and pathway identification, establishing an integrated OA-microbe-gene-pathway axis.

3.4 Acquisition of two biomarkers

In GSE51588, a total of seven XGBoost genes were obtained through XGBoost analysis (**Figure 4a**). A total of five RF genes were obtained through RF analysis (**Figures 4b-c**). Then, there were five common genes among the XGBoost genes and RF genes, and these genes were selected as key genes (**Figure 4d, Supplementary Table 11**). Expression level verification results showed that in both GSE51588 and GSE114007, LDHA was significantly down-regulated and GRIK1 was significantly up-regulated in the OA group, with highly consistent directional changes across both datasets (**Figures 4e-f**), on the basis of which both were designated as candidate biomarkers. ROC curve analysis revealed AUC values of 0.833 and 0.738 for LDHA and GRIK1 in GSE51588, and 0.890 and 0.950 in GSE114007, respectively, all exceeding the diagnostic threshold of 0.7 (**Figures 4g-h**). Therefore, these two genes were identified as biomarkers for OA. A total of 8,421 iterations were performed for the model based on two biomarkers, and the loss function value was 4.16, which indicated that the model had reached a state of stable convergence after being fully trained. The value of the connection weight between the LDHA and GRIK1 nodes in the figure was used to intuitively reflect the difference in the contribution degree of these biomarkers in the prediction model, while the final error value of 4.163647 indicated the fitting accuracy level of the neural network on GSE51588 (**Figure 4i**). The combined model achieved AUC values of 0.963 and 0.880 in GSE51588 and GSE114007, respectively, indicating excellent diagnostic performance in discriminating OA from control samples (**Figures 4j-k**). In summary, this section describes biomarker screening and model validation. XGBoost and RF genes were identified, yielding five overlapping key genes. A total of two stable biomarkers (LDHA and GRIK1) exhibited consistent expression differences and AUC values > 0.7 across datasets. The paired-biomarker model converged steadily and achieved high diagnostic performance.

3.5 A correlation was observed between LDHA and differential immune cells

GSEA showed that LDHA was enriched in 33 pathways and GRIK1 was enriched in 30 pathways, with both being significantly enriched in unfolded protein response (**Figures 5a-b, Supplementary Tables 12-13**). mmune infiltration analysis demonstrated significant inter-group differences in the infiltration levels of multiple immune cell types, including neutrophils, between OA and control groups (**Figure 5c**), with 11 differential immune cells identified, encompassing activated dendritic cells among others (**Figure 5d**). Correlation analysis further revealed that LDHA was significantly correlated with both activated dendritic cells (cor = 0.39, p = 0.005) and central memory CD4 T cells (cor = 0.59, p = 6.82×10^{-6}) (**Figure 5e, Supplementary Table 14**). To sum up, LDHA and GRIK1 differed in pathway enrichment, and LDHA correlated with immune infiltration altered in OA.

3.6 The properties of biomarkers

Chromosomal localization analysis showed that LDHA and GRIK1 were situated on chromosomes 11 and 21, respectively (**Figure 6a**). LDHA and GRIK1 were mainly distributed in the extracellular exosome and cytosol; in addition, they were also distributed in the nucleus, ribosome, nucleoplasm, and other locations (**Figure 6b**). A total of 20 genes were found to be associated with LDHA and GRIK1, including GRIK2, GRIK5, and GRIK4, with related functions such as glutamate receptor activity (**Figure 6c**). miRNA prediction results showed that GRIK1 and LDHA yielded 34 and 26 predicted miRNAs, respectively, with 25 overlapping miRNAs shared between the two biomarkers and a total of 34 predicted miRNAs identified. Among these, 127 lncRNAs were predicted for 34 of the miRNAs. The XIST-hsa-miR-135a-5p-LDHA relationship was included in the lncRNA-miRNA-mRNA network (**Figure 6d, Supplementary Table 15**). Collectively, these results characterize the chromosomal localization, distribution, interacting genes, and lncRNA-miRNA-mRNA networks of LDHA and GRIK1, outlining their comprehensive molecular profiles.

3.7 Most metabolites were stably bound to the biomarkers

As shown in the GGBO network, eight metabolites were found to be associated with the two biomarkers, and their degrees of connectivity were all equal; therefore, these eight metabolites were all used as core metabolites for subsequent analysis (**Figure 7a**). The list included 7-Aminoheptanoic acid, D-Methionine, dihydrocaffeic acid, DL-Norvaline, folic acid, leucylproline, methionine, and n,n,n-Trimethyllysine. The results of drug similarity prediction for the eight metabolites, including molecular weight (MW), lipophilicity (LogP), number of hydrogen bond acceptors/donors, and topological polar surface area (TPSA), were presented in **Table 9**. Folic acid had the largest MW (MW = 441.4) and the highest polarity (TPSA = 213.28), while n,n,n-Trimethyllysine had the lowest LogP (LogP = -2.43). Dihydrocaffeic acid had the strongest lipophilicity (LogP = 0.63). These parameters reflected the drug-like properties of the compounds, such as absorption and penetration potential. According to the results of toxicological prediction in **Table 10**, the risks of cardiotoxicity (hERG) and drug-induced liver injury (DILI) caused by these metabolites were generally low; however, the DILI risks of folic acid and leucylproline were relatively high (> 0.65). Most compounds had strong hydrophilicity (with negative logP/logD values), while dihydrocaffeic acid exhibited moderate lipophilicity. The CYP3A4 inhibitory potential of all compounds was extremely low (< 10^-5), which indicated a low risk of drug-drug interactions. Special attention should be paid to the high DILI risk of folic acid and the potential cardiotoxicity of N,N,N-Trimethyllysine (hERG = 0.116). Molecular docking showed that except for the binding energies of GRIK1 docked with 7-Aminoheptanoic acid, DL-Norvaline, D-Methionine, and methionine, which were higher than -5 kcal/mol, the binding energies of all other dockings were lower than -5 kcal/mol. This indicated that the selected biomarkers had high binding affinity with the core metabolites (**Figures 7b-e, Table 11**). All docking complexes exhibited consistently low RMSD values (below 5 nm), indicating

stable overall structures (Figure 8a). Group trends were comparable, with structural fluctuations mainly localized to the protein N-terminus, C-terminus and specific flexible regions. Relatively high local RMSF peaks were observed in all docking groups except the LDHA-folic acid complex, implying enhanced regional protein flexibility in these binding modes. The LDHA-folic acid complex showed the lowest RMSF, reflecting stronger structural rigidity and superior binding stability (**Figure 8b**). During simulation, total system energy remained low and stable with minimal fluctuations (**Figure 8c**), and dynamic changes in hydrogen bond counts between small molecules and target proteins were also documented (**Figure 8d**). Collectively, these findings characterize metabolite screening, drug-likeness, toxicity, and molecular docking of the two biomarkers, supporting their biomarkermetabolite interactions with preclinical evidence.

3.8 Single-cell transcriptome profiling and pivotal cell screening of OA samples

A total of 25,192 genes and 74,632 cells were included before filtering. After filtering, 25,192 genes and 66,143 cells were retained (**Supplementary Figures 9a-b**). The top 2,000 highly variable genes were screened for subsequent analysis (**Supplementary Figure 9c**). PCA results showed that standard deviation stabilized after the first 30 PCs, all of which had $p < 0.05$; thus, the first 30 PCs were used for downstream analysis (**Supplementary Figures 9d-g**). In total, 12 cell clusters were identified via clustering and annotated into six cell types using canonical marker genes, including Mural cells, lymphoid cells, Mast cells, Myeloid cells, Fibroblasts, and Endothelial cells (**Figures 9a-b, Supplementary Table 16**). Marker genes exhibited high cell-type specific expression (**Figures 9c-d**). Proportions of fibroblasts, myeloid cells and endothelial cells were found to be relatively high in OA and control groups (**Figure 9e**). No expression of GRIK1 was detected in the single-cell dataset, whereas abundant expression of LDHA was observed in endothelial cells; accordingly, endothelial cells were

selected as pivotal cell populations (**Figure 9f**). Taken together, overall cellular composition was clarified and endothelial cells were determined as the core pivotal cell type for subsequent mechanistic exploration based on single-cell transcriptomic profiling results.

3.9 Subtype characterization, trajectory, intercellular communication and metabolic landscape of endothelial cells

After secondary dimensionality reduction and clustering, endothelial cells were divided into nine clusters and were annotated into four subtypes, including Tip-like ECs, Stalk-like ECs, Lymphatic ECs and Tumor ECs (**Figures 10a-b, Supplementary Table 17**). Pseudotime trajectory ordering was performed from left to right, and Tumor ECs together with state1 were positioned at the early differentiation stage; high expression of CD36 and RGCC was detected in Tumor ECs (**Figures 10c-f**). Abundant cell-cell communication with strong interaction strength was identified between endothelial cells and myeloid cells, and such communication was elevated in OA samples relative to control groups (**Figures 11a-d**). Strong CCL14-ACKR1 signaling from endothelial cells toward endothelial cells was detected in both OA and control cohorts (**Figures 11e-h**). Lower metabolic pathway activities were observed in Tumor ECs, whereas enriched expression of the Drug metabolism-other enzymes pathway was detected in the remaining three endothelial subtypes (**Figure 11i**). Collectively, distinct functional subtypes of endothelial cells were distinguished, and their differentiation trajectory, intercellular communication features as well as subtype-specific metabolic patterns were systematically defined.

4. Discussion

4.1 Overview of Background and Key Findings

OA lacks effective disease-modifying therapies and early diagnostic biomarkers, largely due to insufficient understanding of its molecular mechanisms. Although the DG and GM are individually

linked to OA, their synergistic crosstalk mechanisms remain unclear. This study integrated MR and transcriptomic analyses to systematically explore DG- and GM-related OA biomarkers and their underlying regulatory networks.

MR analysis identified 8 OA-associated druggable genes and 4 OA-related gut microbial taxa, with genus.Bifidobacterium.id.436 correlating with 22 metabolites and 461 GM-associated metabolic genes. After intersection analysis, machine learning screening and two independent dataset validations, LDHA and GRIK1 were confirmed as core OA protective biomarkers. The BP neural network model based on the two biomarkers achieved favorable diagnostic efficacy, with AUC values of 0.963 (GSE51588) and 0.880 (GSE114007). Single-cell analysis identified endothelial cells as the pivotal cell type in OA, with intensive intercellular communication between myeloid cells and endothelial cells. Collectively, this study constructed a complete regulatory axis: genus.Bifidobacterium.id.436 - metabolites (folic acid/dihydrocaffeic acid) - biomarkers (LDHA/GRIK1) - metabolic pathways (folate one-carbon cycle/cysteine-methionine metabolism) - endothelial cells, providing systematic evidence for DG-GM synergistic effects in OA pathogenesis.

4.2 In-Depth Mechanisms of Biomarkers and Key Cells

LDHA and GRIK1 function synergistically via a unified GM-metabolite-biomarker axis. genus.Bifidobacterium.id.436-derived folic acid and dihydrocaffeic acid target LDHA and GRIK1 respectively, regulating OA through distinct but complementary downstream metabolic, inflammatory and proteostatic pathways.

4.2.1 LDHA: A Protective Biomarker Regulated by UPR, Metabolic Synergy, Immunity, and ceRNA Network

MR analysis confirmed that LDHA is significantly downregulated in OA cartilage and exerts a protective

effect against OA, serving as a promising therapeutic target[56]. The GM-LDHA association is mediated by genus.Bifidobacterium.id.436-derived folic acid. Molecular docking verified high binding stability between LDHA and folic acid (binding energy < -5 kcal/mol, low RMSF and RMSD), indicating a direct regulatory interaction[57]. Cross-disease evidence further supports this DG-GM regulatory loop, as intestinal LDHA modulates GM composition via hypoxia-inducible factor 2α, forming a conserved metabolic regulatory mechanism across metabolic diseases and OA[56]. Functionally, LDHA synergistically regulates the folate one-carbon cycle and cysteine-methionine metabolism. OA-induced LDHA downregulation disrupts chondrocyte DNA synthesis, methylation and cartilage matrix protein modification, impairing chondrocyte proliferation, matrix synthesis and anti-inflammatory capacity to accelerate OA progression[57]. LDHA also maintains cellular redox homeostasis; its downregulation triggers ROS accumulation and NF-κB pathway activation, upregulating inflammatory factors and matrix-degrading enzymes[58]. Immune infiltration analysis further revealed a close association between *LDHA* and the joint inflammatory microenvironment: *LDHA* expression was significantly positively correlated with activated dendritic cells (cor = 0.39, $p = 0.005$) and central memory $CD4^+$ T cells (cor = 0.59, $p = 6.82\times10^{-6}$), which is consistent with the critical role of dendritic cells in orchestrating joint inflammatory responses[59]. At the post-transcriptional level, the XIST-hsa-miR-135a-5p-LDHA ceRNA network modulates LDHA expression: lncRNA XIST sponges hsa-miR-135a-5p to relieve LDHA mRNA degradation and enhance its protective function[60]. These multi-layered mechanisms covering redox balance, immune regulation and epigenetic modification enrich the DG-GM crosstalk framework, providing a theoretical basis for multi-target OA therapy.

4.2.2 GRIK1: A Novel Protective Biomarker with Site-Specific Glycosylation-Dependent Function

GRIK1 is a newly identified protective OA biomarker with an abnormal upregulation in OA cartilage,

which is considered a compensatory response of chondrocytes to inflammatory stress and glutamatergic excitotoxicity[61]. Despite compensatory GRIK1 elevation, its protective effect is insufficient to halt OA progression, consistent with the causal protective effect confirmed by MR analysis. This finding reveals the unrecognized role of glutamate receptor signaling in maintaining chondrocyte homeostasis.

The GM-GRIK1 crosstalk relies on genus.Bifidobacterium.id.436-derived dihydrocaffeic acid, a metabolite produced from dietary polyphenol metabolism[62]. Molecular docking confirmed stable binding between GRIK1 and dihydrocaffeic acid (binding energy < -5 kcal/mol). Dihydrocaffeic acid exerts anti-inflammatory and cartilage-protective effects via inhibiting the NF-κB/MAPK pathway, and GRIK1 upregulation amplifies these protective effects by enhancing chondrocyte sensitivity to this metabolite[63]. It is thus hypothesized that the *GRIK1*-dihydrocaffeic acid complex may form a protective cascade: *GRIK1* upregulation enhances cellular responsiveness to dihydrocaffeic acid, amplifying its anti-inflammatory and anti-catabolic effects. This mechanism positions dihydrocaffeic acid as a key mediator that enables crosstalk between the GM and the DG to modulate OA pathogenesis.

The protective cascade mediated by the *GRIK1*-dihydrocaffeic acid interaction extends beyond NF-κB/MAPK inhibition to engage a broader proteostatic mechanism. GSEA confirmed the enrichment of *GRIK1* in the unfolded protein response (UPR) pathway, wherein UPR dysregulation impairs chondrocyte homeostasis and promotes OA progression[64]. *GRIK1* upregulation may stabilize the unfolded protein sensor complex and potentiate UPR activation in response to increased protein misfolding under inflammatory conditions[65]. This UPR-mediated protective mechanism is further enhanced by optimized dihydrocaffeic acid metabolism: the anti-inflammatory properties of this metabolite reduce endoplasmic reticulum (ER) stress, creating a synergistic protective loop. A key feature of *GRIK1* function is the site-specific N-glycosylation at asparagine residues at the interface of its amino-

terminal domain and ligand-binding domain, with HNK-1 epitope modification regulating receptor desensitization kinetics and membrane stability[66]. This site-specific glycosylation is critical for the role of *GRIK1* as a glutamate receptor in chondrocyte signal transduction, influencing ligand-binding affinity and receptor trafficking to the cell surface[67]. This glycosylation modification mediates the metabolite-dependent protective effect of GRIK1. As a druggable target linking GM and OA, GRIK1 supports potential OA interventions via microbiota regulation, dietary polyphenol supplementation and targeted small-molecule therapy[1].

.

4.2.3 Endothelial cells: key stromal cells regulating local inflammation and systemic bone-vascular axis in OA

Associations between endothelial cells and OA were verified from systemic regulation and local histopathological perspectives in multiple published studies. First, systemic low-grade inflammation were induced by joint-derived inflammatory cytokines and extracellular vesicles entering the circulation; endothelial dysfunction was subsequently triggered via the bone-vascular axis. In turn, articular degeneration was aggravated by metabolic disorders such as obesity, and a bidirectional vicious cycle was therefore formed between endothelial injury and OA[68]. Second, endothelial cells inside the intermuscular adipose tissue of patients with knee OA were identified via single-nucleus RNA sequencing. Pathogenic inflammatory microenvironments were constructed collectively by endothelial cells, abnormally accumulated mast cells, B lymphocytes and proinflammatory adipocytes[69]. Third, endothelial infiltration was observed in the lesioned synovium and surrounding soft tissues of patients with hip OA. Local inflammatory bursts and articular matrix destruction were synergistically facilitated by endothelial cells together with diverse inflammatory immune cells and fibroblasts within pathological

595 microenvironments[70].

596 **4.3 GM-Metabolite-Biomarker Axis and Clinical Implications**

597 The BP neural network model built upon LDHA and GRIK1 demonstrated satisfactory diagnostic
598 efficacy in both the training and validation sets, with AUC values exceeding 0.88 in both, indicating
599 considerable potential for clinical translation[71]. The high diagnostic efficacy benefits from the
600 mechanistic complementarity of the two biomarkers: folic acid-modulated LDHA regulates metabolic
601 homeostasis, while dihydrocaffeic acid-targeted GRIK1 controls inflammation and proteostasis, jointly
602 covering core pathological dimensions of OA[72, 73]. This DG-GM integrated model provides mechanism-
603 based molecular signatures and references for personalized OA intervention.

604 Notably, this study still requires further validation. Large-scale, multicenter prospective clinical cohorts
605 are needed to verify the model's value in predicting OA progression and treatment prognosis.
606 Additionally, translating these transcriptomic findings into clinical assays faces challenges in
607 standardization, reproducibility and cost control.

608 **4.4 Summary and Precise Limitations**

609 This study systematically explored DG- and GM-associated OA biomarkers via integrated bioinformatic
610 analysis. The core findings include: 1) 8 causal OA-related druggable genes and 4 gut microbial taxa
611 identified by MR analysis; 2) two robust protective biomarkers (LDHA, GRIK1) validated in dual
612 independent datasets; 3) both biomarkers enriched in the UPR pathway and correlated with GM
613 metabolites and immune infiltration; 4) endothelial cells identified as key regulatory cells; 5) a complete
614 "GM-metabolite-biomarker-pathway-key cell" regulatory axis clarified for DG-GM-OA crosstalk.

615 LDHA and GRIK1 exert complementary protective functions. LDHA modulates OA via metabolic
616 synergy, redox balance and immune regulation, while glycosylation-dependent GRIK1 mediates

metabolite-induced anti-inflammatory and proteostatic protection. The constructed diagnostic model and gut microbial metabolites provide novel, safe candidates for OA precise therapy.

This study has several limitations. First, all GWAS datasets were derived from European populations, limiting the cross-ethnic generalizability of the conclusions, which requires multi-ethnic cohort validation. Second, all findings are based on bioinformatic analysis, lacking in vitro and in vivo functional verification of LDHA and GRIK1. Third, this study only analyzed genus-level gut microbiota; future shotgun metagenomic sequencing is needed to identify strain-specific metabolic functions in OA progression.

Conclusion

This study screened and validated two novel protective OA biomarkers (LDHA and GRIK1) by integrating MR and transcriptomic analyses, confirming their tight association with DG and GM. Endothelial cells were identified as the key stromal cells in OA pathogenesis. A complete "microbe-metabolite-biomarker-pathway-key cell" regulatory cascade was successfully constructed, solidifying the mechanistic evidence of DG-GM crosstalk in OA. The dual-biomarker diagnostic model achieved robust diagnostic efficacy (AUC > 0.88), and gut microbial metabolites (especially dihydrocaffeic acid) demonstrated favorable safety and therapeutic potential. These findings update the molecular mechanism of gut-joint axis in OA, providing new insights for OA diagnosis and targeted precise treatment. Further multi-ethnic cohort validation and systematic experimental verification will strengthen these results, accelerate clinical translation, and improve OA patient prognosis.

References

[1] SUN Q, NIU F, WANG L, et al. Decoding Fibroblast Heterogeneity in Osteoarthritis: Identification of a Fibrosis-Associated Subtype and Novel Diagnostic Biomarkers [J]. Mediators

Inflamm, 2025, 2025: 7066432.

[2] COACCIOLI S, SARZI-PUTTINI P, ZIS P, et al. Osteoarthritis: New Insight on Its Pathophysiology [J]. J Clin Med, 2022, 11(20).

[3] WANG Y H, SU C H, CHEN L C, et al. miR-548aj-3p and miR-3127-3p suppress RANKL-facilitated inflammatory cytokines and catabolic factor in osteoarthritis and rheumatoid arthritis [J]. Int J Med Sci, 2025, 22(14): 3650–63.

[4] XIAO J, LEI W, ZHANG H, et al. Identification and Validation of Fibroblast-Associated Genes in Osteoarthritis Based on High-Dimensional Weighted Gene Coexpression Network Analysis [J]. J Immunol Res, 2025, 2025: 5547701.

[5] ZHU Y, DENG L, XIA J, et al. Identification of Key Glycolysis-Related Genes in Osteoarthritis and Their Correlation with Immune Infiltration Using Bioinformatics Analysis and Machine Learning [J]. Open Access Rheumatol, 2025, 17: 157–71.

[6] CAI Y, YU D, MENG J, et al. Multi-omics analysis and validation of autophagy-related diagnostic biomarker in osteoarthritis [J]. Ann Med, 2025, 57(1): 2548045.

[7] YIN K F, CHEN T, GU X J, et al. Systematic druggable genome-wide Mendelian randomization identifies therapeutic targets for sarcopenia [J]. J Cachexia Sarcopenia Muscle, 2024, 15(4): 1324–34.

[8] LI S, PENG Y, YU Y, et al. Investigating the Impact of Circulating MicroRNAs on Knee and Hip Osteoarthritis: Causal Links, Biological Mechanisms, and Drug Interactions [J]. Int J Mol Sci, 2024, 26(1).

[9] LAZAR V, DITU L M, PIRCALABIORU G G, et al. Aspects of Gut Microbiota and Immune System Interactions in Infectious Diseases, Immunopathology, and Cancer [J]. Front Immunol, 2018, 9: 1830.

[10] SCHOTT E M, FARNSWORTH C W, GRIER A, et al. Targeting the gut microbiome to treat the osteoarthritis of obesity [J]. JCI Insight, 2018, 3(8).

[11] GUAN Z, JIA J, ZHANG C, et al. Gut microbiome dysbiosis alleviates the progression of osteoarthritis in mice [J]. Clin Sci (Lond), 2020, 134(23): 3159–74.

[12] PARADA VENEGAS D, DE LA FUENTE M K, LANDSKRON G, et al. Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases [J]. Front Immunol, 2019, 10: 277.

[13] SUN N, ZHAO Y, ZHANG A, et al. Gut microbiota and osteoarthritis: epidemiology, mechanistic analysis, and new horizons for pharmacological interventions [J]. Front Cell Infect Microbiol, 2025, 15: 1605860.

[14] WEI Z, LI F, PI G. Association Between Gut Microbiota and Osteoarthritis: A Review of Evidence for Potential Mechanisms and Therapeutics [J]. Front Cell Infect Microbiol, 2022, 12: 812596.

[15] QI Y, GAO Y, CAI H. Integrated eQTL-pQTL Mendelian randomization and single-cell sequencing reveal therapeutic targets in ovarian clear cell cancer [J]. Discov Oncol, 2025, 16(1): 1249.

[16] GAO H, GAN X, HE J, et al. Mendelian Randomization and Transcriptome Analyses Reveal Important Roles for CEBPB and CX3CR1 in Osteoarthritis [J]. Bioengineering (Basel), 2025, 12(9).

[17] YAN Y, HE J, XU Z, et al. Mendelian randomization based on genome-wide association studies and expression quantitative trait loci, predicting gene targets for the complexity of osteoarthritis as well as the clinical prognosis of the condition [J]. Front Med (Lausanne), 2024, 11: 1409439.

[18] LU Y, ZHOU Z, PAN D. Genetic insights into the roles of fatty acids and gut microbiota in osteoarthritis: A Mendelian randomization study [J]. Medicine (Baltimore), 2024, 103(49): e40674.

683 [19] LI J, LIU Z, ZHANG G, et al. Uncovering the heterogeneity of NK cells on the prognosis of HCC
684 by integrating bulk and single-cell RNA-seq data [J]. Front Oncol, 2025, 15: 1570647.
685 [20] TANG M, XU Y, PAN M. Integrative Analysis of scRNA-Seq and Bulk RNA-Seq Identifies
686 Plasma Cell Related Genes and Constructs a Prognostic Model for Hepatocellular Carcinoma [J]. J
687 Hepatocell Carcinoma, 2025, 12: 427–44.
688 [21] CHOU C H, WU C C, SONG I W, et al. Genome-wide expression profiles of subchondral bone in
689 osteoarthritis [J]. Arthritis Res Ther, 2013, 15(6): R190.
690 [22] FISCH K M, GAMINI R, ALVAREZ-GARCIA O, et al. Identification of transcription factors
691 responsible for dysregulated networks in human osteoarthritis cartilage by global gene expression
692 analysis [J]. Osteoarthritis Cartilage, 2018, 26(11): 1531–8.
693 [23] ZHANG C, HE Y, LIU L. Identifying therapeutic target genes for migraine by systematic
694 druggable genome-wide Mendelian randomization [J]. J Headache Pain, 2024, 25(1): 100.
695 [24] HEMANI G, ZHENG J, ELSWORTH B, et al. The MR-Base platform supports systematic causal
696 inference across the human phenome [J]. Elife, 2018, 7.
697 [25] BOWDEN J, DAVEY SMITH G, BURGESS S. Mendelian randomization with invalid
698 instruments: effect estimation and bias detection through Egger regression [J]. Int J Epidemiol, 2015,
699 44(2): 512–25.
700 [26] BOWDEN J, DAVEY SMITH G, HAYCOCK P C, et al. Consistent Estimation in Mendelian
701 Randomization with Some Invalid Instruments Using a Weighted Median Estimator [J]. Genet
702 Epidemiol, 2016, 40(4): 304–14.
703 [27] BURGESS S, SCOTT R A, TIMPSON N J, et al. Using published data in Mendelian
704 randomization: a blueprint for efficient identification of causal risk factors [J]. Eur J Epidemiol, 2015,

30(7): 543–52.

[28] HARTWIG F P, DAVEY SMITH G, BOWDEN J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption [J]. Int J Epidemiol, 2017, 46(6): 1985–98.

[29] DENG Q, WU Y, HU X, et al. Oolong Tea Consumption and the Risk of Oral Squamous Cell Carcinoma: A Propensity Score-Based Analysis in Southeast China [J]. Front Nutr, 2022, 9: 928840.

[30] HE J, ZHENG Z, LI S, et al. Identification and assessment of differentially expressed necroptosis long non-coding RNAs associated with periodontitis in human [J]. BMC Oral Health, 2023, 23(1): 632.

[31] WEI M, SUN P, LIU X, et al. Expression characterization of the guanylate-binding protein gene family in breast cancer and its association with the immune microenvironment [J]. PeerJ, 2025, 13: e20058.

[32] ZHANG H, FENG T, LIN J, et al. Bibliometric-driven research on chemoresistance in breast cancer: knowledge mapping, hotspot evolution, and emerging insights (1994-2024) [J]. Discov Oncol, 2025, 16(1): 1676.

[33] GU Z, GU L, EILS R, et al. circlize Implements and enhances circular visualization in R [J]. Bioinformatics, 2014, 30(19): 2811–2.

[34] SCHAARSCHMIDT S, GLAUBITZ U, ERBAN A, et al. Differentiation of the High Night Temperature Response in Leaf Segments of Rice Cultivars with Contrasting Tolerance [J]. Int J Mol Sci, 2021, 22(19).

[35] BOSTOCK M, OGIEVETSKY V, HEER J. D^3: Data-Driven Documents [J]. IEEE Trans Vis Comput Graph, 2011, 17(12): 2301–9.

[36] YU G, WANG L G, HAN Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters [J]. Omics, 2012, 16(5): 284–7.

[37] HUANG C, MURUGIAH K, MAHAJAN S, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: A retrospective cohort study [J]. PLoS Med, 2018, 15(11): e1002703.

[38] ZHENG X, HUANG J, MENG J, et al. Identification and Experimental Verification of PDK4 as a Potential Biomarker for Diagnosis and Treatment in Rheumatoid Arthritis [J]. Mol Biotechnol, 2025, 67(10): 3975–92.

[39] LI L, HUANG Y T, WANG L T, et al. ADH4-a potential prognostic marker for hepatocellular carcinoma with possible immune-related implications [J]. BMC Cancer, 2024, 24(1): 927.

[40] LI S, QUE Y, YANG R, et al. Construction of Osteosarcoma Diagnosis Model by Random Forest and Artificial Neural Network [J]. J Pers Med, 2023, 13(3).

[41] QI B, LIU H, ZHOU Q, et al. An immune-related lncRNA signature for the prognosis of pancreatic adenocarcinoma [J]. Aging (Albany NY), 2021, 13(14): 18806–26.

[42] GU Z, HüBSCHMANN D. Make Interactive Complex Heatmaps in R [J]. Bioinformatics, 2022, 38(5): 1460–2.

[43] ZHENG J, ZHANG T, GUO W, et al. Integrative Analysis of Multi-Omics Identified the Prognostic Biomarkers in Acute Myelogenous Leukemia [J]. Front Oncol, 2020, 10: 591937.

[44] BAIDA G, BHALLA P, KIRSANOV K, et al. REDD1 functions at the crossroads between the therapeutic and adverse effects of topical glucocorticoids [J]. EMBO Mol Med, 2015, 7(1): 42–58.

[45] ZHANG H, MELTZER P, DAVIS S. RCircos: an R package for Circos 2D track plots [J]. BMC Bioinformatics, 2013, 14: 244.

[46] ZHANG Y, WANG Y, WAN J, et al. Investigate the potential inhibitors of sphingosine kinase 1 (SphK1) with molecular dynamics and artificial intelligence drug design methods [J]. J Mol Model,

2025, 31(10): 282.

[47] MEI H, HUANG Y, YI J, et al. Molecular Dynamics Simulation of the Thermosensitive Gelation Mechanism of Phosphorylcholine Groups-Conjugated Methylcellulose Hydrogel [J]. Gels, 2025, 11(7).

[48] ZHANG P, LIU J, PEI S, et al. Mast cell marker gene signature: prognosis and immunotherapy response prediction in lung adenocarcinoma through integrated scRNA-seq and bulk RNA-seq [J]. Front Immunol, 2023, 14: 1189520.

[49] CAO G, XUAN X, LI Y, et al. Single-cell RNA sequencing reveals the vascular smooth muscle cell phenotypic landscape in aortic aneurysm [J]. Cell Commun Signal, 2023, 21(1): 113.

[50] GERMAIN P L, LUN A, GARCIA MEIXIDE C, et al. Doublet identification in single-cell sequencing data using scDblFinder [J]. F1000Res, 2021, 10: 979.

[51] TANG S, YAO L, RUAN J, et al. Single-cell atlas of human infrapatellar fat pad and synovium implicates APOE signaling in osteoarthritis pathology [J]. Sci Transl Med, 2024, 16(731): eadf4590.

[52] KIM N, KIM H K, LEE K, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma [J]. Nat Commun, 2020, 11(1): 2285.

[53] DU J, YUAN X, DENG H, et al. Single-cell and spatial heterogeneity landscapes of mature epicardial cells [J]. J Pharm Anal, 2023, 13(8): 894–907.

[54] CHEN Q, SU L, LIU C, et al. PRKAR1A and SDCBP Serve as Potential Predictors of Heart Failure Following Acute Myocardial Infarction [J]. Front Immunol, 2022, 13: 878876.

[55] GRISS J, VITERI G, SIDIROPOULOS K, et al. ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis [J]. Mol Cell Proteomics, 2020, 19(12): 2115–25.

[56] WU Q, LIANG X, WANG K, et al. Intestinal hypoxia-inducible factor 2α regulates lactate levels to shape the gut microbiome and alter thermogenesis [J]. Cell Metab, 2021, 33(10): 1988–2003.e7.

[57] WEN Z H, SUNG C S, LIN S C, et al. Intra-Articular Lactate Dehydrogenase A Inhibitor Oxamate Reduces Experimental Osteoarthritis and Nociception in Rats via Possible Alteration of Glycolysis-Related Protein Expression in Cartilage Tissue [J]. Int J Mol Sci, 2023, 24(13).

[58] GONG X, YANG S Y, WANG Z Y, et al. The role of hypoxic microenvironment in autoimmune diseases [J]. Front Immunol, 2024, 15: 1435306.

[59] ARRA M, SWARNKAR G, KE K, et al. LDHA-mediated ROS generation in chondrocytes is a potential therapeutic target for osteoarthritis [J]. Nat Commun, 2020, 11(1): 3427.

[60] WEI Y, DAI L, DENG Y, et al. LncRNA XIST promotes adjuvant-induced arthritis by increasing the expression of YY1 via miR-34a-5p [J]. Arch Rheumatol, 2023, 38(1): 82–94.

[61] ZHENG L, ZHANG Z, SHENG P, et al. The role of metabolism in chondrocyte dysfunction and the progression of osteoarthritis [J]. Ageing Res Rev, 2021, 66: 101249.

[62] MILLS C E, TZOUNIS X, ORUNA-CONCHA M J, et al. In vitro colonic metabolism of coffee and chlorogenic acid results in selective changes in human faecal microbiota growth [J]. Br J Nutr, 2015, 113(8): 1220–7.

[63] LU R, WANG Y G, QU Y, et al. Dihydrocaffeic acid improves IL-1β-induced inflammation and cartilage degradation via inhibiting NF-κB and MAPK signalling pathways [J]. Bone Joint Res, 2023, 12(4): 259–73.

[64] HUGHES A, OXFORD A E, TAWARA K, et al. Endoplasmic Reticulum Stress and Unfolded Protein Response in Cartilage Pathophysiology; Contributing Factors to Apoptosis and Osteoarthritis [J]. Int J Mol Sci, 2017, 18(3).

[65] ZHOU Z, LU J, YANG M, et al. The mitochondrial unfolded protein response (UPR(mt)) protects against osteoarthritis [J]. Exp Mol Med, 2022, 54(11): 1979–90.

[66] VERNON C G, COPITS B A, STOLZ J R, et al. N-glycan content modulates kainate receptor functional properties [J]. J Physiol, 2017, 595(17): 5913–30.

[67] MAYER M L. Crystal structures of the GluR5 and GluR6 ligand binding cores: molecular mechanisms underlying kainate receptor selectivity [J]. Neuron, 2005, 45(4): 539–52.

[68] LUO J, LI L, DENG Q, et al. Systemic implications of osteoarthritis: from local degeneration to systemic metabolic Dysregulation [J]. J Transl Med, 2026.

[69] ELINGAARD-LARSEN L O, WHYTOCK K L, DIVOUX A, et al. Single-Nuclei Resolution of Intermuscular Adipose Tissue Indicates an Inflammation-Associated Cellular Profile in Individuals With Knee Osteoarthritis: Findings From the SOMMA KOA Ancillary Study [J]. Aging Cell, 2026, 25(1): e70348.

[70] PARALIOV A T, VICAŞ R M, DîRNU R, et al. Hip osteoarthritis - histopathological aspects [J]. Rom J Morphol Embryol, 2025, 66(1): 217–27.

[71] DU J, ZHOU T, ZHANG W, et al. Developing the new diagnostic model by integrating bioinformatics and machine learning for osteoarthritis [J]. J Orthop Surg Res, 2024, 19(1): 832.

[72] ER A, AY I E, HOROZOĞLU CERAN T. Evaluating Turkish Readability and Quality of Strabismus-Related Websites [J]. Cureus, 2024, 16(4): e58603.

[73] XIA J, QIAO Z, HAO X, et al. LDHA-induced histone lactylation mediates the development of osteoarthritis through regulating the transcription activity of TPI1 gene [J]. Autoimmunity, 2024, 57(1): 2384889.

Statements & Declarations

Authors' contributions

Zheng Wangyang was responsible for conceptualization, formal analysis, and writing the manuscript, Fei Ji was responsible for raising questions, reviewing, guiding analysis, and revising the article, Chen Shuang was responsible for data collection and organization, Zhao Lang, Wang Lingli and Liu Peng are responsible for data collation and analysis. All authors read and approved the final manuscript. The authors declare that all data were generated in-house, that no paper mill was used and that no AI tool has been used for the generation of text or figures.

Disclosure

The authors declare no conflict of interest.

Funding

Not applicable.

Ethics approval and informed consent

This study does not involve human or animal samples, so ethical approval is not required.

Consent for publication

N/A

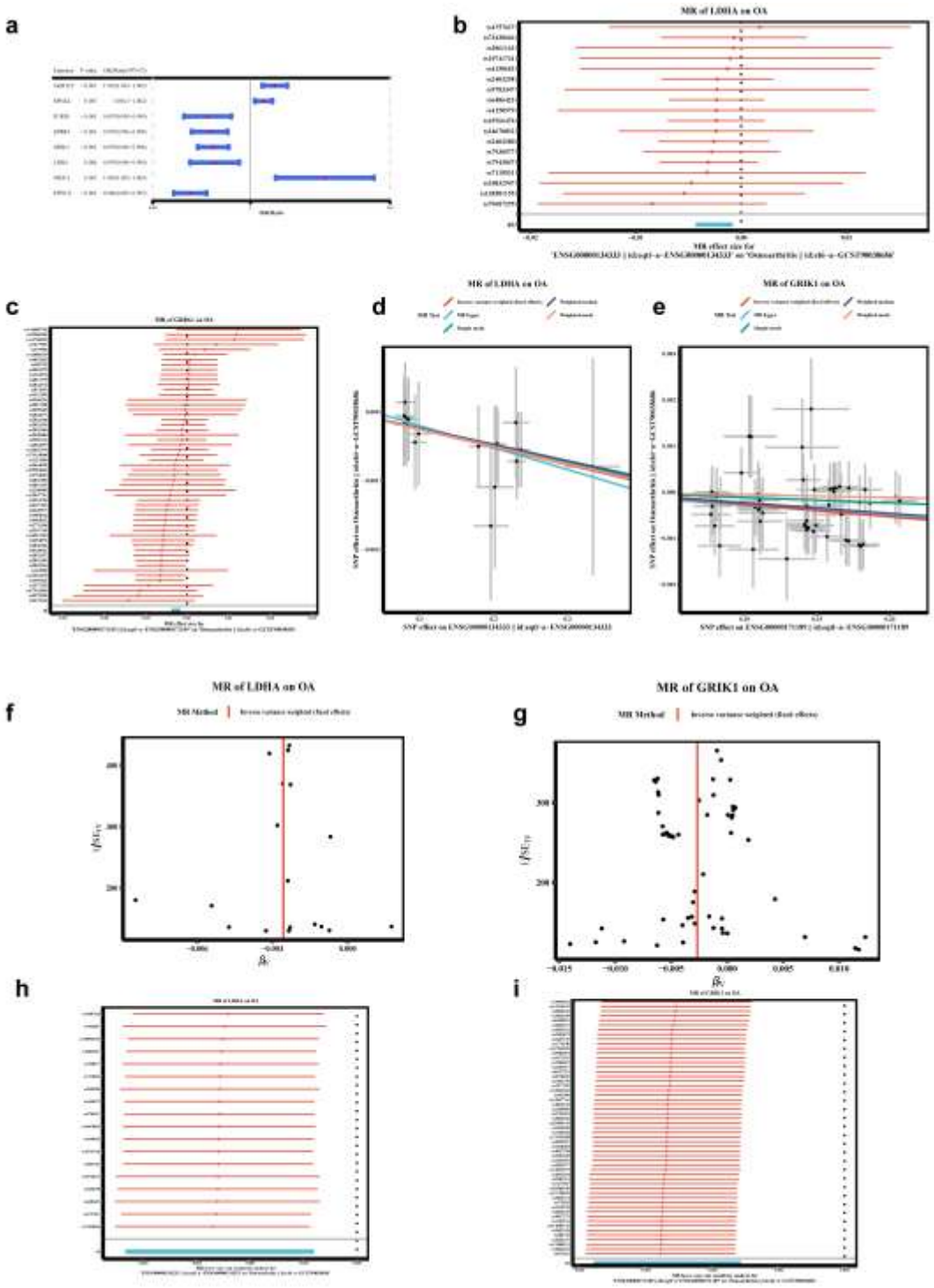
Data availability

All data generated and analyzed in this study are included in this published article and its supplementary information files. The public datasets used in this study are sourced as follows: the osteoarthritis transcriptome datasets (GSE51588, GSE114007) are available from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/); the osteoarthritis genetic variation dataset (ebi-a-GCST90038686) and druggable genome-related eQTL data are from the IEU OpenGWAS platform (https://gwas.mrcieu.ac.uk/); gut microbiome eQTL data are from the MiBioGen database (https://mibiogen.gcc.rug.nl/); gut microbiome-related metabolite information is from the gutMGene

v1.0 database (http://bio-computing.hrbmu.edu.cn/gutmgene/#/home); metabolite target prediction data are from the SEA platform (https://sea.bkslab.org/) and STP platform (http://www.swisstargetprediction.ch/); protein structure data are from the UniProt database (https://www.uniprot.org/); metabolite molecular structure and related property data are from the PubChem database (https://pubchem.ncbi.nlm.nih.gov/), SwissADME database (http://www.swissadme.ch/index.php), and ADMETlab database (https://admetmesh.scbdd.com/); the reference gene set for gene set enrichment analysis is from the Molecular Signatures Database (MSigDB, https://www.gsea-msigdb.org/gsea/msigdb/); protein-protein interaction data are from the STRING database (https://string-db.org/). The download dates for all datasets are detailed in the "Materials and Methods" section of the paper. For further access to relevant data, please contact the corresponding author.

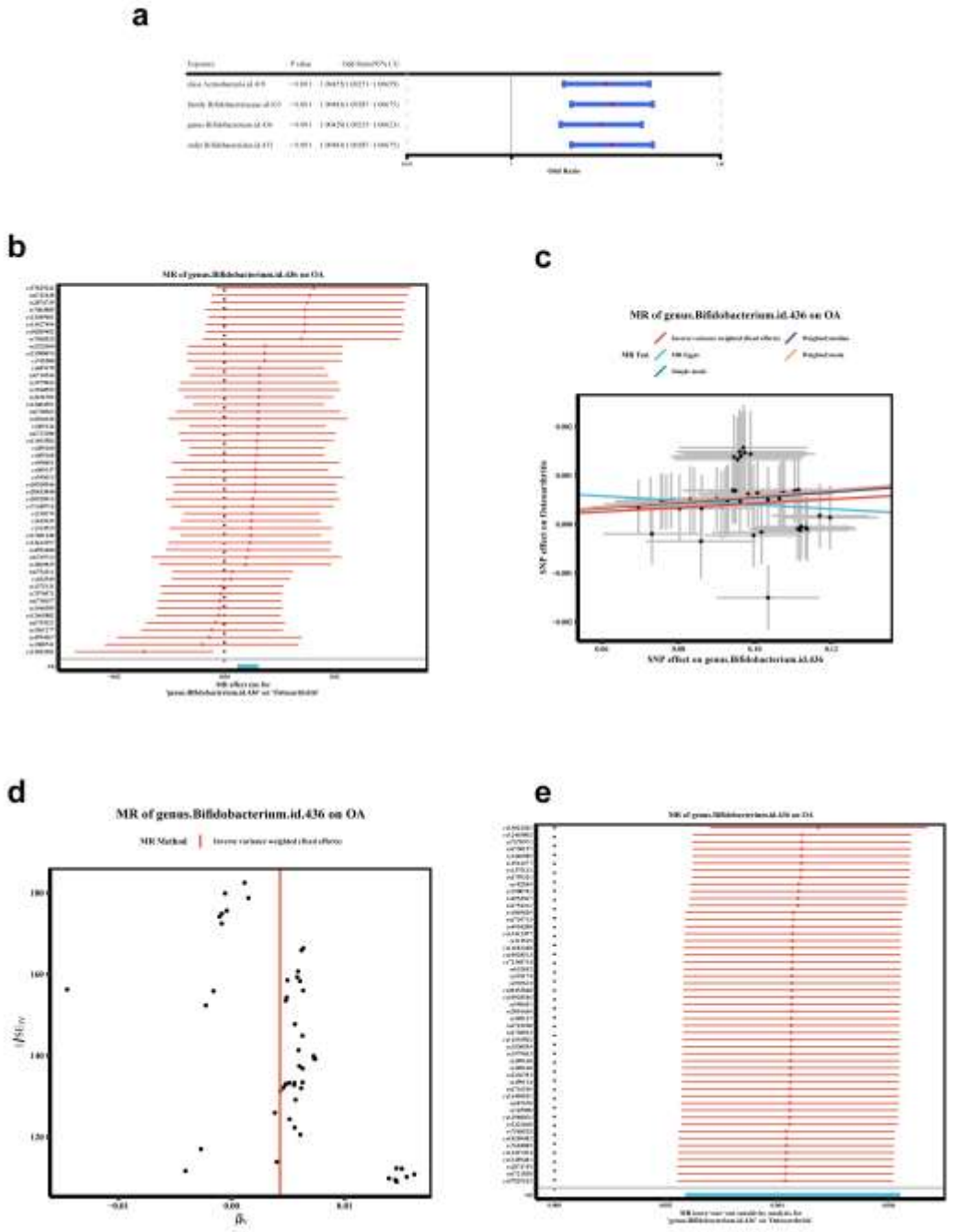
Figures

Figure 1 Acquisition of eight OA-related DRGs



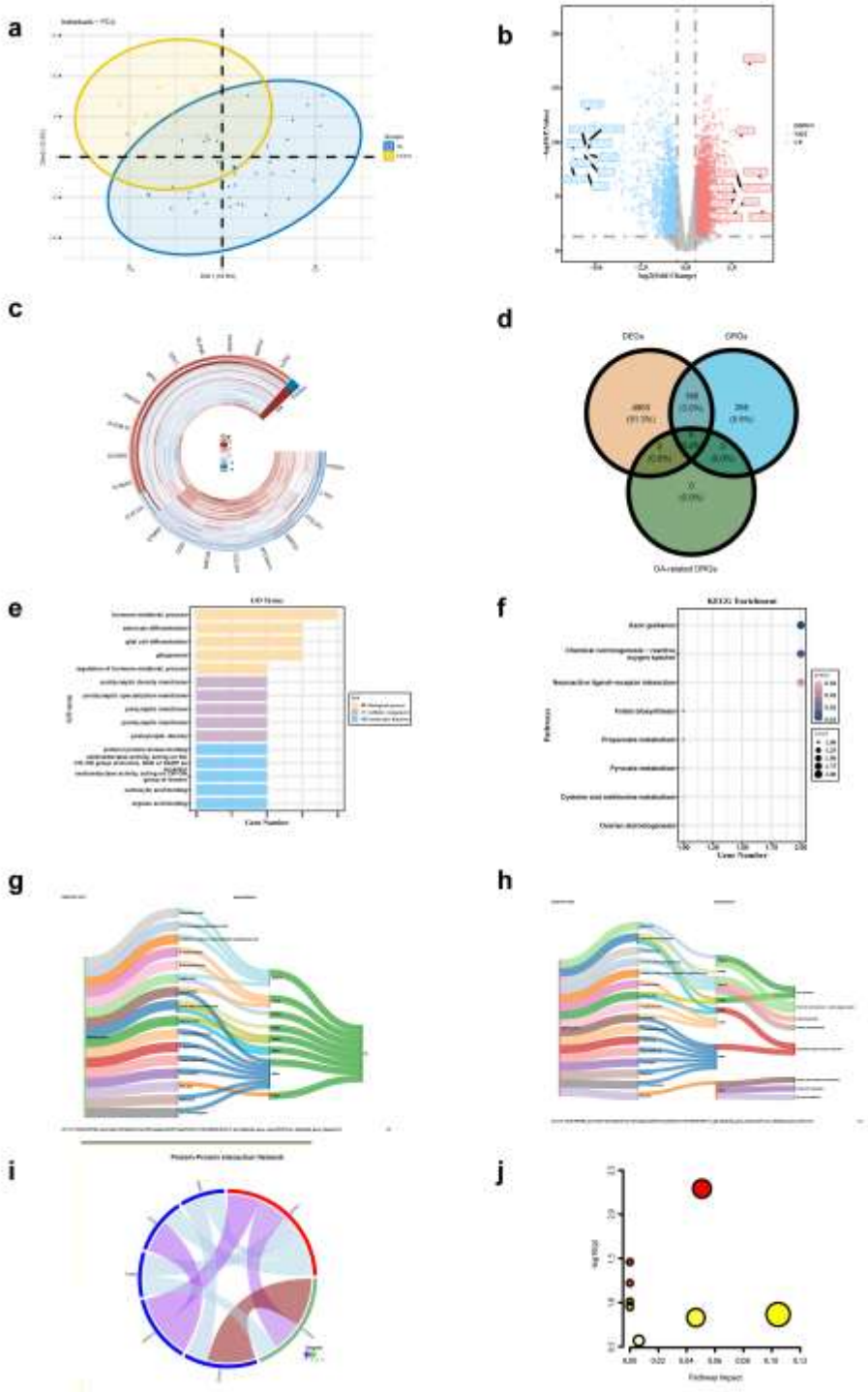
(a): Forest plot for IVW was presented; (b-c): Forest plots for MR results were displayed; (d-e): Scatter plots for MR results were shown; (f-g): Funnel plots for MR analysis were presented; (h-i): Forest plots for LOO test were displayed.

Figure 2 Acquisition of four OA-related GMs



(a): Forest plot for IVW was presented; (b): Forest plots for MR results were displayed; (c): Scatter plots for MR results were shown; (d): Funnel plots for MR analysis were presented; (e): Forest plots for LOO test were displayed.

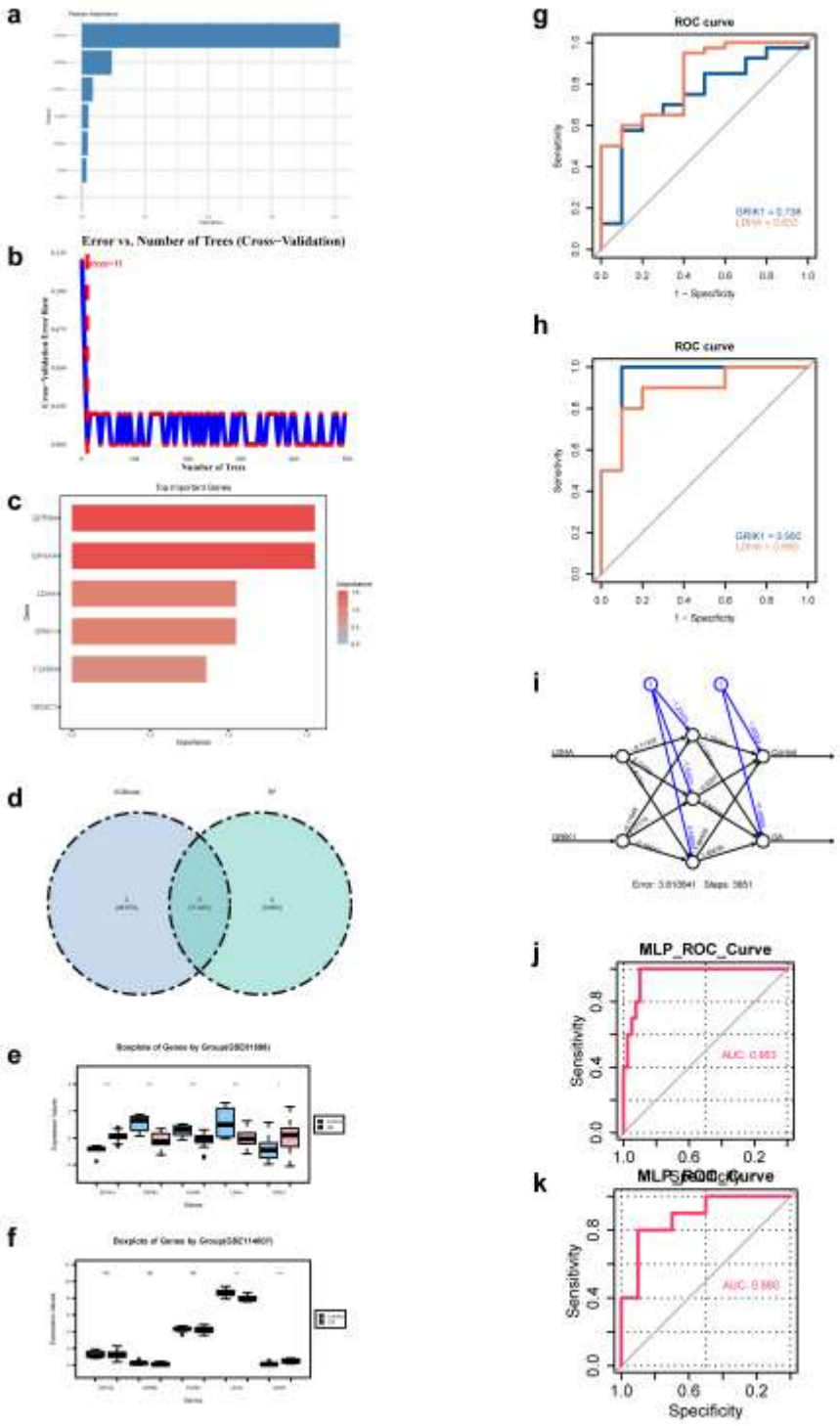
Figure 3. Acquisition of candidate genes



(a): The visualization of expression differences between the disease group and the control group was performed; (b): A volcano plot of DEGs was generated; (c): A heatmap of DEGs was created; (d): Acquisition of candidate genes; (e): GO enrichment analysis of candidate genes; (f): KEGG enrichment analysis of candidate genes; (g): GMCO network; (h) GMCK network; (i): PPI network of candidate

genes; (j): Bubble chart of metabolic pathway enrichment analysis.

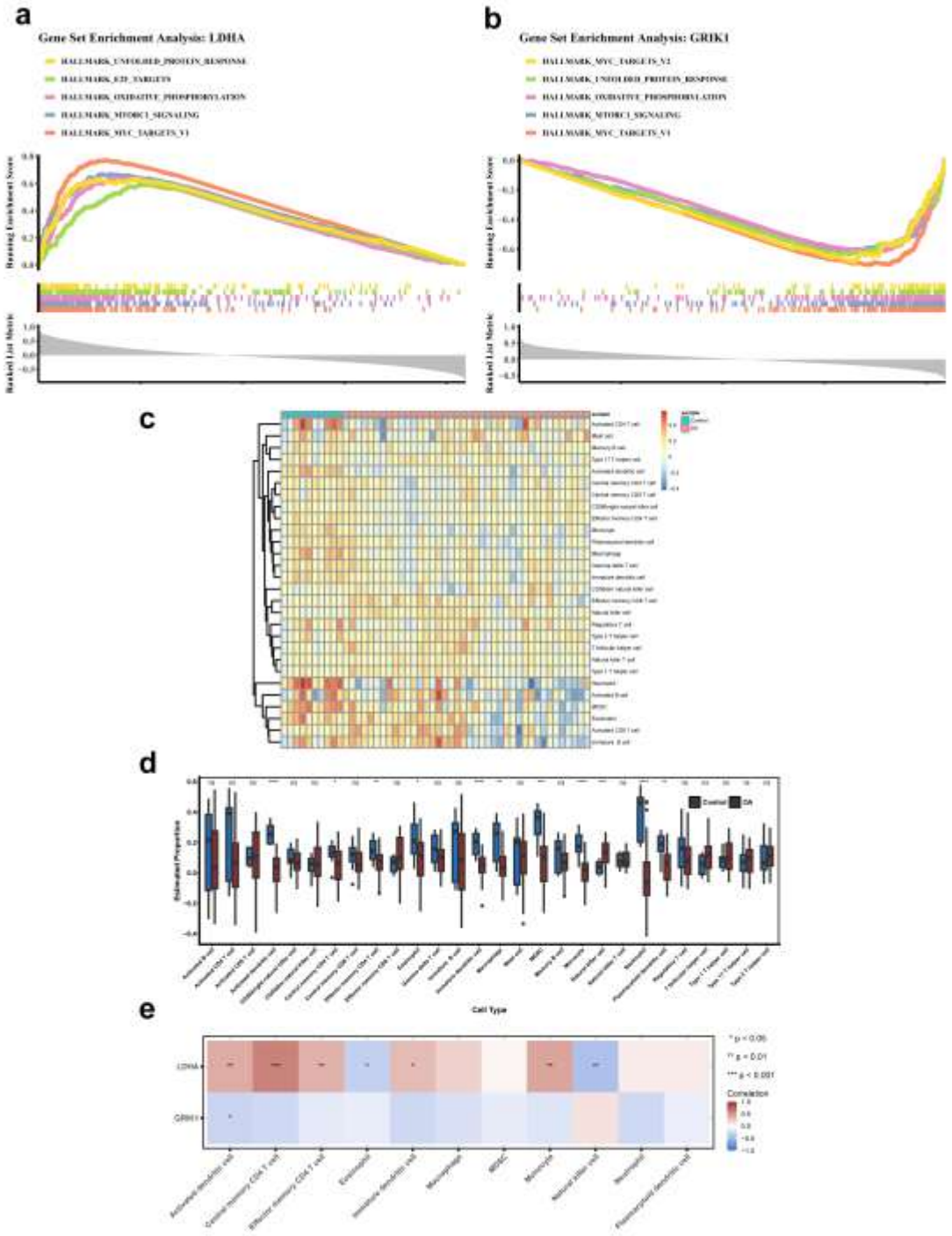
Figure 4. Acquisition of biomarkers:



(a): Results of XGBoost analysis were presented; (b-c): Results of RF analysis were shown; (d): A Venn diagram for key gene screening was displayed; (e): The expression profiles of key genes in the GSE51588

dataset were illustrated; (f): The expression profiles of key genes in the GSE114007 dataset were demonstrated; (g): The results of ROC analysis for candidate biomarkers in the GSE51588 dataset were described; (h): The results of ROC analysis for candidate biomarkers in the GSE114007 dataset were presented; (i): The BP neural network model in GSE51588 was depicted; (j): The ROC curve of the model in GSE51588 was shown; (k): The ROC curve of the model in GSE114007 was displayed.

Figure 5. A correlation was observed between biomarkers and differential immune cells



875

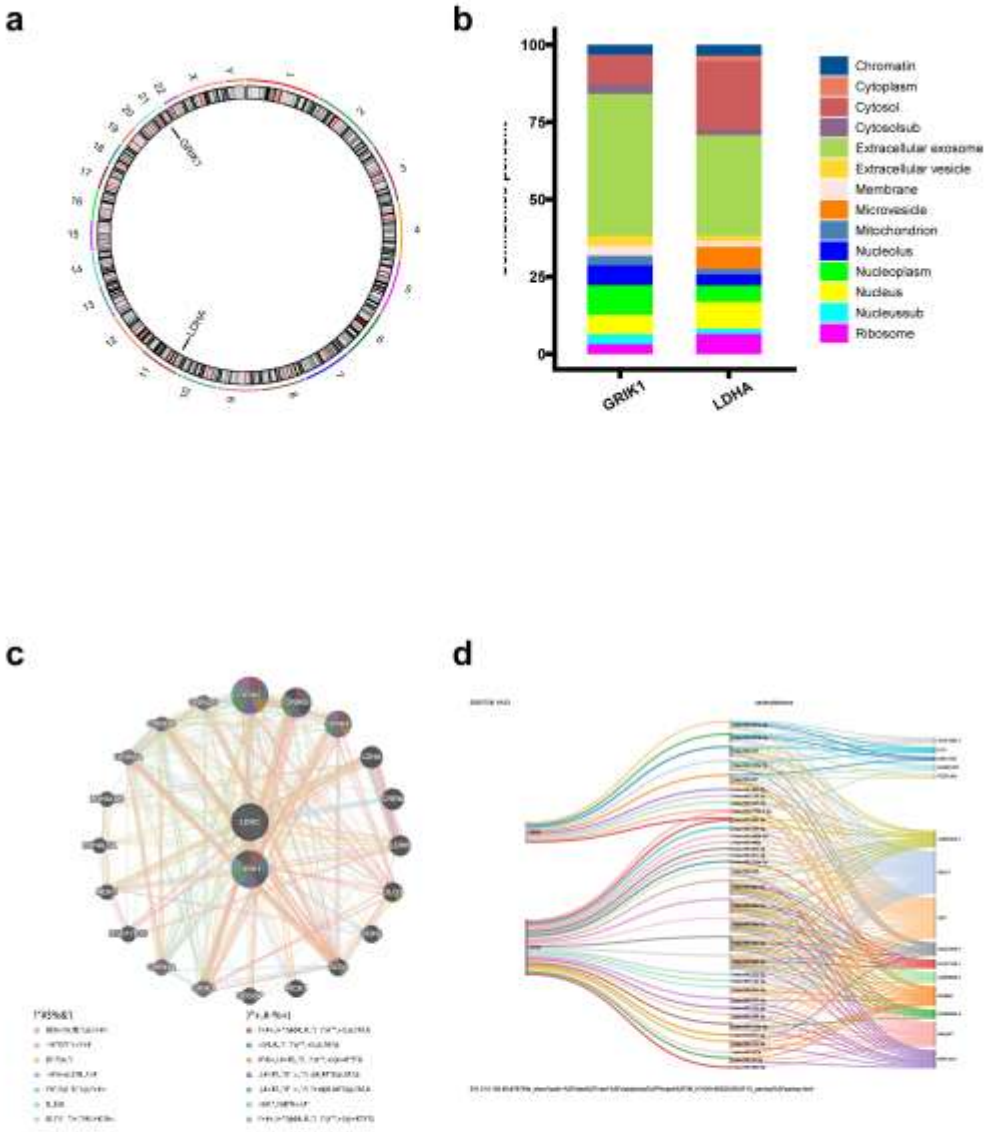
876 (a): GSEA enrichment analysis of LDHA was performed; (b): GSEA enrichment analysis of GRIK1 was

877 performed; (c): A heatmap of the inferred content of different cells via ssGSEA was generated; (d):

878 Differential immune cells in control and OA samples were identified; (e): Correlation analysis between

879 biomarkers and differential immune cells was conducted.

880 **Figure 6. The properties of biomarkers**



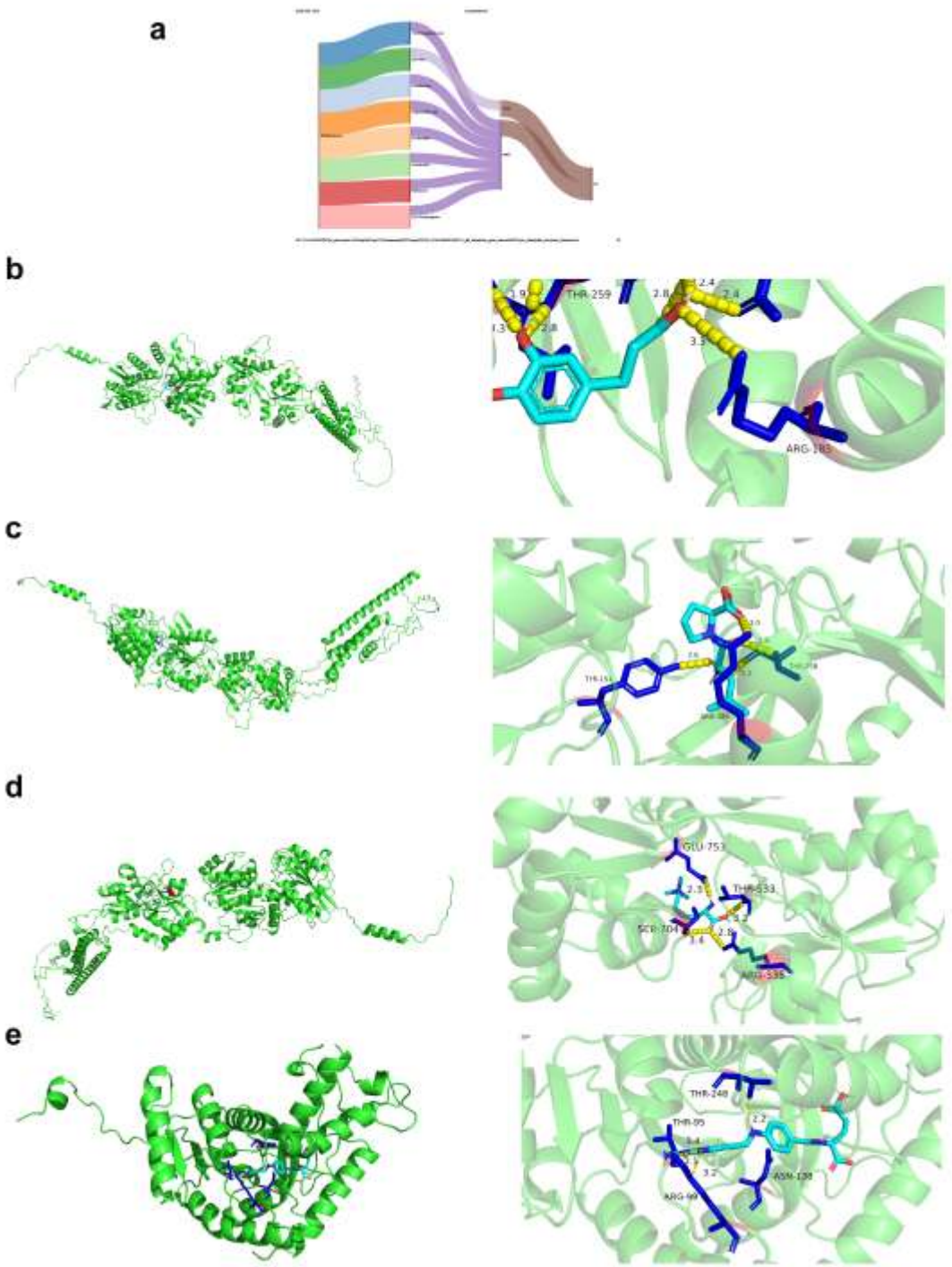
881

882 (a): The chromosome localizations of the biomarkers were identified; (b): The subcellular localizations

883 of the biomarkers were identified; (c): The GeneMANIA analysis of the biomarkers was carried out; (d):

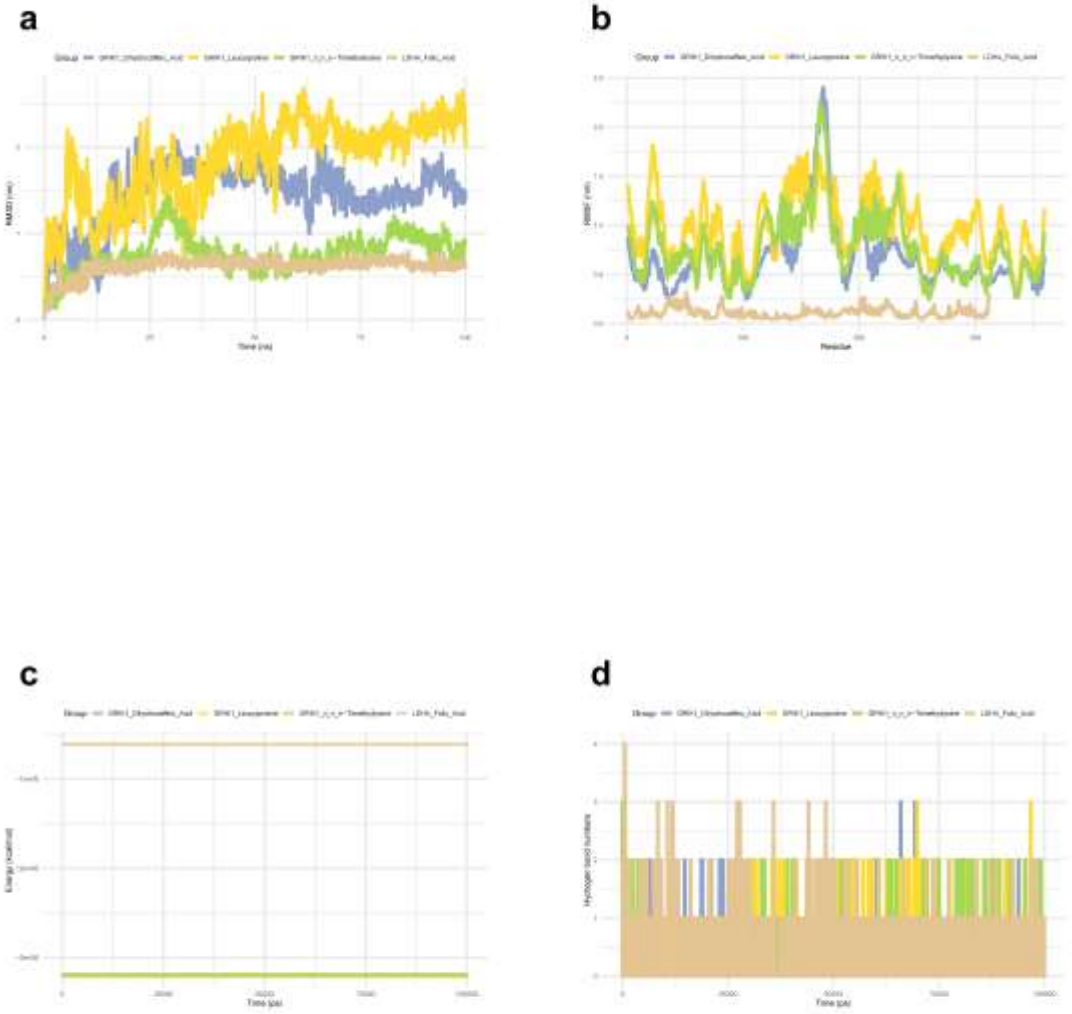
884 The lncRNA-miRNA-mRNA regulatory network was constructed.

Figure 7. Most metabolites were stably bound to the biomarkers



(a): GGBO network; (b): The molecular docking of GRIK1 with dihydrocaffeic acid; (c): The molecular docking of GRIK1 with leucylproline; (d): The molecular docking of GRIK1 with n,n,n-Trimethyllysine; (e): The molecular docking of LDHA with folic acid.

890 **Figure 8. Results of molecular dynamics simulation**



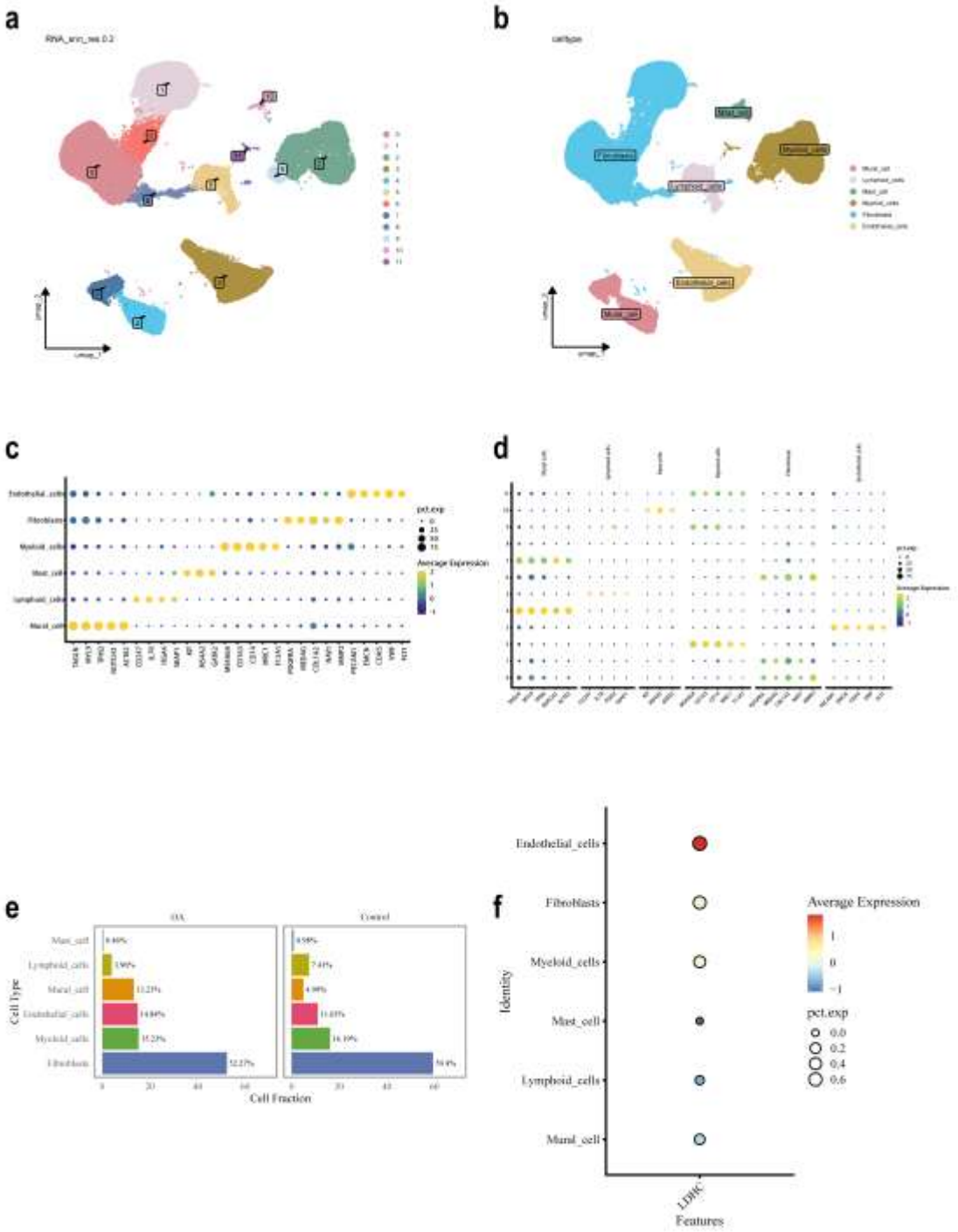
891

892 (a): The graph of RMSD analysis; (b): The graph of RMSF analysis; (c): The changes in energy during

893 molecular dynamics simulation; (d): The changes in hydrogen bonds during molecular dynamics

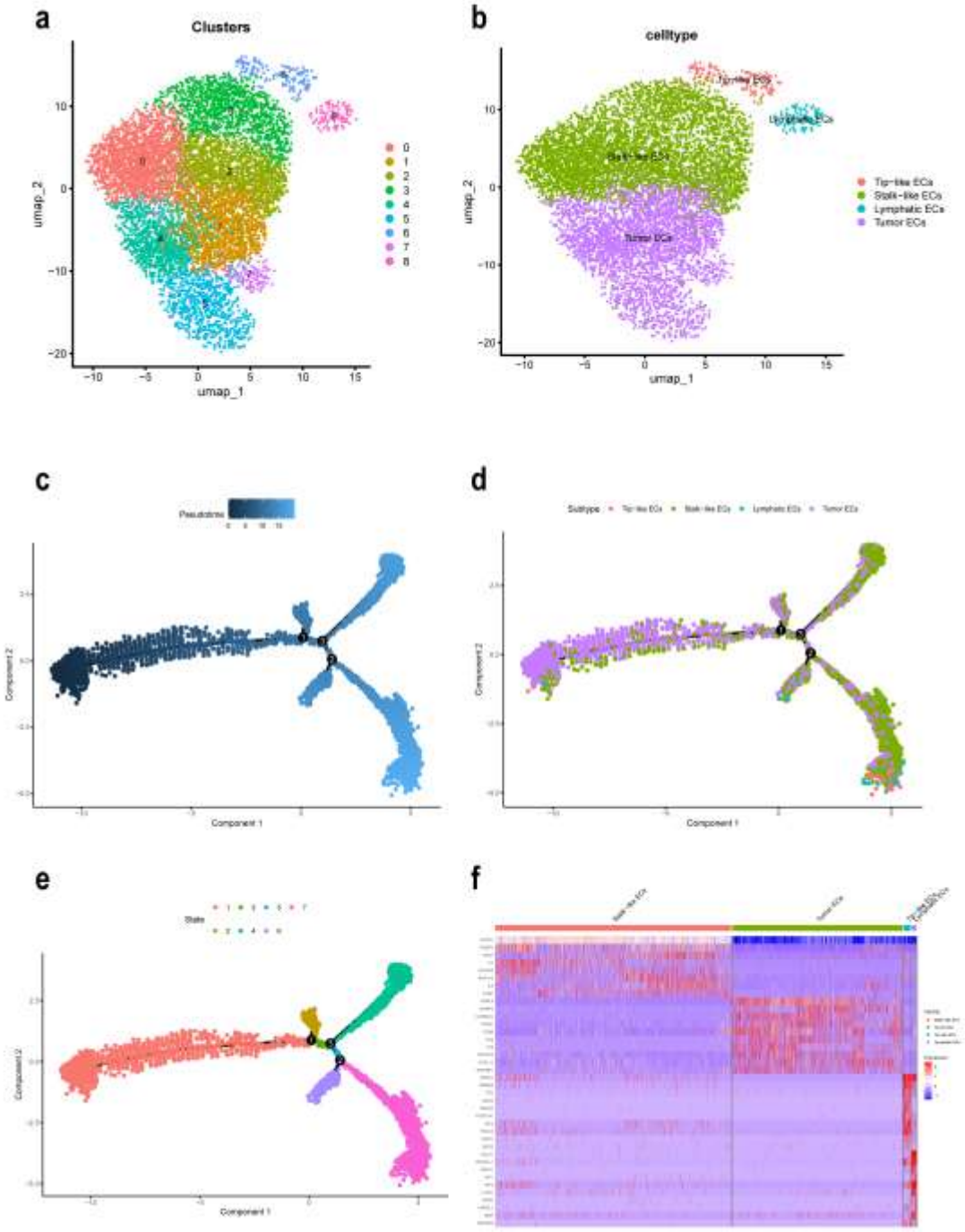
894 simulation.

Figure 9. Single-cell transcriptomic profiling and biomarker screening in OA



(a): UMAP clustering plot; (b): Cell-type annotation plot; (c-d): Marker expression heatmaps; (e): Cell proportion distribution; (f): Biomarker expression landscape in distinct cells.

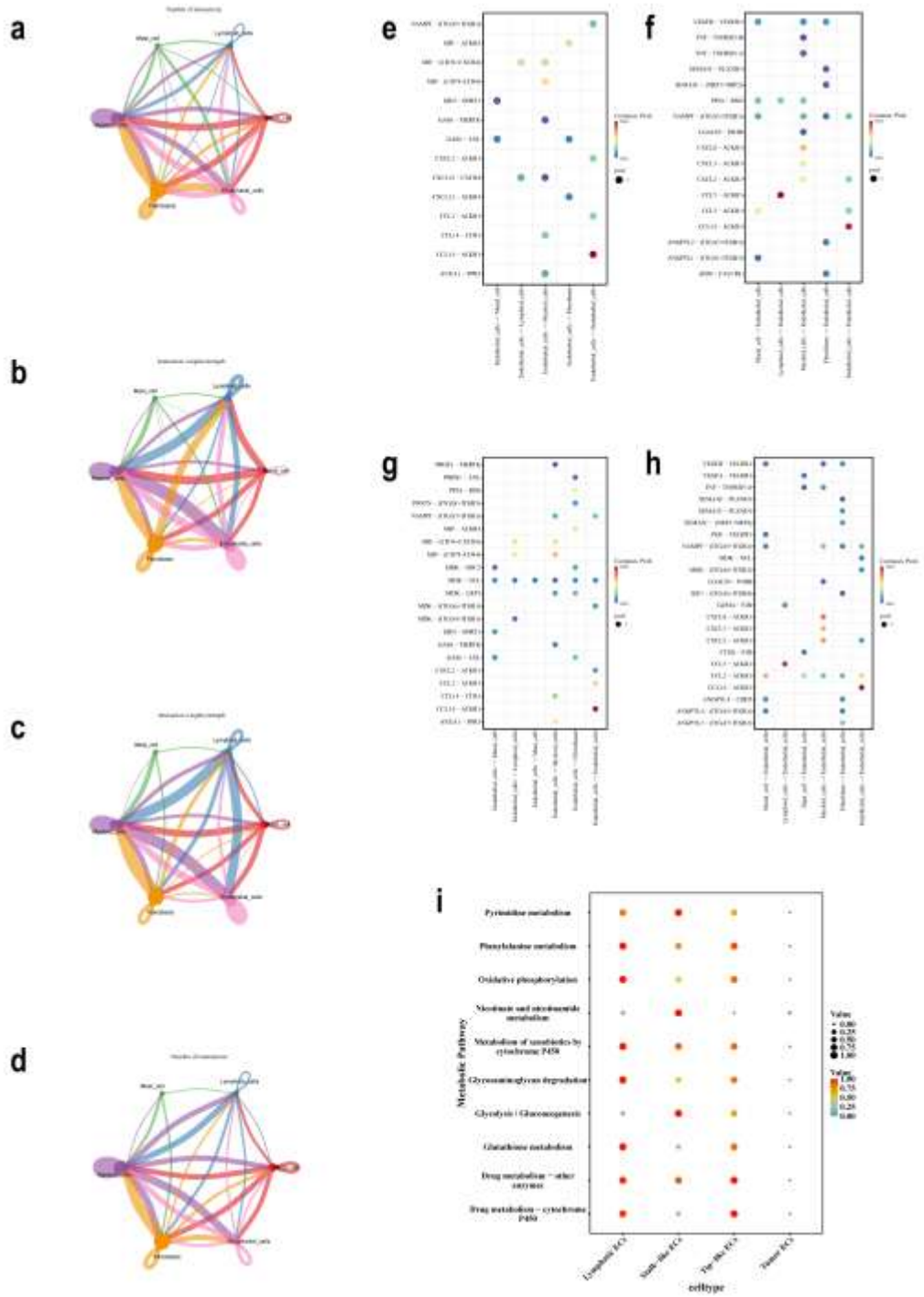
Figure 10. Subtype classification and pseudotime trajectory analysis of pivotal endothelial cells



(a): Secondary dimensionality reduction and clustering plot of pivotal cells; (b): Secondary cell annotation plot of pivotal cells; (c-e): Pseudotime trajectory analysis; (f): Differential gene expression plot of pivotal cell subsets.

Figure 11. Intercellular communication and metabolic profiling of endothelial cell subtypes in OA

and control groups



(a) Number of cell–cell communications in OA group; (b) Strength of cell-cell communications in OA group; (c) Number of cell-cell communications in control group; (d) Strength of cell-cell communications in control group; (e) Signal output of pivotal cells via ligand-receptor pairs (control); (f)

910 Signal input of pivotal cells via ligand-receptor pairs (control); (g) Signal output of pivotal cells via

911 ligand-receptor pairs (OA); (h) Signal input of pivotal cells via ligand-receptor pairs (OA); (i) Metabolic

912 pathway enrichment analysis.

913 **Tables**

SYMBOL	id.exposure	id.outcome	Pvalue	or
AKR1C3	eqtl-a-ENSG00000196139	ebi-a-GCST90038686	<0.001	1.001753285
EGLN3	eqtl-a-ENSG00000129521	ebi-a-GCST90038686	0.041	1.001948428
EPHA4	eqtl-a-ENSG00000116106	ebi-a-GCST90038686	0.003	1.00094728
FURIN	eqtl-a-ENSG00000140564	ebi-a-GCST90038686	<0.001	0.996939633
GPR84	eqtl-a-ENSG00000139572	ebi-a-GCST90038686	<0.001	0.997082666
GRIK1	eqtl-a-ENSG00000171189	ebi-a-GCST90038686	<0.001	0.997338561
ITGA2B	eqtl-a-ENSG00000005961	ebi-a-GCST90038686	0.039	1.00119748
LDHA	eqtl-a-ENSG00000134333	ebi-a-GCST90038686	0.004	0.997443301

MMP9	eqtl-a-ENSG00000100985	ebi-a-GCST90038686	0.049	1.001560507
NR3C1	eqtl-a-ENSG00000113580	ebi-a-GCST90038686	0.003	1.005329074
MAPKAPK2	eqtl-a-ENSG00000162889	ebi-a-GCST90038686	0.045	0.996367143
PTPN11	eqtl-a-ENSG00000179295	ebi-a-GCST90038686	<0.001	0.995692537

914

915 **Table 1:** There was a significant causal relationship between exposure and OA;

SYMBOL	id.exposure	egger_intercept	se	pval
AKR1C3	eqtl-a-ENSG00000196139	5.38E-05	0.000174304	0.759021225
EGLN3	eqtl-a-ENSG00000129521	0.000561905	0.000434572	0.218514801
EPHA4	eqtl-a-ENSG00000116106	-0.000247763	0.000186823	0.190925182
FURIN	eqtl-a-ENSG00000140564	0.000793417	0.000387057	0.053070876
GPR84	eqtl-a-ENSG00000139572	-0.000562977	0.000329174	0.096896286
GRIK1	eqtl-a-ENSG00000171189	-5.62E-05	0.00032495	0.86344001

ITGA2B	eqtl-a-ENSG00000005961	0.000324162	0.000243773	0.191958076
LDHA	eqtl-a-ENSG00000134333	0.000116847	0.000369568	0.755959805
MAPKAPK2	eqtl-a-ENSG00000162889	0.000551517	0.001236495	0.674225646
MMP9	eqtl-a-ENSG00000100985	-0.000612704	0.000332714	0.075450331
NR3C1	eqtl-a-ENSG00000113580	-0.002067043	0.002487189	0.422159121
PTPN11	eqtl-a-ENSG00000179295	-0.000489483	0.000308809	0.116997855

Table 2: Horizontal pleiotropy test;

id.exposure	id.outcome	snp_r2.exposure	snp_r2.outcome	correct_causal_direction	steiger_pval
AKR1C3	Osteoarthritis||id:ebi-a-GCST90038686	0.110502967	9.74E-07	TRUE	< 0.05
EPHA4		0.280594363	9.13E-07	TRUE	< 0.05
FURIN		0.029872877	5.99E-07	TRUE	1.44E-191
GPR84		0.053596003	8.14E-07	TRUE	< 0.05
GRIK1		0.07526346	1.51E-06	TRUE	< 0.05
LDHA		0.028001121	2.41E-07	TRUE	5.54E-146
NR3C1		0.007123246	2.58E-07	TRUE	2.44E-47
PTPN11		0.064285215	1.47E-06	TRUE	6.64E-194

Table 3: Steiger directionality test;

SYMBOL	id.exposure	method	Q	Q_df	Q_pval
AKR1C3	eqtl-a-ENSG00000196139	Inverse variance weighted(fixedeffects)	27.67903195	44	0.974094604
EGLN3	eqtl-a-ENSG00000129521	Inverse variance weighted(fixedeffects)	9.913070741	14	0.76851096
EPHA4	eqtl-a-ENSG00000116106	Inverse variance weighted(fixedeffects)	29.49789044	50	0.990767476
FURIN	eqtl-a-ENSG00000140564	Inverse variance weighted(fixedeffects)	13.23297033	22	0.92645952
GPR84	eqtl-a-ENSG00000139572	Inverse variance weighted(fixedeffects)	15.07435832	33	0.996827298
GRIK1	eqtl-a-ENSG00000171189	Inverse variance weighted(fixedeffects)	40.24153638	54	0.917895809
ITGA2B	eqtl-a-ENSG00000005961	Inverse variance weighted(fixedeffects)	63.75639182	37	0.0040578
LDHA	eqtl-a-ENSG00000134333	Inverse variance weighted(fixedeffects)	2.321873295	17	0.999989385
MAPKAPK2	eqtl-a-ENSG00000162889	Inverse variance weighted(fixedeffects)	1.952858717	6	0.923982685
MMP9	eqtl-a-ENSG00000100985	Inverse variance weighted(fixedeffects)	32.47992971	31	0.393770314

NR3C1	eqtl-a-ENSG00000113580	Inverse variance weighted(fixedeffects)	2.341630355	13	0.999454667
PTPN11	eqtl-a-ENSG00000179295	Inverse variance weighted(fixedeffects)	11.47170368	79	1

Table 4: Heterogeneity test;

id.exposure	id.outcome	Pvalue	or
class.Actinobacteria.id.419	ebi-a-GCST90038686	1.21E-05	1.004546661
family.Bifidobacteriaceae.id.433	ebi-a-GCST90038686	1.14E-06	1.004808215
genus.Bifidobacterium.id.436	ebi-a-GCST90038686	1.36E-05	1.004289608
order.Bifidobacteriales.id.432	ebi-a-GCST90038686	1.14E-06	1.004808215
phylum.Actinobacteria.id.400	ebi-a-GCST90038686	0.041638626	1.003451824

Table 5: There was a significant causal relationship between exposure and OA;

id.exposure	id.outcome	egger_intercept	se	pval
class.Actinobacteria.id.419	Osteoarthritis||id:ebi-a-GCST90038686	0.000819917	0.000720822	0.26087153
family.Bifidobacteriaceae.id.433		0.000739008	0.000697189	0.294245877

genus.Bifidobacterium.id.436	0.000843555	0.000681478	0.221676875
order.Bifidobacteriales.id.432	0.000739008	0.000697189	0.294245877
phylum.Actinobacteria.id.400	0.002455423	0.00196925	0.224989438

Table 6: Horizontal pleiotropy test;

id.exposure	id.outcome	snp_r2.exposure	snp_r2.outcome	correct_causal_direction	steiger_pval
class.Actinobacteria.id.419	Osteoarthritis||id:ebi-a-GCST90038686	0.15972392	0.000119436	TRUE	< 0.05
family.Bifidobacteriaceae.id.433		0.161353141	0.000131357	TRUE	< 0.05
genus.Bifidobacterium.id.436		0.157553827	0.000111792	TRUE	< 0.05
order.Bifidobacteriales.id.432		0.161353141	0.000131357	TRUE	< 0.05

Table 7: Steiger directionality test;

id.exposure	id.outcome	method	Q	Q_df	Q_pval
class.Actinobacteria.id.419	Osteoarthritis||id:ebi-a-GCST90038686	Inversevariance weighted	32.57339164	50	0.973260569
family.Bifidobacteriaceae.id.433		Inversevariance weighted	31.43304364	51	0.98585456
genus.Bifidobacterium.id.436		Inversevariance weighted	29.13843495	50	0.991976084
order.Bifidobacteriales.id.432		Inversevariance weighted	31.43304364	51	0.98585456

phylum.Actinobacteria.id.400	Inversevariance weighted	14.25266838	24	0.940736245

928

929 **Table 8:** Heterogeneity test;

Compound	MW	ConsensusLogP	H-bondacceptors	H-bonddonors	TPSA
7-Aminoheptanoic acid	145.2	0.14	3	2	63.32
D-Methionine	149.21	-0.61	3	2	88.62
Dihydrocaffeic acid	182.17	0.63	4	3	77.76
DL-Norvaline	117.15	-0.71	3	2	63.32
Folic Acid	441.4	-0.36	9	6	213.28
Leucylproline	228.29	0.04	4	2	83.63
Methionine	149.21	-0.59	3	2	88.62
n,n,n-Trimethyllysine	189.28	-2.43	3	2	63.32

930

931 **Table 9:** Drug similarity prediction results;

Compound	hERG	DILI	logP	logD	CYP3A4-inh
7-Aminoheptanoic acid	0.050634261	0.08756049	-2.47372151	-0.51319351	4.29E-07
D-Methionine	0.030217214	0.007744119	-0.872964061	0.131298837	1.85E-08
Dihydrocaffeic acid	0.028109441	0.296994269	0.958225869	1.118220454	7.83E-06
DL-Norvaline	0.067193203	0.023832843	-1.908756163	-0.489334986	2.12E-07
Folic Acid	0.022792693	0.999894619	-1.761076432	0.585902745	1.77E-10
Leucylproline	0.066699393	0.65387398	-0.814685999	-2.58E-05	8.51E-06
Methionine	0.038352817	0.02846921	-1.744853977	-0.374448837	2.54E-08

n,n,n-Trimethyllysine	0.116341867	3.37E-05	-2.403245236	-0.93970919	4.27E-11

Table 10: Physical property prediction results;

biomarkers	protein conformation ID	Core metabolites	binding energy
GRIK1	AF-P39086-F1	7-Aminoheptanoic acid	-4.5
GRIK1	AF-P48443-F1	D-Methionine	-5
GRIK1	AF-P46439-F1	Dihydrocaffeic acid	-6.4
GRIK1	AF-P39086-F1	DL-Norvaline	-4.8
LDHA	AF-P00338-F1	Folic Acid	-8.2
GRIK1	AF-P39086-F1	Leucylproline	-6.5
GRIK1	AF-P39086-F1	Methionine	-4.4
GRIK1	AF-P39086-F1	n,n,n-Trimethyllysine	-5.6

Table 11: Binding energy of core metabolites and biomarkers;

Online Resources

Supplementary Figure 1: OA-related DRGs scatter plots were plotted;

Supplementary Figure 2: OA-related DRGs forest plots were plotted;

Supplementary Figure 3: OA-related DRGs funnel plots were plotted;

Supplementary Figure 4: OA-related DRGs LOO plots were plotted;

Supplementary Figure 5: OA-related GMs scatter plots were plotted;

Supplementary Figure 6: OA-related GMs forest plots were plotted;

Supplementary Figure 7: OA-related GMs funnel plots were plotted;

Supplementary Figure 8: OA-related GMs LOO plots were plotted;

Supplementary Figure 9： Single-Cell RNA-seq Data Quality Control and Dimensionality Reduction Analysis;

Supplementary Table 1: DRGs;

Supplementary Table 2: SNPs for MR analysis of OA and DRGs were selected;

Supplementary Table 3: SNPs for MR analysis of OA and GMs were selected;

Supplementary Table 4: OA-related metabolites were identified;

Supplementary Table 5: GRGs;

Supplementary Table 6: The results of differential expression analysis;

Supplementary Table 7: Candidate genes;

Supplementary Table 8: The results of GO enrichment analysis;

Supplementary Table 9: The results of KEGG enrichment analysis;

Supplementary Table 10: Results of metabolic pathway enrichment analysis;

Supplementary Table 11: Key genes;

Supplementary Table 12: The results of GSEA enrichment analysis for LDHA;

Supplementary Table 13: The results of GSEA enrichment analysis for GRIK1;

Supplementary Table 14: Correlation between biomarkers and differential immune cells;

Supplementary Table 15: lncRNA-miRNA-mRNA network.

Supplementary Files

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

- OnlineResources.zip