ACCSCIENCE PUBLISHING

Eurasian Journal of Medicine and Oncology

ORIGINAL RESEARCH ARTICLE

Identification and validation of relevant diagnostic biomarkers for osteoporosis by Weighted Gene Co-expression Network Analysis and machine learning

Cuicui Zhou[1,2], Zarina Awang[1], and Farra Aidah Jumuddin[1]*

[1]Department of Clinical Medicine, Faculty of Medicine, Lincoln University College, Petaling Jaya, Selangor, Malaysia

[2]Department of Orthopaedic Surgery, The Second Affiliated Hospital of Nanyang Medical College, Nanyang, Henan, China

Abstract

Introduction: Osteoporosis (OP) is a systemic metabolic bone disease characterized by complex pathogenesis and high prevalence. Current diagnostic and therapeutic approaches have limited effectiveness, and new biomarkers are needed to improve the treatment and diagnosis of OP.

Objective: The present study aimed to identify novel diagnostic biomarkers for OP through integrated bioinformatics analysis.

Methods: We performed an integrative bioinformatics analysis combining Weighted Gene Co-expression Network Analysis and machine learning on two Gene Expression Omnibus datasets (GSE35958, GSE35956). Differentially expressed genes (DEGs) were identified using "limma" package of R software, followed by module construction and key gene screening via Least Absolute Shrinkage and Selection Operator (LASSO) regression. Functional enrichment, immune infiltration, and drug prediction analyses were conducted to explore biological mechanisms and therapeutic potential.

Results: Differential expression analysis identified 1,020 DEGs, from which 10 co-expression modules were constructed. The blue module demonstrated the strongest correlation with OP ($r = 0.99$, $p<0.0001$). LASSO regression analysis prioritized seven candidate genes (LOC286177, nucleobindin 1 [NUCB1], peroxisomal biogenesis factor 19 [PEX19], metastasis associated 1 [MTA1], DRA aassociated protein 1 [DRAP1], protocadherin gamma A1 [PCDHGA1], and pre-mRNA processing factor 39 [PRPF39]), with subsequent validation confirming NUCB1, PEX19, MTA1, DRAP1, and PCDHGA1 as robust diagnostic biomarkers (Area under the curve > 0.85). Functional enrichment implicated these genes in endoplasmic reticulum stress, Wnt/β-catenin signaling, and immune regulatory pathways. Immune profiling further revealed significant perturbations in T-cell and macrophage populations in OP. The Coremine Medical database was leveraged to predict potential therapeutic agents, including both small-molecule and phytochemical candidates.

Conclusion: The present study identified NUCB1, PEX19, MTA1, DRAP1, and PCDHGA1 as promising OP diagnostic markers and explored their roles in bone metabolism. The findings offer insights for early diagnosis and targeted therapy but require further clinical validation.

Keywords: Osteoporosis; Gene Expression Omnibus; Weighted Gene Co-expression Network Analysis; Least Absolute Shrinkage and Selection Operator; Biomarker

***Corresponding author:**
Farra Aidah Jumuddin
(farraaiadah@lincoln.edu.my)

Citation: Zhou C, Awang Z, Jumuddin FA. Identification and validation of relevant diagnostic biomarkers for osteoporosis by Weighted Gene Co-expression Network Analysis and machine learning. *Eurasian J Med Oncol.* 2025;9(3):261-276.
doi: 10.36922/EJMO025240252

Received: June 09, 2025

Revised: July 25, 2025

Accepted: July 29, 2025

Published online: September 9, 2025

1. Introduction

Osteoporosis (OP) is a systemic metabolic bone disorder characterized by decreased bone mineral density (BMD) and microarchitectural deterioration, predisposing patients to fragility fractures at high-risk sites including the hip, spine, and wrist.[1] These fractures frequently lead to chronic pain, functional impairment, and irreversible skeletal damage, collectively contributing to reduced quality of life.[2-4] OP currently affects approximately 200 million individuals globally, with a prevalence of 30% in women and 20% in men aged over 50 years.[5] Furthermore, demographic aging is projected to escalate the annual healthcare costs of osteoporotic fractures to hundreds of billions of USD by 2050.[6] Together, these pathophysiological and epidemiological features establish OP as a major global public health challenge requiring urgent intervention.

The pathogenesis of OP is multifactorial, with immune and inflammatory responses playing pivotal roles in disrupting bone homeostasis.[5] Chronic inflammatory states activate the nuclear factor kappa B (NF-κB) signaling pathway and promote the release of pro-inflammatory cytokines such as interleukin-6 (IL-6) and tumor necrosis factor-alpha (TNF-α), which enhance osteoclast activity and suppress osteoblast function, thereby accelerating bone loss.[6,7] In addition, genetic predisposition, hormonal changes, nutritional status, and lifestyle factors contribute significantly to the development of OP. [8-10] In recent years, therapeutic strategies for OP have primarily included bisphosphonates, selective estrogen receptor modulators (SERMs), and monoclonal antibodies. While these agents effectively inhibit bone resorption, their long-term use may be associated with adverse effects and fails to fully restore bone microstructure and function.[11,12] Moreover, the absence of definitive diagnostic criteria and targeted therapies complicates early intervention and personalized treatment.[13] Therefore, identifying diagnostic biomarkers holds substantial clinical significance for early screening, risk prediction, and the discovery of therapeutic targets in OP.[14]

Bioinformatics approaches have become instrumental in elucidating core disease mechanisms. Weighted Gene Co-expression Network Analysis (WGCNA) enables the systematic identification of disease-associated modules and hub genes from high-throughput data, with demonstrated applications in OP and other metabolic disorders.[15,16] As a classical machine learning method, Least Absolute Shrinkage and Selection Operator (LASSO) regression employs L1 regularization to enhance model interpretability by selecting the most predictive features while reducing overfitting.[17] This approach has been widely adopted in biomarker discovery due to its ability to handle high-dimensional omics data and identify robust disease-associated signatures.[17] In particular, LASSO regression has demonstrated scientific reliability in screening biomarkers for complex diseases, including oncology, cardiovascular disorders, and, more recently, bone metabolic diseases.[18,19] In the present study, we aim to combine WGCNA and LASSO regression to systematically analyze gene expression data related to OP, identifying and validating diagnostic biomarkers and potential drug targets. Through this research, we seek to provide novel scientific insights and references for the clinical diagnosis and treatment of OP. The analysis flowchart is shown in Figure 1.

2. Materials and methods

2.1. Sample sources

Transcriptome profiles were retrieved from the NCBI Gene Expression Omnibus (GEO) repository via the GEOquery R package (v2.66.0). Two datasets generated on the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) were selected to serve as discovery and validation cohorts, respectively: GSE35958, which included four healthy controls and five OP cases, and GSE35956, used as the validation set with five controls and five cases.

Raw CEL files were downloaded and subjected to stringent quality control using the affyQC-Report R package (v1.72.0), with no samples excluded on the basis of normalized unscaled standard errors (>1.05) or relative log expression dispersion criteria. Probe-level data were background-corrected, quantile-normalized, and log2-transformed using the affy and "limma" packages. Non-specific filtering was applied to retain only probes with an average intensity ≥4 in at least 20% of arrays. Probe sets were collapsed to unique Entrez gene symbols using the annotate package (max mean probe per gene). The two datasets were processed independently to preserve their mutual independence for downstream machine learning modeling and subsequent enrichment analyses.

2.2. Differential analysis

To identify differentially expressed genes (DEGs) associated with OP, we conducted a comprehensive analysis of the GSE35958 and GSE35956 datasets utilizing the "limma" package in R. The criteria for selecting DEGs were established based on a stringent threshold of |log fold change (FC)| > 0.5 and a significance level of $p<0.05$. Specifically, genes with log FC > 0.5 and $p<0.05$ were classified as upregulated, whereas those with log FC < −0.5 and $p<0.05$ were categorized as downregulated. For visualization, we generated heatmaps using the "pheatmap" package, and volcano plots were constructed using "ggplot2" to illustrate the distribution of DEGs.

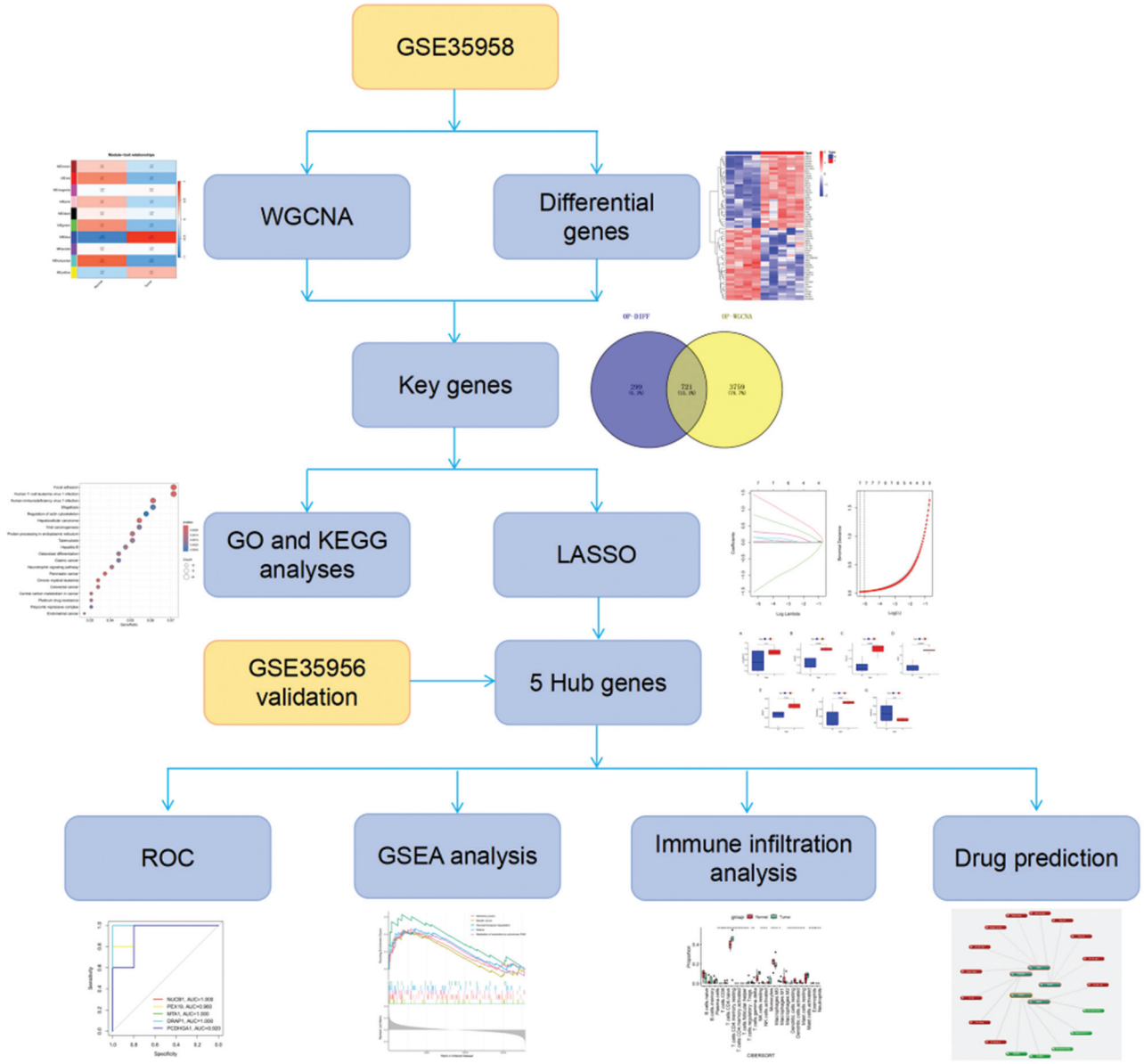

Figure 1. The flowchart of the integrative bioinformatics analysis
Abbreviations: GO: Gene ontology; GSEA: Gene Set Enrichment Analysis; KEGG: Kyoto encyclopedia of genes and genomes; LASSO: Least Absolute Shrinkage and Selection Operator; ROC: Receiver operating characteristic; WGCNA: Weighted Gene Co-expression Network Analysis.

2.3. WGCNA

Co-expression network analysis was conducted using the WGCNA package (v1.72) in R on the GSE35958 expression dataset. The adjacency matrix was constructed with a soft-threshold power (β) of 7, determined by scale-free topology fitting index ($R^2 > 0.85$). Network topology was calculated using unsigned topological overlap matrices (TOMType = "unsigned") with the following parameters: minModuleSize = 50 to ensure biological relevance, mergeCutHeight = 0.6 for module consolidation, and recreateThreshold = 0 to maintain original clustering. Non-expressed genes (standard deviation = 0) were excluded from the analysis. Module-trait relationships were assessed through Pearson's correlation, and modules with significant associations ($|r| > 0.5$, false discovery rate [FDR]-adjusted $p<0.01$) were selected for downstream biomarker analysis.[15] Finally, DEGs were intersected with genes contained within the key WGCNA modules using the "Venn" package in R. The intersected genes were analyzed in the next step of the study.

2.4. Enrichment analysis

To systematically dissect the biological relevance of the identified gene lists, we conducted gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses. Gene symbols were first converted to Entrez IDs using the org.Hs.eg.db annotation package (v3.17.0). The clusterProfiler package (v4.8.1) was then used to query three GO domains—biological process (BP), molecular function, and cellular component (CC)—as well as the KEGG PATHWAY database (release 104.0). Enrichment analysis was performed using the enrichGO and enrichKEGG functions, with a p-value cutoff of 0.05 after Benjamini–Hochberg FDR correction, and gene sets restricted to sizes between 10 and 500 genes. Redundant GO terms were collapsed with the simplify function (q-value cutoff = 0.05). To quantify the enrichment strength, normalized enrichment scores (NES) and gene ratios were extracted for each term.

The resulting GO and KEGG terms were visualized using the "ggplot2" (v3.4.2) and "enrichplot" (v1.18.3) packages. Bubble plots and cnet plots were generated to depict the top 20 most significantly enriched terms, ranked by adjusted p-value.

2.5. Screening and validation of key genes through machine learning

To identify the minimal set of genes that robustly discriminates OP from healthy bone, we applied LASSO penalized logistic regression to the GSE35958 discovery cohort (n = 9). Prior to modeling, the expression matrix was variance-stabilized using the voom transformation and z-score normalized across samples. The optimal λ value was determined by 10-fold cross-validation (CV) with 1,000 repetitions, selecting the λ within one standard error of the minimum mean cross-validated deviance. Genes retaining non-zero regression coefficients at the optimal λ were designated as core genes. All LASSO analyses were implemented using the glmnet package (v4.1-8) in R (v4.3.0).

For internal validation, expression levels of the core genes in GSE35958 dataset were visualized using box plots and dot plots generated with the "ggpubr" package (v0.6.0). For external validation, the same core gene set was projected onto the independent GSE35956 dataset (n = 10). Differential expression between OP and control groups was quantified by the two-sided Wilcoxon rank-sum test and corrected for multiple comparisons using the Benjamini–Hochberg (FDR < 0.05).

To assess the discriminatory capacity of the core gene signatures, receiver operating characteristic (ROC) curves were constructed for both datasets using the "pROC" package (v1.18.4). The area under the ROC curve (area under the curve [AUC]) and the corresponding 95 % confidence intervals were computed by DeLong's method. An AUC >0.85 was considered indicative of excellent predictive accuracy.

2.6. Gene Set Enrichment Analysis (GSEA)

GSEA was performed to systematically explore the potential BPs and signaling pathways associated with the hub gene of interest. The entire transcriptome was first ranked according to the Pearson's correlation coefficient between the hub gene and all other expressed genes. Next, singlegene GSEA was conducted with the "GSEA" function implemented in the clusterProfiler R package against the Molecular Signatures Database (MSigDB, v7.x) "c2.cp.kegg" and "c5.go" collections. To ensure statistical robustness, 10,000 phenotype-based permutations were executed, and gene set sizes were restricted to between 10 and 200 genes. Pathways with a ($|NES|$) ≥ 1 and an adjusted p-value (FDR) <0.05 were considered significantly enriched. The leading-edge genes driving each enriched gene set were extracted, and the results were visualized using the "gseaplot2" function from the "enrichplot" R package.

2.7. Immune infiltration analysis

To quantify the landscape of tumor-infiltrating immune cells, we applied "CIBERSORT" R package calculation to perform *de novo* enumeration of 22 human immune cell subsets through a ν-support vector regression algorithm trained on the LM22 signature matrix, which consisted of 547 genes. The analysis was carried out using default parameters: 1,000 permutations, quantile normalization disabled, and batch correction activated. Samples with a deconvolution p<0.05 were retained for downstream analysis to ensure high fidelity of the inferred fractions.

Subsequently, the relative proportions of immune cell subsets, such as $CD8^+$ T cells, M2 macrophages, and regulatory T cells, were compared between normal and OP samples using a two-sided Wilcoxon rank-sum test with Benjamini–Hochberg correction (FDR < 0.05). Compositional shifts were visualized using the "ggplot2" and "pheatmap" R packages, whereas intercellular correlation networks ($|r|$ > 0.4 and p<0.05) were constructed using the "igraph" package to explore potential immune cell crosstalk.

2.8. Potential drug prediction

The OP core genes (Entrez IDs) were imported into Coremine Medical (http://www.coremine.com/medical), where the platform's text-mining engine was queried under the "herb–gene–disease" relationship layer to retrieve botanical entities statistically linked to both the gene set

and the MeSH term "osteoporosis." Statistical significance was assessed by a one-tailed Fisher's exact test and corrected for multiple comparisons using the Benjamini–Hochberg method; an adjusted $p<0.05$ was required for retention. For CV, the Latin binomials of the prioritized herbs were subsequently queried in the HERB database (http://herb.ac.cn) to confirm their documented therapeutic relevance to OP. Only records that were manually curated from peer-reviewed publications within the last decade were retained, ensuring the highest level of evidence.

3. Results

3.1. Identification of DEGs

In the study, we first performed batch effect correction and normalization on the raw data extracted from the GEO dataset GSE35958. The results showed that the distribution of expression profiles of all the samples generally converged after batch correction and log-normalization (Figure 2A). Subsequently, differential expression analysis was conducted between the control and OP groups, resulting

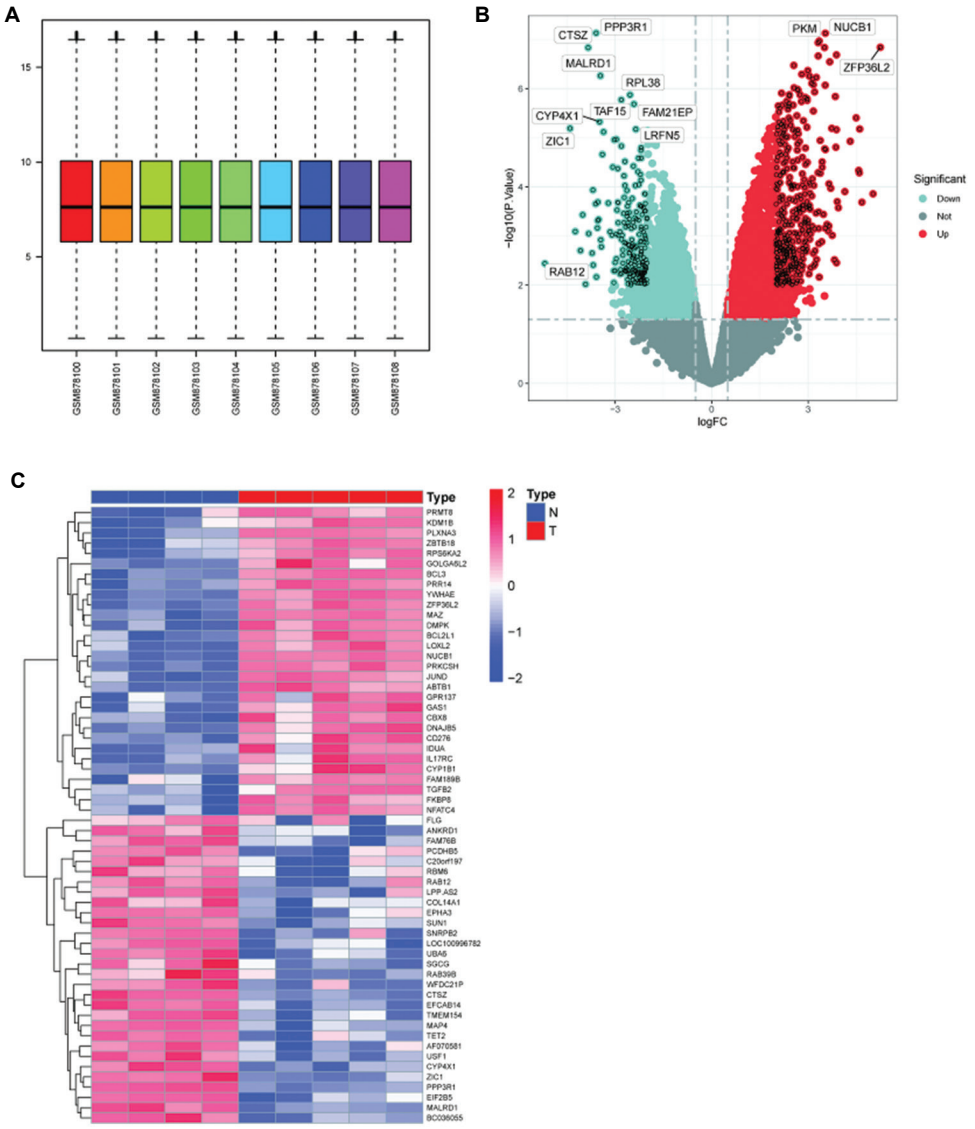

Figure 2. Identification of DEGs. (A) Normalized gene expression data from the GSE35958 dataset. (B) Volcano plot of DEGs in GSE35958. Red dots indicate upregulated genes, blue dots indicate downregulated genes, and gray dots indicate genes without significant differential expression. (C) Heatmap of DEGs in the GSE30528 dataset.
Abbreviations: DEGs: Differentially expressed genes. Group description: N, normal group; T, test group comprising osteoporosis (OP) patients.

in the identification of a total of 1,020 DEGs. In addition, a volcano plot (Figure 2B) and a heatmap (Figure 2C) were then generated to visualize the distribution of DEGs.

3.2. Construction of weighted gene co-expression networks

To investigate the association among OP-associated DEGs, we constructed a weighted gene co-expression network using the WGCNA package. The optimal soft-threshold power was determined to be 7 (Figure 3A), and a total of 10 distinct co-expression modules were identified (Figure 3B-D). Among these, the black module exhibited the strongest correlation with OP (correlation coefficient = 0.99, $p<0.0001$), establishing it as the key module. Subsequent intersection analysis between the OP-associated DEGs and the genes within the WGCNA module revealed 721 core candidate genes (Figure 3E).

3.3. Functional enrichment analysis

Subsequently, we performed GO and KEGG enrichment analysed on the 721 key genes. In the BP assessment, they were mainly associated with functions such as regulation of response to endoplasmic reticulum (ER) stress and small GTPase-mediated signal transduction. In the CC assessment, they were mainly enriched in the ER-Golgi intermediate compartment and the RNA polymerase II transcription regulator complex. In the MM assessment, they were mainly associated with DNA-binding transcription factor binding and chromatin DNA binding (Figure 4A). KEGG pathway analysis revealed significant enrichment in pathways including human T-cell leukemia virus 1 infection, human immunodeficiency virus 1 infection, and focal adhesion (Figure 4B-D).

3.4. Machine learning screening and expression validation of pivotal genes

To further screen the pivotal genes, we analyzed 721 key genes using the LASSO regression algorithm, identifying seven pivotal genes, namely LOC286177, nucleobindin 1 (NUCB1), peroxisomal biogenesis factor 19 (PEX19), metastasis associated 1 (MTA1), DRA aassociated protein 1 (DRAP1), protocadherin gamma A1 (PCDHGA1), and pre-mRNA processing factor 39 (PRPF39) (Figure 5A and B).

Next, we validated these seven hub genes in both the GSE35958 dataset and an independent OP dataset, GSE35956. In the GSE35958 dataset, the expression levels of LOC286177, NUCB1, PEX19, MTA1, DRAP1, and PCDHGA1 were significantly elevated in the OP group compared with normal controls, while PRPF39 expression was significantly decreased (Figure 6A-G). In the GSE35956 dataset, consistent elevation of NUCB1, PEX19, MTA1, DRAP1, and PCDHGA1 was also observed in the OP group, whereas LOC286177 and PRPF39 shosed no significant differential expression (Figure 7A-G). Therefore, subsequent analyses focused on the five validated genes, including NUCB1, PEX19, MTA1, DRAP1, and PCDHGA1.

3.5. ROC curve analysis

Next, we performed ROC curve analysis to evaluate the diagnostic performance of the five hub genes. In the GSE35958 dataset, all five genes such as NUCB1, PEX19, MTA1, DRAP1, and PCDHGA1 demonstrated excellent diagnostic accuracy, each achieving an AUC of 1.000 (Figure 8A). In the GSE35956 dataset, NUCB1, MTA1, and DRAP1 also achieved AUCs of 1.000, while PEX19 and PCDHGA1 showed AUCs of 0.960 and 0.920, respectively (Figure 8B). In both datasets, the AUC values of all five genes were >0.7, suggesting strong diagnostic potential.

3.6. GSEA

To explore the potential pathways of these five hub genes, we performed single-gene GSEA. The results showed that the high-expression group of NUCB1 was highly enriched in the galactose metabolism, starch and sucrose metabolism pathways (Figure 9A). The high-expression group of PEX19 was highly enriched in the adherens junction and metabolism of xenobiotics by cytochrome P450 pathways (Figure 9B). Both MTA1 and DRAP1 high expression groups were highly enriched in phenylalanine metabolism and SNARE interactions in vesicular transport pathways (Figure 9C and D). The high-expression group of PCDHGA1 was highly enriched in glycosaminoglycan degradation and vasopressin-regulated water reabsorption pathways (Figure 9E).

3.7. Immune infiltration analysis

To analyze the differences in immune levels between osteoporotic and normal conditions, we analyzed immune cell infiltration in the control and OP groups using the GSE35958 dataset. The results showed that T cell subsets and macrophage subsets were the major subpopulations of immune cells in both groups (Figure 10A). Notably, the expression levels of $CD4^+$ T cells, regulatory T cells, M0, M1, and M2 macrophages, and several other immune cell types were significantly altered in OP samples compared to normal samples (Figure 10B).

3.8. Potential targeted drug prediction and identification of core components

To explore potential therapeutic drugs for OP, we predicted potential Chinese or Western therapeutic drugs for these five genes by screening them with $p<0.05$ as a criterion through Coremine Medical database. The results showed

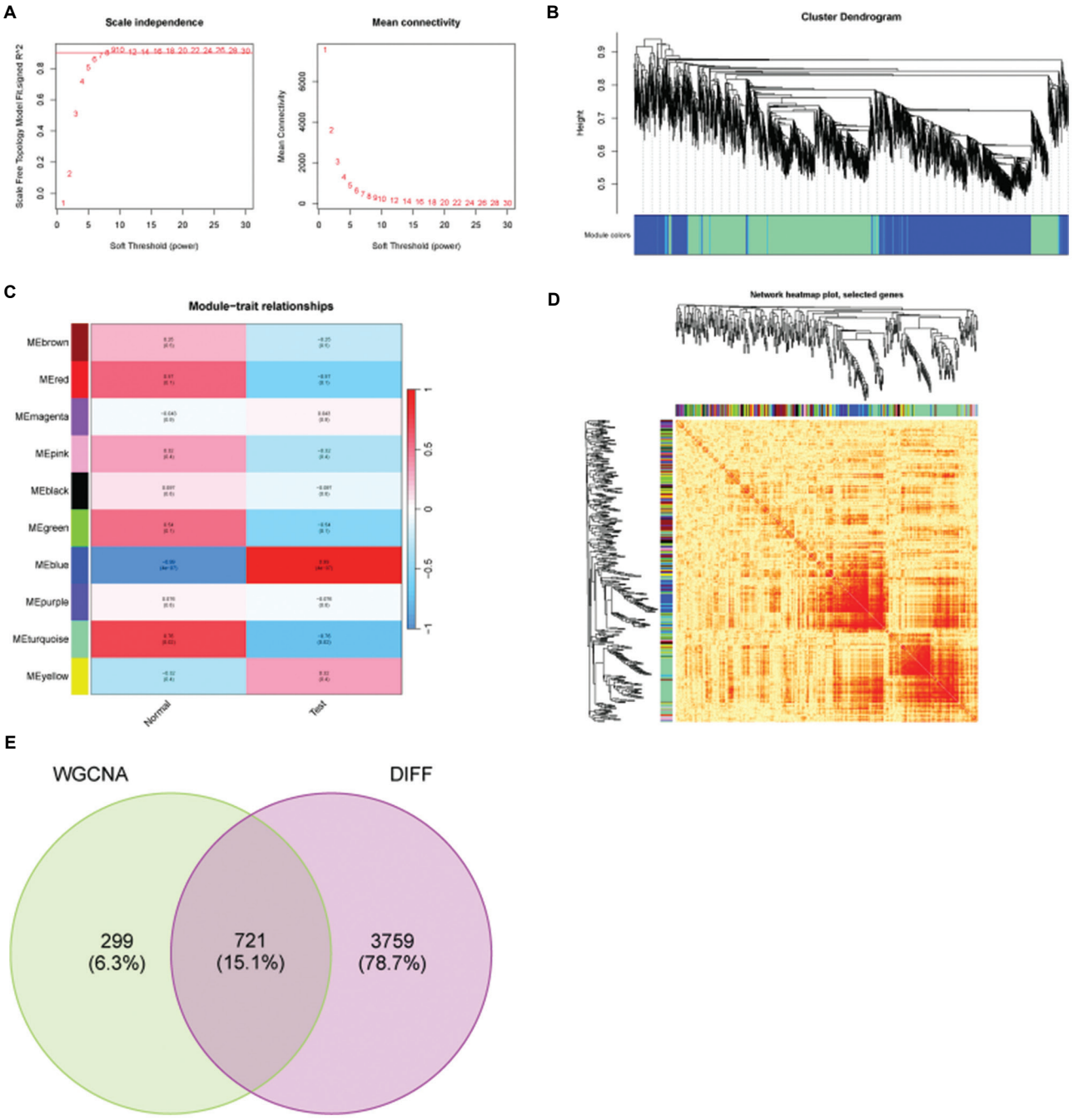

Figure 3. Construction of weighted gene co-expression networks. (A) Scale independence and mean connectivity were used to determine the optimal soft-threshold power. (B) Hierarchical clustering tree dendrogram showing co-expression modules identified by WGCNA. (C) Module-trait relationship heatmap displaying 10 gene modules (y-axis) and clinical modules (x-axis). The numbers in each cell indicate the Pearson's correlation coefficient (r) and the corresponding *p*-value. (D) Heatmap displaying the TOM among all genes. (E) Venn diagrams of 721 DEGs.
Abbreviations: DEGs: Differentially expressed genes; OP: Osteoporosis; TOM: Topological overlap matrices; WGCNA: Weighted Gene Co-expression Network Analysis.

that NUCB1, MTA1, DRAP1, and PCDHGA1 were associated with four, five, one, and two potential therapeutic Chinese medicines, respectively, while PEX19 was not associated with any. Notably, PEX19 was associated with five potential therapeutic Western medicines, as identified through Western medicine screening (Figure 11).

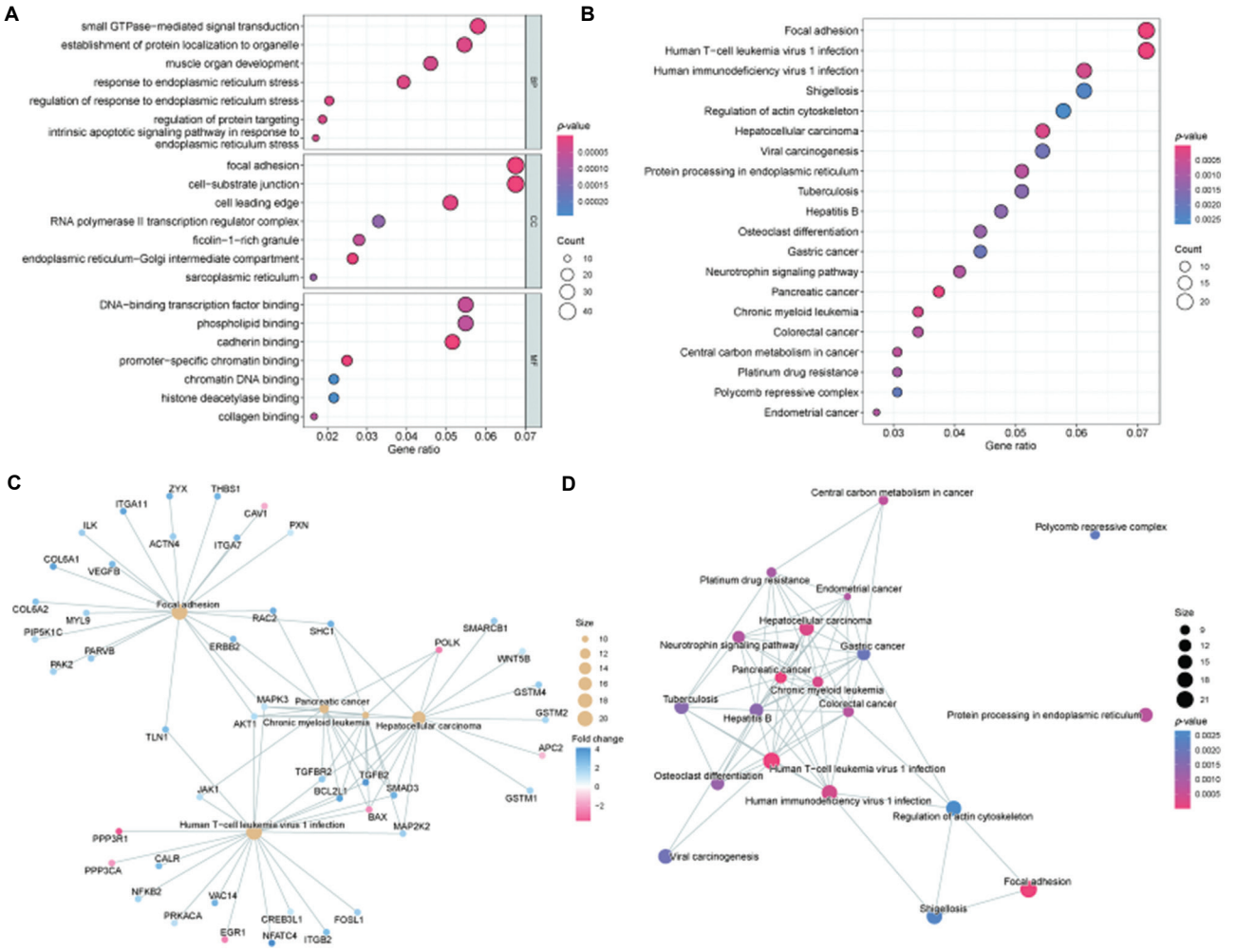

Figure 4. Functional enrichment analysis. (A) GO enrichment analysis of key genes. (B) KEGG pathway enrichment analysis. (C and D) Visualization of KEGG-enriched pathways and the related genes.
Abbreviations: GO: Gene ontology; KEGG: Kyoto encyclopedia of genes and genomes.

4. Discussion

Current diagnostic methods for OP primarily rely on dual-energy X-ray absorptiometry for BMD measurement and clinical fracture risk assessment tools such as FRAX. However, these approaches exhibit limitations, including insufficient sensitivity, high cost, and an inability to detect early-stage bone metabolism imbalances.[20] Moreover, due to the insidious nature of early OP symptoms, delayed intervention can lead to irreversible bone loss and fragility fractures, increasing disability, mortality, and socioeconomic burdens.[11] Consequently, identifying sensitive and specific molecular diagnostic markers is critical for early screening and targeted intervention. In the present study, we integrated WGCNA and LASSO regression to screen transcriptomic data from OP patients, identifying five key genes (*NUCB1*, *PEX19*, *MTA1*, *DRAP1*, and *PCDHGA1*). These genes were significantly upregulated in OP datasets ($p<0.01$) and demonstrated high diagnostic efficacy in ROC curve analysis (AUC > 0.85). While these findings suggest their potential as novel diagnostic biomarkers, the limited sample size (training: $n = 9$; validation: $n = 10$) necessitates further large-scale validation to evaluate possible overfitting.

NUCB1 is a calcium-binding protein predominantly localized in the Golgi apparatus of neurons.[21] It functions as a calcium-dependent guanine nucleotide dissociation inhibitor (GDI) for Gαi1, regulating G-protein-mediated signaling pathways.[22] In addition, NUCB1 modulates the unfolded protein response by inhibiting ATF6 activation, thereby influencing ER stress adaptation.[23] Emerging evidence suggests that dysregulated calcium signaling and ER stress contribute to bone metabolism disorders,

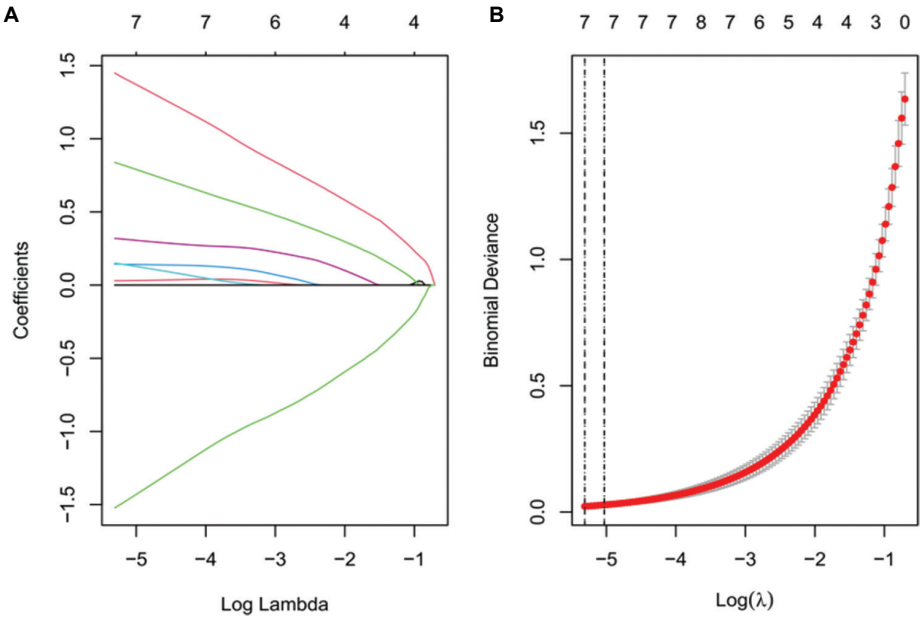

Figure 5. Machine learning screening using LASSO regression model. (A) The coefficient profiles were plotted based on the log (lambda) sequence. (B) A 10-fold CV plot for optimal λ selection.
Abbreviations: CV: Cross-validation; LASSO: Least Absolute Shrinkage and Selection Operator.

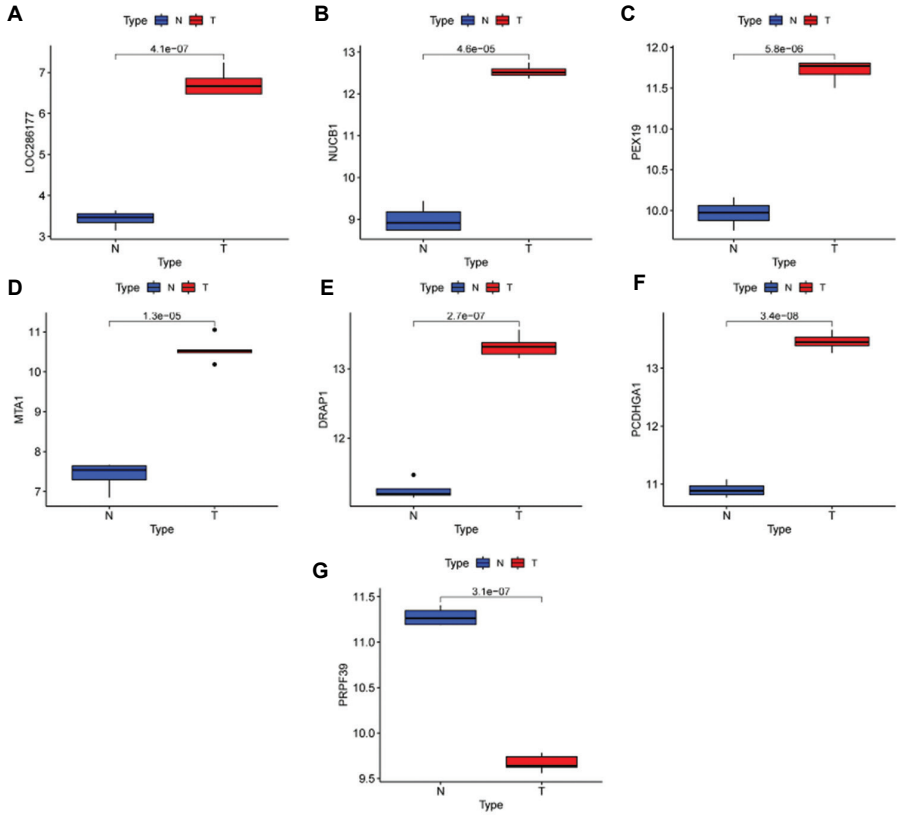

Figure 6. Validation of hub gene expression in the GSE35958 dataset. Differential analysis of (A) LOC286177, (B) NUCB1, (C) PEX19, (D) MTA1, (E) DRAP1, (F) PCDHGA1, and (G) PRPF39 expresison between healthy and OP groups. Group description: N, normal group; T, test group comprising osteoporosis (OP) patients.
Abbreviations: NUCB1: Nucleobindin 1; PEX19: Peroxisomal biogenesis factor 19; MTA1: Metastasis associated 1; DRAP1: DRA aassociated protein 1; PCDHGA1: Protocadherin gamma A1; PRPF39: Pre-mRNA processing factor 39.

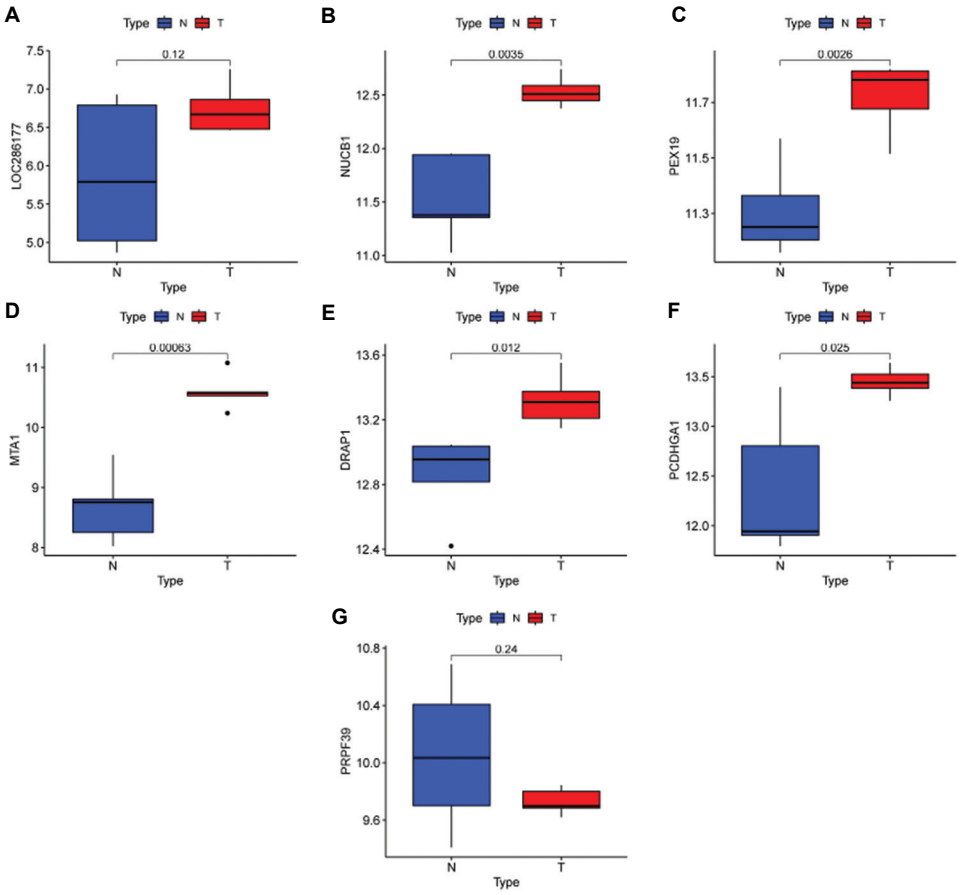

Figure 7. Validation of hub gene expression in an external GSE35956 dataset. Differential analysis of (A) LOC286177, (B) NUCB1, (C) PEX19, (D) MTA1, (E) DRAP1, (F) PCDHGA1, and (G) PRPF39 expresison between healthy and OP groups. Group description: N, normal group; T, test group comprising osteoporosis (OP) patients.
Abbreviations: NUCB1: Nucleobindin 1; PEX19: Peroxisomal biogenesis factor 19; MTA1: Metastasis associated 1; DRAP1: DRA aassociated protein 1; PCDHGA1: Protocadherin gamma A1; PRPF39: Pre-mRNA processing factor 39.

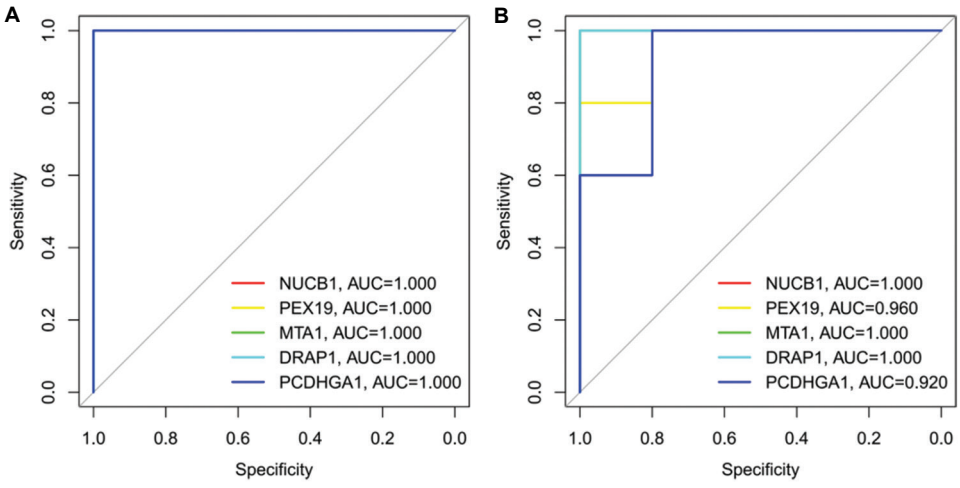

Figure 8. ROC curve analysis. ROC curve evaluating the diagnostic performance of hub genes in the (A) GSE35958 and (B) GSE35956 datasets.
Abbreviations: AUC: Area under the curve; ROC: Receiver operating characteristic.

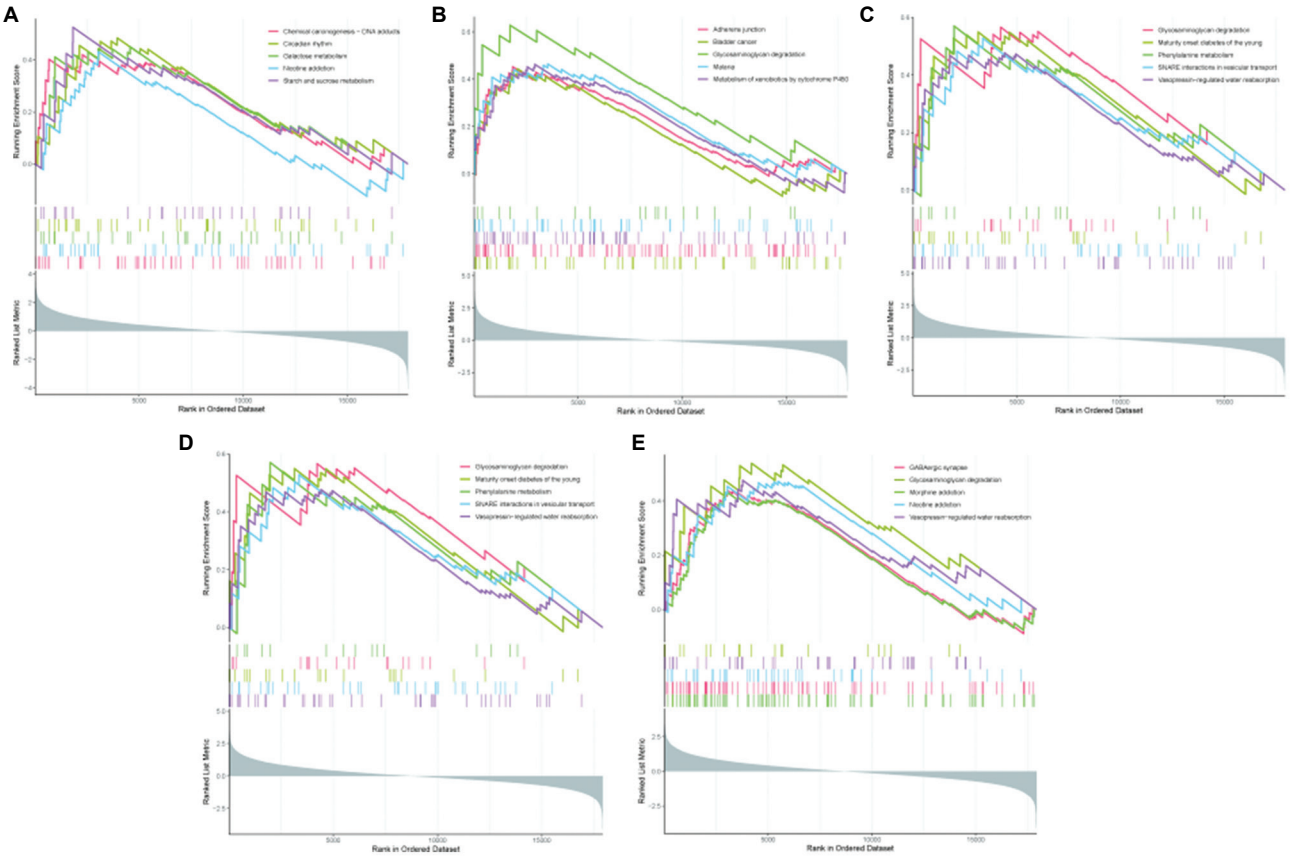

Figure 9. Enrichment results of GSEA to identify the potential pathways that differentiate the healthy and OP groups. Overview of the upper five pathways obtained from GSEA of (A) NUCB1, (B) PEX19, (C) MTA1, (D) DRAP1, and (E) PCDHGA1.
Abbreviations: GSEA: Gene Set Enrichment Analysis; OP: Osteoporosis; NUCB1: Nucleobindin 1; PEX19: Peroxisomal biogenesis factor 19; MTA1: Metastasis associated 1; DRAP1: DRA aassociated protein 1; PCDHGA1: Protocadherin gamma A1; PRPF39: Pre-mRNA processing factor 39.

including OP.[24,25] Although NUCB1 has not been directly linked to bone remodeling, its role in calcium homeostasis and stress response pathways raises the possibility of its involvement in osteoclast/osteoblast regulation. Further studies are required to determine whether NUCB1-mediated calcium signaling or ER stress modulation affects bone mineralization processes.

PEX19 is an essential chaperone for peroxisome membrane assembly and protein import, critical for peroxisome biogenesis.[26] Recent studies have demonstrated that PEX19 deficiency impairs peroxisomal β-oxidation of very-long-chain fatty acids, resulting in intracellular lipid accumulation and excessive reactive oxygen species (ROS) generation.[27] Emerging evidence indicates that ROS promote osteoclast differentiation and enhance bone resorption activity,[28] suggesting that PEX19 dysfunction may contribute to OP through ROS-mediated osteoclast activation. Furthermore, impaired peroxisomal metabolism may reduce endogenous fatty acid availability, potentially compromising osteoblast bioenergetics and bone formation.[29]

MTA1 is a core component of the nucleosome remodeling and histone deacetylation complex, which regulates gene expression through histone deacetylation (HDAC) modification.[30] HDAC plays an important role in regulating osteogenic differentiation and bone formation in bone marrow stromal cells,[31] which may provide research clues for MTA1 to activate osteoclastogenesis through epigenetic inhibition of osteogenic differentiation. In addition, modern studies have found that MTA1 can be involved in mediating the inactivation of the Wnt/β-catenin pathway,[32] and the imbalance of the Wnt/β-catenin pathway is closely related to the occurrence and development of OP,[33] which may also be a potential mechanism of MTA1-mediated OP.

DRAP1 is a transcriptional coregulator that forms a functional complex with DR1 to modulate gene expression programs.[34] Recent mechanistic studies revealed that the DRAP1/DR1 complex potentiates mTOR signaling activity, thereby promoting tumor progression and influencing therapeutic responses in triple-negative

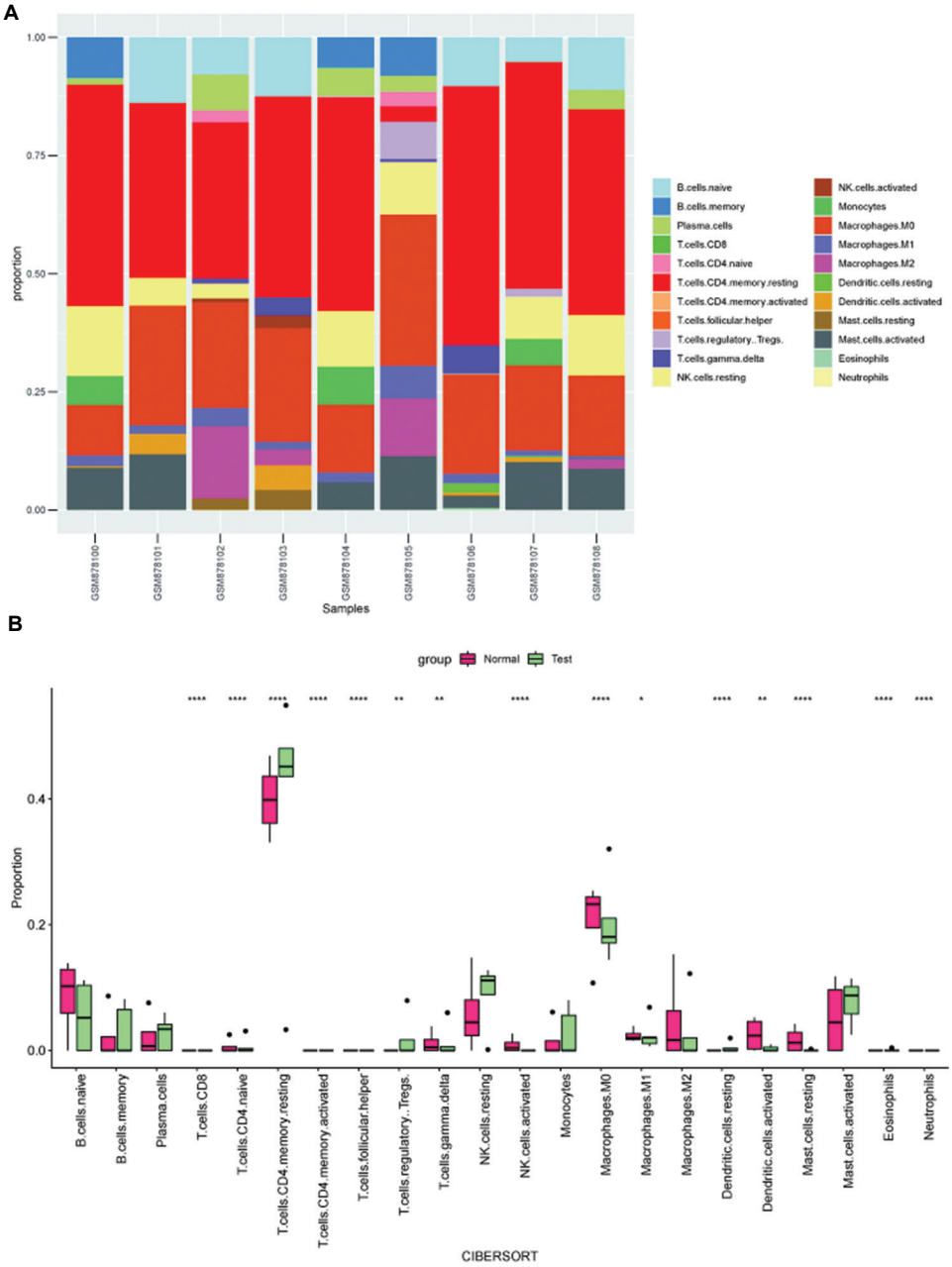

Figure 10. Immune infiltration analysis. (A) The bar plot showing the content of immune cell subsets in the OP and the control groups. Different colors represent different immune cell types. (B) Box plot showing the different distribution of immune cells between the OP and control groups.
Abbreviation: OP: Osteoporosis.

breast cancer.[34] The mTOR pathway is known to regulate fundamental cellular processes, including autophagy and mitophagy,[35] which are critically involved in maintaining bone homeostasis. Evidence from bone biology research demonstrates that mTOR signaling coordinates osteoblast-mediated bone formation and osteoclast-mediated bone resorption.[35] While these findings suggest a potential regulatory role for DRAP1 in bone metabolism through mTOR pathway modulation, direct experimental evidence establishing this connection remains to be documented.

PCDHGA1, a neural cadherin family member, mediates calcium-dependent cell adhesion.[36] Recent studies identified its expression in osteoblasts and its association with mineralization capacity.[37] Genetic analyses have linked protocadherin loci to BMD variations.[38]

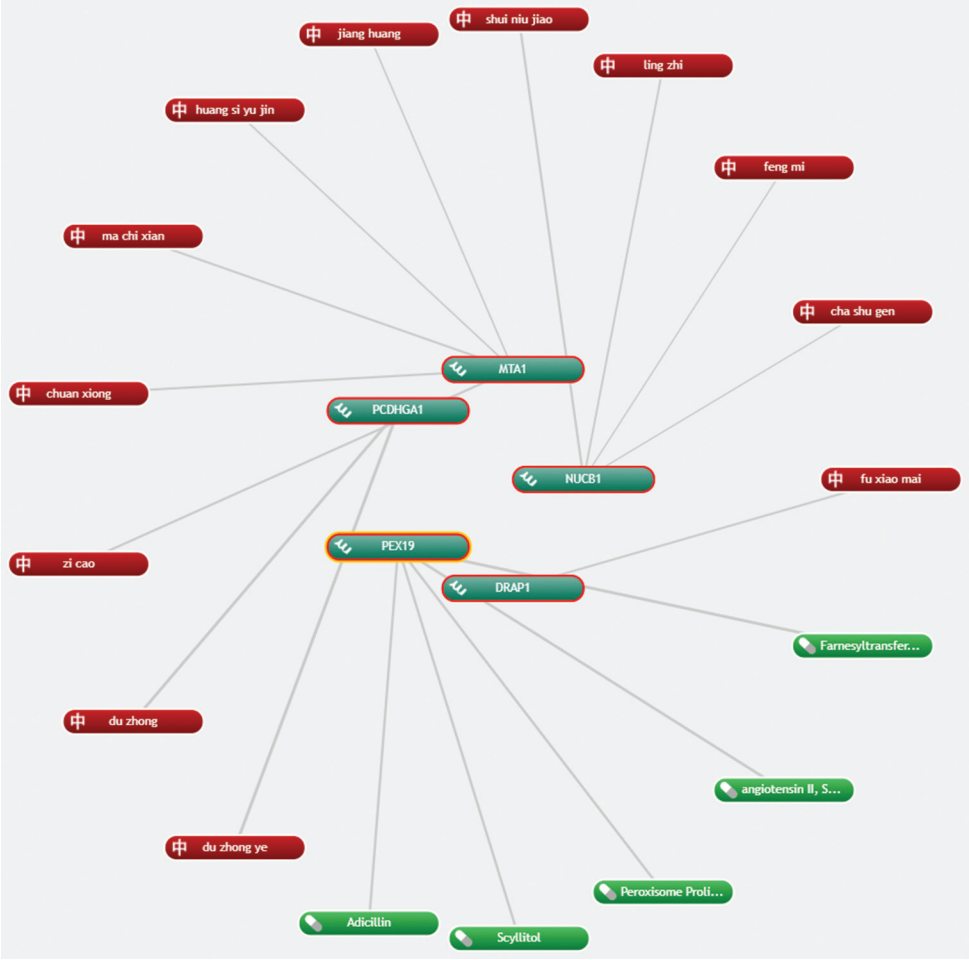

Figure 11. Potential targeted drug prediction and core components identification. Dark green boxes represent biomarkers; red boxes represent potentially therapeutic herbal medicines; green boxes represent potentially therapeutic Western medicines.

While these findings suggest PCDHGA1 involvement in bone metabolism, direct evidence for its role in OP pathogenesis remains unavailable. Future studies should characterize its function in bone remodeling and examine potential interactions with osteogenic pathways.

The present study has several limitations that should be considered. First, the exclusive reliance on computational analyses of public datasets (GSE35958: n = 9; GSE35956: n = 10) without wet laboratory or clinical validation represents a significant constraint. Second, the small sample sizes may compromise the statistical power and generalizability of the identified biomarkers (NUCB1, PEX19, MTA1, DRAP1, PCDHGA1). Third, the transcriptomic focus precluded integration with proteomic or metabolomic data, limiting systems-level understanding. Despite these limitations, this work provides the first computational evidence for these five biomarkers in OP diagnosis. To address these gaps, we propose a three-phase validation roadmap: (1) *in vitro* functional validation using osteoblast/osteoclast cultures to assess biomarker roles in bone remodeling; (2) preclinical validation in OP animal models with longitudinal biomarker monitoring; and (3) clinical validation through prospective cohort studies measuring biomarker performance against gold-standard diagnostics. These studies should be complemented by multiomics integration to elucidate the biomarkers' mechanistic pathways and regulatory networks.

5. Conclusion

This study identifies NUCB1, PEX19, MTA1, DRAP1, and PCDHGA1 as potential diagnostic biomarkers for OP through integrative bioinformatics and machine learning. These genes, found to be upregulated in OP and linked to ER stress, Wnt/β-catenin signaling, and immune dysregulation, and they demonstrate strong diagnostic accuracy (AUC > 0.85). Immune infiltration analysis further supports their role in OP-associated microenvironmental changes. While computational predictions suggest therapeutic

candidates, further experimental and clinical validation is essential to confirm their mechanistic roles and translational value. These findings provide potential directions for developing diagnostic and therapeutic strategies for OP.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Farra Aidah Jumuddin
Formal analysis: Zarina Awang
Investigation: Cuicui Zhou
Methodology: Cuicui Zhou, Farra Aidah Jumuddin
Writing–original draft: Cuicui Zhou
Writing–review & editing: Cuicui Zhou, Zarina Awang

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data is available from the corresponding author upon reasonable request.

References

1. Gómez O, Talero AP, Zanchetta MB, *et al*. Diagnostic, treatment, and follow-up of osteoporosis-position statement of the Latin American federation of endocrinology. *Arch Osteoporos.* 2021;16(1):114.

   doi: 10.1007/s11657-021-00974-x

2. Gao S, Zhao Y. Quality of life in postmenopausal women with osteoporosis: A systematic review and meta-analysis. *Qual Life Res.* 2023;32(6):1551-1565.

   doi: 10.1007/s11136-022-03281-1

3. Compston JE, McClung MR, Leslie WD. Osteoporosis. *Lancet.* 2019;393(10169):364-376.

   doi: 10.1016/s0140-6736(18)32112-3

4. Clynes MA, Harvey NC, Curtis EM, Fuggle NR, Dennison EM, Cooper C. The epidemiology of osteoporosis. *Br Med Bull.* 2020;133(1):105-117.

   doi: 10.1093/bmb/ldaa005

5. Wu D, Cline-Smith A, Shashkova E, Perla A, Katyal A, Aurora R. T-cell mediated inflammation in postmenopausal osteoporosis. *Front Immunol.* 2021;12:687551.

   doi: 10.3389/fimmu.2021.687551

6. Fischer V, Haffner-Luntzer M. Interaction between bone and immune cells: Implications for postmenopausal osteoporosis. *Semin Cell Dev Biol.* 2022;123:14-21.

   doi: 10.1016/j.semcdb.2021.05.014

7. Tyagi AM, Srivastava K, Mansoori MN, Trivedi R, Chattopadhyay N, Singh D. Estrogen deficiency induces the differentiation of IL-17 secreting Th17 cells: A new candidate in the pathogenesis of osteoporosis. *PLoS One.* 2012;7(9):e44552.

   doi: 10.1371/journal.pone.0044552

8. Management of osteoporosis in postmenopausal women: The 2021 position statement of The North American menopause society. *Menopause.* 2021;28(9):973-997.

   doi: 10.1097/gme.0000000000001831

9. Reid IR. Vitamin D effect on bone mineral density and fractures. *Endocrinol Metab Clin North Am.* 2017;46(4):935-945.

   doi: 10.1016/j.ecl.2017.07.005

10. Weaver CM, Gordon CM, Janz KF, *et al*. The national osteoporosis foundation's position statement on peak bone mass development and lifestyle factors: A systematic review and implementation recommendations. *Osteoporos Int.* 2016;27(4):1281-1386.

    doi: 10.1007/s00198-015-3440-3

11. LeBoff MS, Greenspan SL, Insogna KL, *et al*. The clinician's guide to prevention and treatment of osteoporosis. *Osteoporos Int.* 2022;33(10):2049-2102.

    doi: 10.1007/s00198-021-05900-y

12. Johnston CB, Dagar M. Osteoporosis in older adults. *Med Clin North Am.* 2020;104(5):873-884.

    doi: 10.1016/j.mcna.2020.06.004

13. Eastell R, Szulc P. Use of bone turnover markers in postmenopausal osteoporosis. *Lancet Diabetes Endocrinol.* 2017;5(11):908-923.

    doi: 10.1016/s2213-8587(17)30184-5

14. Li H, Xiao Z, Quarles LD, Li W. Osteoporosis: Mechanism, molecular target and current status on drug development. *Curr Med Chem.* 2021;28(8):1489-1507.

    doi: 10.2174/0929867327666200330142432

15. Lo HJ, Tsai CH, Huang TW. Apoptosis-associated genetic mechanisms in the transition from rheumatoid arthritis to osteoporosis: A bioinformatics and functional analysis approach. *APL Bioeng.* 2024;8(4):046107.

    doi: 10.1063/5.0233961

16. Xu M, Zhou H, Hu P, *et al*. Identification and validation of immune and oxidative stress-related diagnostic markers for diabetic nephropathy by WGCNA and machine learning. *Front Immunol.* 2023;14:1084531.

doi: 10.3389/fimmu.2023.1084531

17. Jiang F, Zhou H, Shen H. Identification of critical biomarkers and immune infiltration in rheumatoid arthritis based on WGCNA and LASSO algorithm. *Front Immunol.* 2022;13:925695.

doi: 10.3389/fimmu.2022.925695

18. Wang T, Dai L, Shen S, *et al*. Comprehensive molecular analyses of a macrophage-related gene signature with regard to prognosis, immune features, and biomarkers for immunotherapy in hepatocellular carcinoma based on WGCNA and the LASSO algorithm. *Front Immunol.* 2022;13:843408.

doi: 10.3389/fimmu.2022.843408

19. Liu F, Huang Y, Liu F, Wang H. Identification of immune-related genes in diagnosing atherosclerosis with rheumatoid arthritis through bioinformatics analysis and machine learning. *Front Immunol.* 2023;14:1126647.

doi: 10.3389/fimmu.2023.1126647

20. Carey JJ, Chih-Hsing Wu P, Bergin D. Risk assessment tools for osteoporosis and fractures in 2022. *Best Pract Res Clin Rheumatol.* 2022;36(3):101775.

doi: 10.1016/j.berh.2022.101775

21. Tulke S, Williams P, Hellysaz A, Ilegems E, Wendel M, Broberger C. Nucleobindin 1 (NUCB1) is a golgi-resident marker of neurons. *Neuroscience.* 2016;314:179-188.

doi: 10.1016/j.neuroscience.2015.11.062

22. Kapoor N, Gupta R, Menon ST, Folta-Stogniew E, Raleigh DP, Sakmar TP. Nucleobindin 1 is a calcium-regulated guanine nucleotide dissociation inhibitor of G{alpha}i1. *J Biol Chem.* 2010;285(41):31647-31660.

doi: 10.1074/jbc.M110.148429

23. Tsukumo Y, Tomida A, Kitahara O, *et al*. Nucleobindin 1 controls the unfolded protein response by inhibiting ATF6 activation. *J Biol Chem.* 2007;282(40):29264-29272.

doi: 10.1074/jbc.M705038200

24. Proudfoot D. Calcium signaling and tissue calcification. *Cold Spring Harb Perspect Biol.* 2019;11(10):a035303.

doi: 10.1101/cshperspect.a035303

25. Li HJ, Goff A, Rudzinskas SA, *et al*. Altered estradiol-dependent cellular Ca^{2+} homeostasis and endoplasmic reticulum stress response in premenstrual dysphoric disorder. *Mol Psychiatry.* 2021;26(11):6963-6974.

doi: 10.1038/s41380-021-01144-8

26. Agrawal G, Shang HH, Xia ZJ, Subramani S. Functional regions of the peroxin Pex19 necessary for peroxisome biogenesis. *J Biol Chem.* 2017;292(27):11547-11560.

doi: 10.1074/jbc.M116.774067

27. Sarkar C, Lipinski MM. Role and function of peroxisomes in neuroinflammation. *Cells.* 2024;13(19):1655.

doi: 10.3390/cells13191655

28. Agidigbi TS, Kim C. Reactive oxygen species in osteoclast differentiation and possible pharmaceutical targets of ROS-mediated osteoclast diseases. *Int J Mol Sci.* 2019;20(14):3576.

doi: 10.3390/ijms20143576

29. Nandy A, Helderman RCM, Thapa S, *et al*. Lipolysis supports bone formation by providing osteoblasts with endogenous fatty acid substrates to maintain bioenergetic status. *Bone Res.* 2023;11(1):62.

doi: 10.1038/s41413-023-00297-2

30. Li DQ, Pakala SB, Nair SS, Eswaran J, Kumar R. Metastasis-associated protein 1/nucleosome remodeling and histone deacetylase complex in cancer. *Cancer Res.* 2012;72(2):387-394.

doi: 10.1158/0008-5472.Can-11-2345

31. Wang J, Wang CD, Zhang N, *et al*. Mechanical stimulation orchestrates the osteogenic differentiation of human bone marrow stromal cells by regulating HDAC1. *Cell Death Dis.* 2016;7(5):e2221.

doi: 10.1038/cddis.2016.112

32. Lu Y, Wei C, Xi Z. Curcumin suppresses proliferation and invasion in non-small cell lung cancer by modulation of MTA1-mediated Wnt/β-catenin pathway. *In Vitro Cell Dev Biol Anim.* 2014;50(9):840-850.

doi: 10.1007/s11626-014-9779-5

33. Gao Y, Chen N, Fu Z, Zhang Q. Progress of wnt signaling pathway in osteoporosis. *Biomolecules.* 2023;13(3):483.

doi: 10.3390/biom13030483

34. Huang MY, Hu SY, Dong J, *et al*. The DRAP1/DR1 repressor complex increases mTOR activity to promote progression and confer everolimus sensitivity in triple-negative breast cancer. *Cancer Res.* 2024;84(16):2660-2673.

doi: 10.1158/0008-5472.Can-23-2781

35. Wang S, Deng Z, Ma Y, *et al*. The role of autophagy and mitophagy in bone metabolic disorders. *Int J Biol Sci.* 2020;16(14):2675-2691.

doi: 10.7150/ijbs.46627

36. Wu Q, Maniatis T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell.* 1999;97(6):779-790.

doi: 10.1016/s0092-8674(00)80789-8

37. Guan M, Pan D, Zhang M, Leng X, Yao B. The aqueous extract of eucommia leaves promotes proliferation, differentiation,

and mineralization of osteoblast-like MC3T3-E1 cells. *Evid Based Complement Alternat Med*. 2021;2021:3641317.

doi: 10.1155/2021/3641317

38. Estrada K, Styrkarsdottir U, Evangelou E, *et al*. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet*. 2012;44(5):491-501.

doi: 10.1038/ng.2249