Iran J Public Health, Vol. 54, No.8, Aug 2025, pp.1742-1753

Original Article

Identification of Potential Biomarkers for Osteoarthritis

Franko Shehaj [1], ***Ahmed Jasim Mahmood Al-Mashhadani*** [2], ****Haohuan Li*** [1]

1. *Department of Orthopedics, Renmin Hospital of Wuhan University, Wuhan, China*
2. *Department of Ophthalmology, Renmin Hospital of Wuhan University, Wuhan, China*

***Corresponding Author:** Email: lihaohuan@whu.edu.cn

(Received 10 Feb 2025; accepted 21 May 2025)

Abstract

Background: We aimed to identify biomarkers associated with Osteoarthritis (OA) and evaluate their predictive capabilities.

Methods: Four synovial tissue datasets (GSE1919, GSE12021, GSE55235, GSE55457) and one peripheral blood mononuclear rcells (PBMC) dataset (GSE48556) were obtained. GSE55235 and GSE55457 were merged to conduct differential expression analysis and train machine learning algorithms. Predictive models were trained using a subset of genes and then validated on the other datasets. In addition, PBMC dataset was used to train predictive models using the same subset of genes, with the synovial tissue datasets serving as validation datasets. Finally, immune infiltration analysis was performed in the merged synovial tissue dataset using CIBERSORT.

Results: RPA3, LAMA5, SAT1, and UCP2 were used to train machine learning algorithms. Predictive models performed well in synovial tissue datasets but faced challenges in the PBMC dataset, as models achieved high sensitivity but moderate specificity. However, models trained on the PBMC dataset exhibited high sensitivity and specificity in the four external validation datasets. SAT1 exhibited the highest impact on the model performance. Immune infiltration analysis revealed significant differences in the expression of several immune cells, such as mast cells, between OA and control groups. In general, the four genes showed moderate to strong correlations with mast cells.

Conclusion: While promising, our findings point to the need for further studies to validate biomarkers and improve the models' predictive power across diverse sample types.

Keywords: Orthopedics; Osteoarthritis; Differentially expressed genes; Machine learning

Introduction

Osteoarthritis (OA) is the most prevalent form of arthritis, affecting millions of people worldwide. It is characterized by the degeneration of joint cartilage and the underlying bone, leading to pain, stiffness, and impaired movement (1). Epidemiological studies estimate that OA affects more than 240 million people globally, with a higher prevalence in older adults and women (2). Approximately 30%-65% of the risk of OA is genetically determined (3). The exploration of gene expression profiles in OA has opened a promising frontier in understanding the molecular mechanisms of the disease and its progression (4).

Recent advances in genomics and bioinformatics have enabled researchers to identify specific gene expression patterns associated with OA, thereby explaining complex regulatory networks that lead to cartilage degradation and joint inflammation (5,6). By comparing the genetic signatures of affected tissues with those of healthy controls, scientists have begun to identify potential biomarkers that could serve as early indicators of the disease (5,7). This molecular approach not only improves traditional diagnostic methods but also holds significant promise for the development of personalized therapeutic strategies. Early detection through gene expression profiling could facilitate timely interventions that slow the progression of OA, mitigating irreversible joint damage. In addition, understanding the genetic factors that contribute to OA can aid in the stratification of patients, as treatment regimens can be tailored to individual risk profiles (8). Taken together, integrating genomic data into clinical practice represents a critical step towards a more proactive and preventive approach to managing OA, which ultimately results in improved patient outcomes and quality of life.

Despite advances in clinical diagnostics and imaging, early detection remains problematic because conventional methods fail to capture the molecular events that precede overt joint damage. The integration of machine learning algorithms in genomic research has the potential to deepen our understanding of OA. Machine learning can handle vast amounts of genetic data, uncovering complex patterns and relationships that traditional statistical methods might miss (9,10). Thus, this study aimed to identify potential biomarkers for OA through a comprehensive analysis of gene expression profiles and immune cell interactions, thereby paving the way for improved early detection and personalized therapeutic strategies.

Materials and Methods

Data Collection

The flowchart of the study is shown in Fig. 1A. The study utilized five gene expression datasets obtained from the Gene Expression Omnibus. The primary dataset used for training the models was the merged dataset consisting of GSE55235 and GSE55457, each having 10 OA and 10 control samples. The first independent validation dataset (GSE1919) included 10 samples: 5 OA and 5 controls. GSE12021 was used as the second independent validation dataset, it consisted of 10 OA and 9 control samples. The third independent validation set, GSE48556, included gene expression data of peripheral blood mononuclear cells (PBMC) obtained from 106 OA and 33 control samples.

Differential Expression Analysis

Differentially expressed genes (DEGs) were obtained from the training dataset (merged GSE55235&GSE55457) using the following cut-off values: adjusted *P*-value (Benjamini-Hochberg procedure) <0.05 and |log FC| >0.5. In addition, gene ontology (GO) analysis (biological processes and cellular components) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis were performed using clusterProfiler and org.Hs.eg.db. T-test was utilized to calculate the *P*-value of gene expression levels between the OA and control groups (rstatix package).

Machine Learning Analysis and Model Interpretation

Two methods were used for feature selection: the Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest (RF). The common genes identified by both feature selection methods were selected for model training. The following machine learning algorithms were used: logistic regression (LR), support vector machines (SVM), K-nearest neighbors (KNN), bagging with LR as the base estimator, and AdaBoost with LR as the base estimator. Hyperparameter tuning was performed using GridSearchCV to identify the optimal parameters for each model. Models were evaluated based on the following metrics: area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and F1 score (11). Permutation feature importance,

SHapley Additive explanation (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) were used to interpret predictions.

Immune Infiltration Analysis

Immune infiltration analysis was performed using CIBERSORT. Immune cell types that were expressed in ≤20% of samples were excluded. Wilcoxon signed-rank test was used to calculate differences in cell expression between the OA and control groups of the training data. Spearman's rank correlation was used to calculate the correlation between paired immune cell types and between the most relevant genes identified by feature selection methods and immune cell types.

Ethics approval

Ethical review and approval were waived for this study because no humans were involved in this study.

Results

Differential Expression Analysis

Prior to data analysis, five datasets were preprocessed. Individually preprocessed GSE55235 and GSE55457 were merged into one dataset to increase the sample size of the training dataset. To assess the presence of batch effects within the merged dataset, a PCA plot was constructed. As batch effects were clearly present, batch effects correction was conducted. Data distribution before and after preprocessing are shown in Fig. 1B.

Overall, 880 DEGs, among which 364 were downregulated and 516 were upregulated, were screened (Fig. 2A-B). GO analysis revealed that the identified DEGs were mainly enriched in cytokine production, taxis as well as cell adhesion, and were mostly found in extracellular matrix and organelles' membranes and lumens. According to KEGG analysis, genes were predominantly enriched in signaling pathways (MAPK and PI3K-Akt), lipid and atherosclerosis, and rheumatoid arthritis (Fig. 2C).

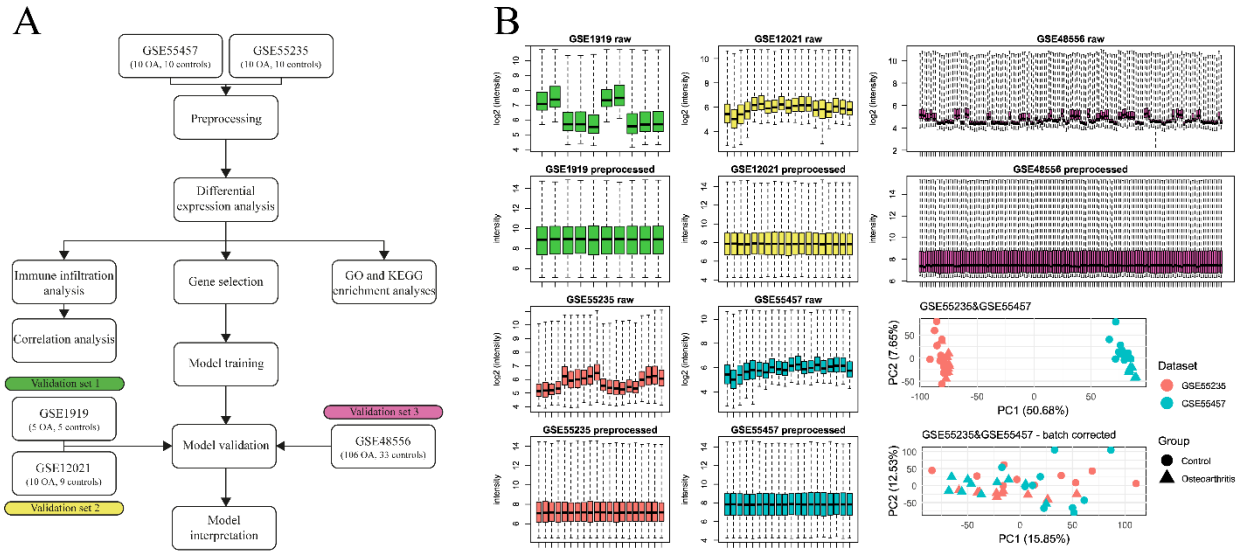

Fig. 1: (A) Flowchart of the study. (B) Boxplots of GSE1919, GSE12021, GSE48556, GSE55235, GSE55457, datasets before and after preprocessing (background correction, normalization, log2 transformation, gene annotation and filtering). Bottom right corner: principal component analysis scatter plots of the merged dataset (GSE55235&GSE55457) before and after batch effects correction

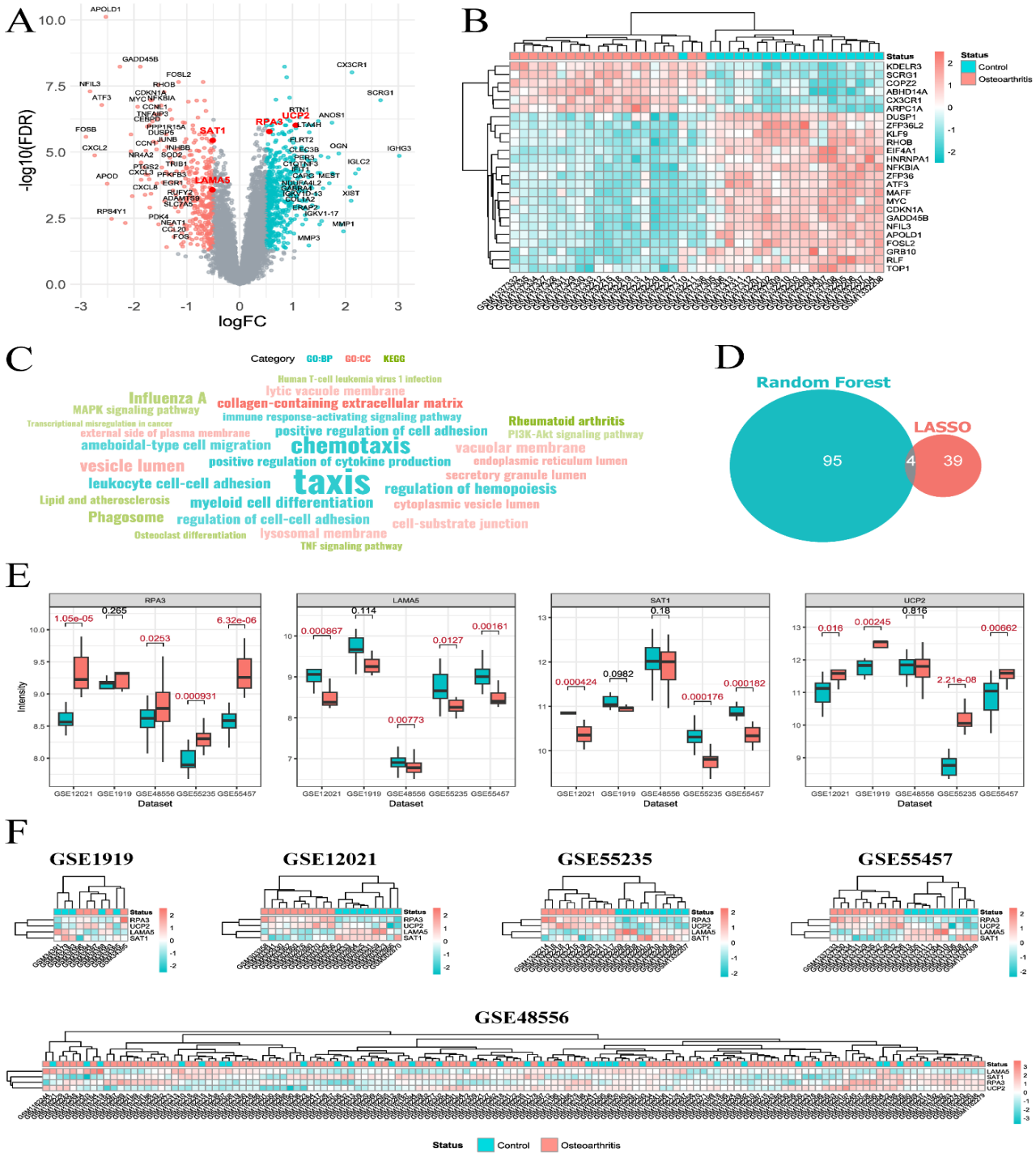

Fig. 2: (A) Volcano plot. (B) Heatmap of the top 25 differentially expressed genes (sorted by *P*-value) (C) Enrichment analysis word cloud of Gene Ontology biological processes (GO: BP), cellular components (GO: CC), and Kyoto Encyclopedia of Genes and Genomes (KEGG). Size represents gene count, color represents category, transparency represents -log10 (P-value). All categories have a -log10 (*P*-value) > 2. (D) Venn diagram of genes identified by two feature selection tools. Four identified genes are RPA3, LAMA5, SAT1, and UCP2. (E) Boxplots of expression levels of four genes between the osteoarthritis group (red) and the control group (blue) across all datasets. P-values < 0.05 are highlighted in red. (F) Heatmaps of fold changes of RPA3, LAMA5, SAT1, and UCP2 across all datasets

Model Training and Evaluation

LASSO model identified 39 relevant genes, and RF classifier identified 95 genes. The common genes identified by both methods were selected for model training. There were four common genes found between both methods: RPA3, LAMA5, SAT1, and UCP2 (Fig. 2D). UCP2 was found to be significantly upregulated in OA patients compared to controls in synovial tissues but not in PBMC (Fig. 2E). Although expression levels of several genes, namely RPA3 and LAMA5, were found to be significantly different between the OA and control groups in all datasets except for GSE1919. Additionally, heatmap of expression fold changes of the identified genes revealed a clear separation between OA and control samples in synovial tissue datasets with the exception of GSE1919 (Fig. 2F).

Five different machine learning algorithms were trained on the merged dataset using RPA3, LAMA5, SAT1, and UCP2 (Table 1 and Fig. 3A-C). In the first and second validation datasets, all five models, especially SVM and KNN, exhibited great predictive performance, characterized by high F1 score, specificity, sensitivity, and AUC.

Table 1: Performance of machine learning algorithms trained on the merged GSE55235&GSE55457 dataset (synovial tissue)

Validation dataset	Metrics	Logistic Regression	Support Vector Machines	k-Nearest Neighbors	Bagging (Logistic Regression)	AdaBoost
GSE1919	Sensitivity	0.8	1.0	0.8	1.0	0.8
	Specificity	1.0	1.0	1.0	1.0	1.0
	F1 score	0.83	0.91	0.91	0.83	0.83
	AUC	0.96	1.0	0.98	1.0	0.96
GSE12021	Sensitivity	1.0	1.0	1.0	1.0	1.0
	Specificity	1.0	1.0	1.0	1.0	1.0
	F1 score	1.0	1.0	1.0	1.0	1.0
	AUC	1.0	1.0	1.0	1.0	1.0
GSE48556	Sensitivity	0.72	0.7	0.42	0.43	0.72
	Specificity	0.64	0.64	0.67	0.88	0.64
	F1 score	0.83	0.80	0.55	0.61	0.83
	AUC	0.68	0.67	0.57	0.67	0.67

However, in the GSE48556 dataset, models exhibited lower predictive performance, with moderate AUC score. Next, we explored how models would perform when trained on the GSE48556 dataset (Table 2 and Fig. 3D-G). Compared to models trained on synovial tissue that consistently showed high performance, the PBMC-trained models had slightly more variability. However, the strong results obtained by several algorithms, such as LR (sensitivity, specificity, AUC and F1 score of above 0.8 in all datasets), indicate their potential for future applications.

Table 2: Performance of machine learning algorithms trained on the GSE48556 dataset (peripheral blood mononuclear cells)

Validation dataset	Metric	Logistic Regression	Support Vector Machines	k-Nearest Neighbors	Bagging (Logistic Regression)	AdaBoost
GSE1919	Sensitivity	0.8	1.0	0.0	0.8	0.8
	Specificity	0.8	0.2	1.0	0.8	0.8
	F1 score	0.8	0.71	0.0	0.8	0.8
	AUC	0.84	0.28	0.28	0.76	0.80
GSE12021	Sensitivity	1.0	0.0	0.9	1.0	0.9
	Specificity	1.0	1.0	0.78	0.89	1.0
	F1 score	1.0	0.0	0.86	0.95	0.95
	AUC	1.0	0.10	0.76	0.98	0.99
GSE55235	Sensitivity	0.8	0.0	0.7	0.9	0.8
	Specificity	1.0	1.0	0.8	0.9	1.0
	F1 score	0.89	0.0	0.74	0.9	0.89
	AUC	0.97	0.01	0.75	0.96	0.97
GSE55457	Sensitivity	1.0	1.0	0.9	1.0	1.0
	Specificity	1.0	0.0	0.7	0.9	1.0
	F1 score	1.0	0.0	0.82	0.95	1.0
	AUC	1.0	0.13	0.74	0.98	1.0

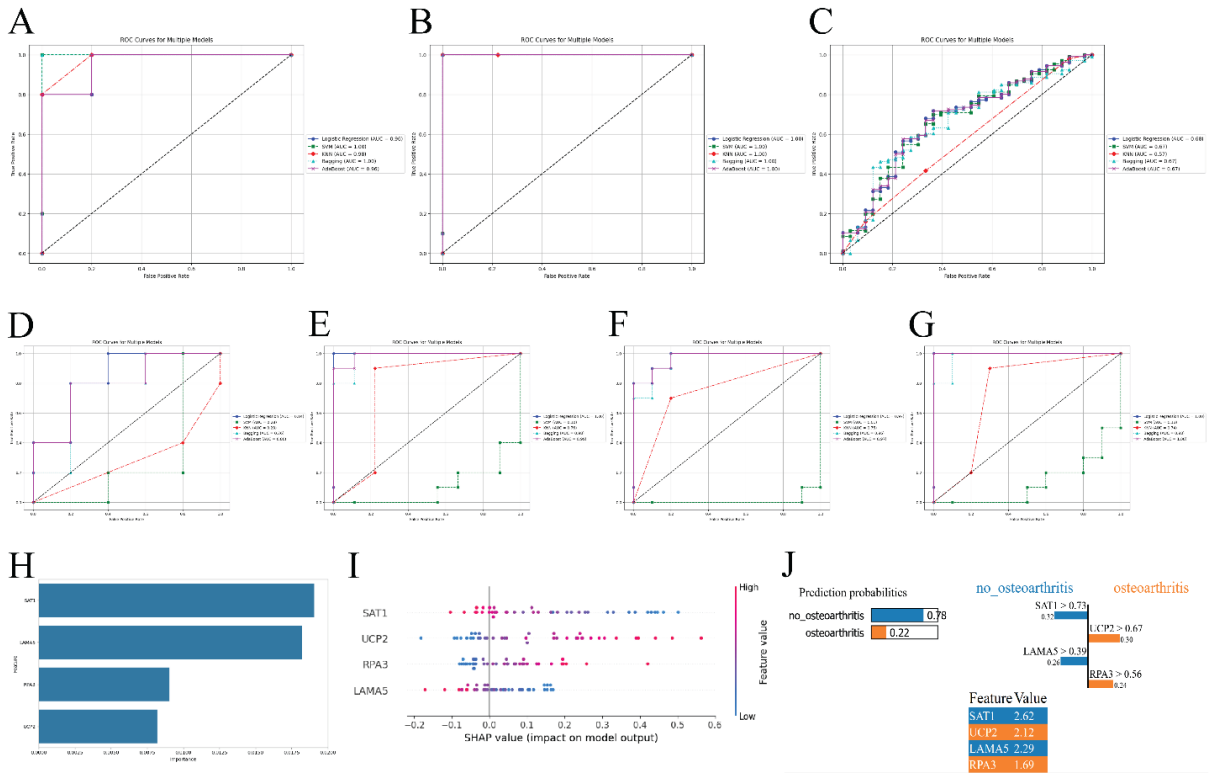

Fig. 3: Top panel: Receiver operating characteristic (ROC) curves of five models trained on the merged GSE55235&GSE55457 dataset (synovial tissue). (A) GSE1919. (B) GSE12021. (C) GSE48556. Middle panel: ROC curves of five models trained on the GSE48556 dataset (synovial tissue). (D) GSE1919. (E) GSE12021. (F) GSE55235. (G) GSE55457. Bottom panel: (H) Feature permutation plot. (I) SHapley Additive exPlanations (SHAP) plot. (J) Local Interpretable Model-agnostic Explanations (LIME) plot

Model Interpretation

Permutation feature importance was calculated for each gene, and genes were ranked based on their impact on the model performance. The importance of genes was determined by the magnitude of their predictive impact. Hence, genes with higher significance values were more influential in predicting OA. SAT1 and LAMA5 had the highest feature importance score reaching above 0.018. (Fig. 3H).

According to SHAP plot, SAT1 was the most influential gene (Fig. 3I). The SHAP values for all four genes showed a broad range of impacts on the model's output. Lower values of SAT1 (blue dots) were associated with higher SHAP values and thus higher impact on the model. Conversely, high values of SAT1 (red dots) contributed to the prediction of "control", albeit to a smaller degree as evidenced by SHAP values clustered around -0.2. Samples with lower values of SAT1 were more likely to be identified as having OA. RPA3 and UCP2 also had a substantial effect on the model. High values of these genes were associated with positive SHAP values, which demonstrates that their increased expressions were indicative of OA. Compared to the other three genes, LAMA5 values were less spread out. Lower values of LAMA5 had positive SHAP values (prediction of OA).

Based on the LIME plot, SAT1 and LAMA5 negatively influenced the decision tree model's prediction of OA (Fig. 3J). In other words, higher values of SAT1 and LAMA5 contributed to the prediction of "no OA" (control). SAT1, in particular, had the highest impact (feature value – 2.62), suggesting its critical role in the model's decision-making process. On the other hand, UCP2 and RPA3 most heavily influenced the OA prediction score. Although they also exhibited a rather strong influence on predicting OA, their impact was lower compared to SAT1 and LAMA5.

Immune Infiltration Analysis

Four cell types were excluded from the immune infiltration analysis as they were not expressed in a sufficient number of samples: native CD4+ T cells (expressed in one sample), activated memory CD4+ T cells (expressed in five samples), resting NK cells (expressed in one sample), and eosinophils (expressed in one sample). Expression levels of memory B cells and resting mast cells were significantly higher in the OA group compared to the control group. In contrast, resting memory CD4+ memory T cells, monocytes, and activated mast cells were significantly more abundant in the control group (Fig. 4A). M2 macrophages were the predominant cell type in almost all samples followed by mast cells and CD8+ T cells (Fig. 4C). Activated mast cells were strongly positively correlated with resting memory CD4+ T cells and very strongly negatively correlated with resting mast cells, whereas resting mast cells in turn were strongly negatively correlated with resting memory CD4+ T cells (Fig. 4B). All genes were found to be significantly correlated with activated and resting mast cells. RPA3 exhibited the strongest strength of correlation followed by SAT1, UCP2, and LAMA5 (Fig. 4D-E).

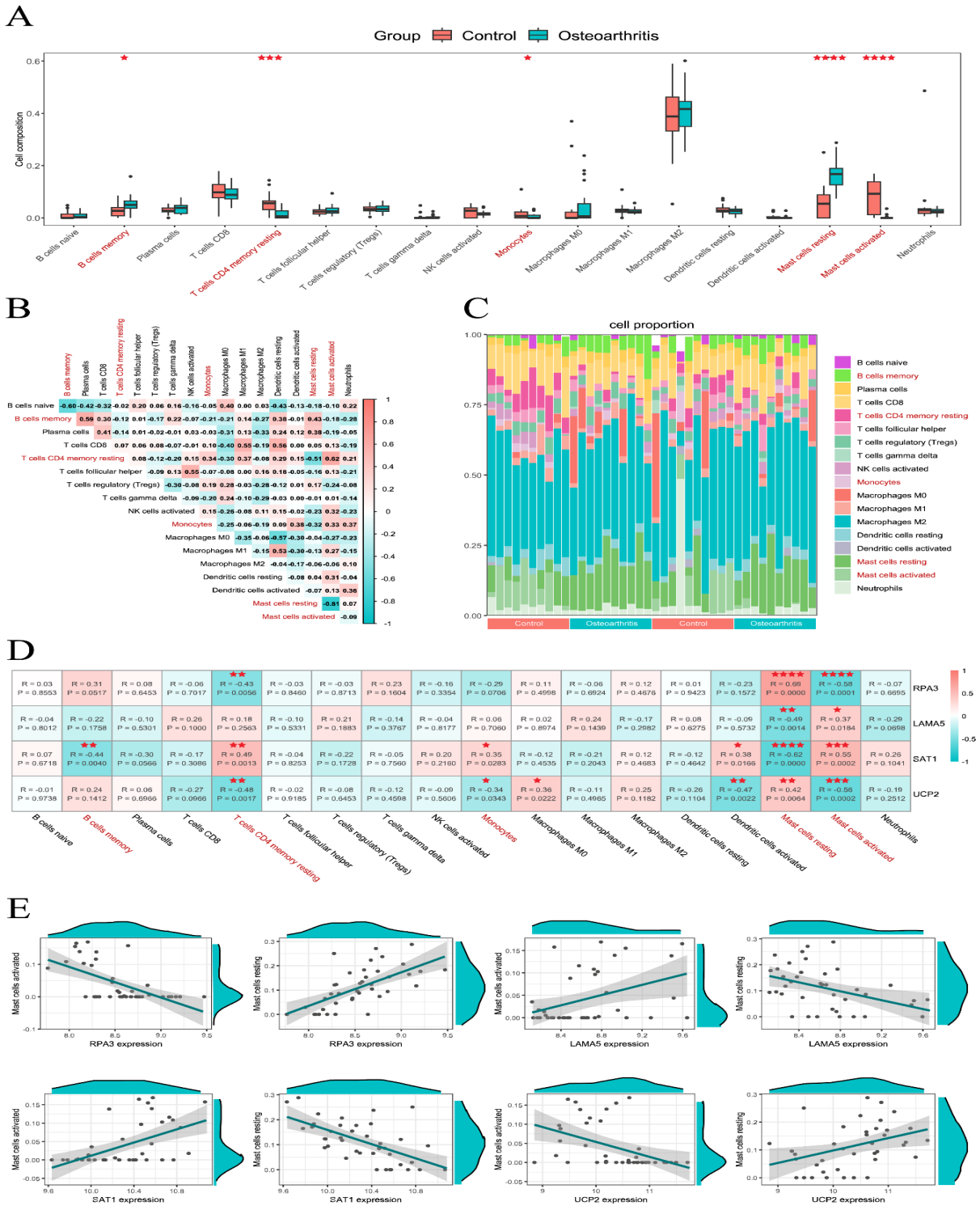

Fig. 4: (A) Boxplots of immune cell types between osteoarthritis and control samples within the training data. (B) Correlation matrix between paired 18 immune cell types. (C) Ratio of the immune cell types, where each column represents a sample within the merged dataset. (D) Heatmap of the Spearman correlation of paired genes and immune cell types. (E) Correlation between RPA3, LAMA5, SAT1 and UCP2 with resting and activated mast cells (R – correlation coefficient, P – P-value).
Note. *: *P*-value ≤ 0.05, **: *P*-value ≤ 0.01, ***: *P*-value ≤ 0.001, ****: *P*-value ≤ 0.0001. Names of all cell types with significant differences in expression levels between the two groups are highlighted in red

Discussion

Numerous studies have performed bioinformatics analyses of OA using datasets from the Gene Expression Omnibus. Several research works even utilized some datasets used in our study(12-15). However, there are some major differences. Firstly, while many similar studies either relied on a single dataset split into training and validation sets (13,15), or employed only a very limited number of external validation sets (16), this study included thorough validation of constructed predictive models across multiple independent datasets. Secondly, datasets were derived from different tissue types, which is a novel approach that improves the generalizability of the predictive models. Thirdly, while some studies applied machine learning solely for gene identification without assessing predictive performance (12), this study comprehensively evaluated model performance on different external validation sets using various metrics. Moreover, advanced machine learning methods and two feature selection methods were used to identify key candidate genes and construct predictive models. Finally, other than machine learning analysis, the paper explored correlations between gene expression and immune cells, providing more information on gene expression profiles of OA.

RPA3, LAMA5, SAT1, and UCP2 were identified as the most relevant for predictive modeling. The RPA3 gene encodes a subunit of the replication protein A complex, essential for DNA replication and repair. It stabilizes single-stranded DNA intermediates during these processes, playing a crucial role in maintaining genomic stability and integrity (17). It is mainly associated with tumorigenesis (18,19) and rheumatoid arthritis-associated interstitial lung disease (20). However, its association with OA is unclear and requires further investigation. LAMA5 (laminin subunit alpha 5) is crucial in the structure and function of the extracellular matrix. Due to the critical role of LAMA5, alterations in its expression levels have been implicated in various disorders, including OA (21,22). SAT1 (spermidine/spermine N^1-acetyltransferase 1) is downstream of P53, which as the name suggests, plays a crucial role in the conversion of spermidine and spermine back to putrescine (23). Inhibition of SAT1 could inhibit OA in murine models via suppressing chondrocyte ferroptosis and inflammation as well as the production of reactive oxygen species (24). UCP2 (uncoupling protein 2) was reported to regulate insulin secretion in the pancreas and reactive oxygen species production. Previous studies have demonstrated its widespread presence in the lymphoid system, macrophages, and osteoblasts (25). Moreover, single nucleotide polymorphisms of several genes were found to be associated with healthy aging (26), OA progression (27), etc.

Prediction models trained on high-dimensional small sample-sized data are often associated with bias and poor generalizability (28). Thus, we decided to merge two datasets, GSE55235 and GSE55457, to increase the sample size of the training dataset. The models were validated on three separate datasets, one of which included samples from PBMC. The findings of this study show good predictive performance of various models in predicting OA using gene expression data from RPA3, LAMA5, SAT1, and UCP2. Only UCP2 was found to be significantly differentially expressed between the OA group and the control group in synovial tissues. Machine learning models were trained on a combination of four genes rather than each gene separately. Models can identify subtle patterns or interactions between genes that collectively contribute to the prediction of OA(29). Thus, statistically significant differences in expression levels of each individual gene between the two groups are not as important in machine learning analysis compared to conventionally used statistical tools.

In the GSE48556 dataset, models exhibited good sensitivity but moderate specificity. Such significant differences in the results compared to the other datasets could be caused by the inability of algorithms to generalize well to expression patterns in PBMC, which are different from the molecular changes occurring in synovial tissues,

which are the primary site of inflammation and in OA (30). To investigate this further, we trained predictive models using the GSE48556 dataset and evaluated their performance on four datasets derived from synovial tissue samples. When trained on the GSE48556 datasets, several models such as LR and AdaBoost demonstrated good predictive performance in all external validation datasets as evidenced by high sensitivity, specificity, F1 score, and AUC. Thus, the most likely reason why models trained on GSE55235 and GSE55457 did not achieve high performance when validated on GSE48556, whereas models trained on GSE48556 performed well across all datasets, is the difference in sample size rather than significant biological differences between tissue types. A markedly larger sample size in GSE48556 provides a more representative training set that better captures the gene expression variability. The improvement in predictive performance when training on blood-based samples and validating on synovial tissue datasets shows the broader applicability of RPA3, LAMA5, SAT1, and UCP2 as systemic biomarkers.

Immune infiltration analysis revealed the complex immune dynamics associated with OA. M2 macrophages were the predominant cell type in almost all samples, followed by mast cells and CD8+ T cells. The expression levels of several immune cells, such as mast cells and memory T and B cells, were significantly different between the OA and control groups, which is consistent with findings made in earlier reports (31,32). Memory B cells and resting mast cells were significantly higher in the OA group, suggesting chronic inflammation and immune dysregulation, whereas CD4+ memory T cells, monocytes, and activated mast cells were more abundant in controls, which is indicative of a more regulated immune response (33). Notably, the strongest level of correlation was observed between resting and activated mast cells. LAMA5, UCP2, SAT1 and RPA3 were found to be significantly correlated with activated and resting mast cells exhibiting mostly moderate levels of correlation.

This study has several limitations. First, datasets come from different studies, which introduced heterogeneity due to biological and technological differences. Second, the sample sizes of validation datasets were relatively small, potentially limiting the statistical power of the analysis. Although multiple validation sets were used, larger datasets would enhance the generalizability of the findings. Third, the study focused on transcriptomic data without incorporating proteomics or metabolomics. A multi-omics approach could offer a deeper understanding of molecular mechanisms associated with OA. Fourth, this is a purely bioinformatics analysis of available datasets. Although machine learning analysis and immune infiltration analysis provided novel findings, the exact molecular mechanisms or relevance of the identified gene-immune cell correlations remain unclear.

To address these limitations, future studies should aim to integrate larger and more diverse datasets to improve the generalizability of the findings. The inclusion of multi-omics data, such as proteomics, metabolomics, and single-cell RNA sequencing, could provide a more comprehensive understanding of OA pathophysiology. Moreover, prospective cohort studies are needed to validate the identified biomarkers and predictive models in independent populations. Laboratory experiments can be conducted to understand the biological roles of RPA3, LAMA5, SAT1, and UCP2 in OA pathogenesis. Finally, translating these findings into clinical applications requires the development of non-invasive biomarkers. Since PBMC-based models exhibited variable performance, further optimization and validation are necessary to allow early OA detection or develop therapeutic interventions.

Conclusion

Overall, our results demonstrate the potential of machine learning in OA research while highlighting the necessity for continued exploration of additional biomarkers and improved model robustness across diverse datasets. Future work should focus on refining these models to enhance their applicability in clinical practice, addressing

the variability in performance and ensuring reliable diagnostic outcomes.

Journalism Ethics considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

The authors would like to thank Bulat Abdrakhimov for his tremendous help with the code and valuable comments while conducting the study and writing the manuscript.

Conflict of Interests

The authors declare no conflicts of interest.

References

1. Giorgino R, Albano D, Fusco S, et al (2023). Knee Osteoarthritis: Epidemiology, Pathogenesis, and Mesenchymal Stem Cells: What Else Is New? An Update. *Int J Mol Sci*, 24(7):6405.
2. Safiri S, Kolahi AA, Smith E, et al (2020). Global, regional and national burden of osteoarthritis 1990-2017: a systematic analysis of the Global Burden of Disease Study 2017. *Ann Rheum Dis*, 79 (6):819-828.
3. Vina ER, Kwoh CK (2018). Epidemiology of osteoarthritis: literature update. *Curr Opin Rheumatol*, 30 (2):160-167.
4. Young DA, Barter MJ, Soul J (2022). Osteoarthritis year in review: genetics, genomics, epigenetics. *Osteoarthritis Cartilage*, 30 (2):216-225.
5. Hu X, Ni S, Zhao K, et al (2022). Bioinformatics-Led Discovery of Osteoarthritis Biomarkers and Inflammatory Infiltrates. *Front Immunol*, 13: 871008.
6. Huber S, Günther S, Cambria E, et al (2022). Physiological stretching induces a differential extracellular matrix gene expression response in acetabular labrum cells. *Eur Cell Mater*, 44:90-100.
7. Lu Y, Zhang H, Pan H, et al (2023). Expression pattern analysis of m6A regulators reveals IGF2BP3 as a key modulator in osteoarthritis synovial macrophages. *J Transl Med*, 21:(1):339.
8. Evans CH, Robbins PD (1999). Potential treatment of osteoarthritis by gene therapy. *Rheum Dis Clin North Am*, 25:(2):333-44.
9. Libbrecht MW, Noble WS (2015). Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16 (6):321-332.
10. Yoo C, Ramirez L, Liuzzi J (2014). Big data analysis using modern statistical and machine learning methods in medicine. *Int Neurourol J*, 18 (2):50-7.
11. Owusu-Adjei M, Ben Hayfron-Acquah J, Frimpong T, et al (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digit Health*, 2 (11):e0000290.
12. Li H, Cui Y, Wang J, et al (2024). Identification and validation of biomarkers related to lipid metabolism in osteoarthritis based on machine learning algorithms. *Lipids Health Dis*, 23 (1):111.
13. Li Y, Dong B (2024). Exploring liquid-liquid phase separation-related diagnostic biomarkers in osteoarthritis based on machine learning algorithms and experiment. *Immunobiology*, 229 (5):152825.
14. Xia L, Gong N (2022). Identification and verification of ferroptosis-related genes in the synovial tissue of osteoarthritis using bioinformatics analysis. *Front Mol Biosci*, 29:9:992044.
15. Yongming L, Yizhe X, Zhikai Q, et al (2024). Identification of ion channel-related genes as diagnostic markers and potential therapeutic targets for osteoarthritis through bioinformatics and machine learning-based approaches. *Biomarkers*, 29(5):285-297.
16. Hou Y, Yang Z, Ma J, et al (2025). Screening of potential biomarkers of osteoarthritis: a bioinformatics analysis. *Clin Rheumatol*, 44 (1):453-463.
17. Byrne BM, Oakley GG (2019). Replication protein A, the laxative that keeps DNA

regular: The importance of RPA phosphorylation in maintaining genome stability. *Semin Cell Dev Biol*, 86:112-120.
18. Dai Z, Wang S, Zhang W, et al (2017). Elevated Expression of RPA3 Is Involved in Gastric Cancer Tumorigenesis and Associated with Poor Patient Survival. *Dig Dis Sci*, 62 (9):2369-2375.
19. Qu C, Zhao Y, Feng G, et al (2017). RPA3 is a potential marker of prognosis and radioresistance for nasopharyngeal carcinoma. *J Cell Mol Med*, 21 (11):2872-2883.
20. Juge PA, Sparks JA, Gazal S, et al (2024). RPA3-UMAD1 rs12702634 and rheumatoid arthritis-associated interstitial lung disease in European ancestry. *Rheumatol Adv Pract*, 8(2):rkae059.
21. Sampaolo S, Napolitano F, Tirozzi A, et al (2017). Identification of the first dominant mutation of LAMA5 gene causing a complex multisystem syndrome due to dysfunction of the extracellular matrix. *J Med Genet*, 54 (10):710-720.
22. Wang YX, Zhao ZD, Wang Q, et al (2020). Biological potential alterations of migratory chondrogenic progenitor cells during knee osteoarthritic progression. *Arthritis Res Ther*, 22 (1):62.
23. Ou Y, Wang SJ, Li D, et al (2016). Activation of SAT1 engages polyamine metabolism with p53-mediated ferroptotic responses. *Proc Natl Acad Sci U S A*, 113(44):E6806-E6812.
24. Xu J, Ruan Z, Guo Z, et al (2024). Inhibition of SAT1 alleviates chondrocyte inflammation and ferroptosis by repressing ALOX15 expression and activating the Nrf2 pathway. *Bone Joint Res*, 13 (3):110-123.
25. Mukherjee S, Yun JW (2021). Novel regulatory roles of UCP1 in osteoblasts. *Life Sci*, 276:119427.
26. Kim S, Myers L, Ravussin E, et al (2016). Single nucleotide polymorphisms linked to mitochondrial uncoupling protein genes UCP2 and UCP3 affect mitochondrial metabolism and healthy aging in female nonagenarians. *Biogerontology*, 17 (4):725-36.
27. Pelsma ICM, Claessen K, Slagboom PE, et al (2021). Variants of FOXO3 and RPA3 genes affecting IGF-1 levels alter the risk of development of primary osteoarthritis. *Eur J Endocrinol*, 184 (1):29-39.
28. Zubair IM, Kim B (2022). A Group Feature Ranking and Selection Method Based on Dimension Reduction Technique in High-Dimensional Data. *IEEE Access*, 10 125136-47.
29. Alvari G, Furlanello C, Venuti P (2021). Is Smiling the Key? Machine Learning Analytics Detect Subtle Patterns in Micro-Expressions of Infants with ASD. *J Clin Med*, 10(8):1776.
30. Katschke KJ Jr, Rottman JB, Ruth JH, et al (2001). Differential expression of chemokine receptors on peripheral blood, synovial fluid, and synovial tissue monocytes/macrophages in rheumatoid arthritis. *Arthritis Rheum*, 44 (5):1022-32.
31. de Lange-Brokaar BJ, Ioan-Facsinay A, van Osch GJ, et al (2012). Synovial inflammation, immune cells and their cytokines in osteoarthritis: a review. *Osteoarthritis Cartilage*, 20 (12):1484-99.
32. Li YS, Luo W, Zhu SA, et al (2017). T Cells in Osteoarthritis: Alterations and Beyond. *Front Immunol*, 8:356.
33. Zhang Z, Ma X, Zha Z, et al (2022). The protective effects of allopurinol against IL-17A-induced inflammatory response in mast cells. *Mol Immunol*, 141:53-59.