Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Identification and verification of four candidate biomarkers for early diagnosis of osteoarthritis by machine learning

Xinyu Wang[a,b,1], Tianyi Liu[a,c,d,1], Yueyang Sheng[a], Yanzhuo Zhang[a], Cheng Qiu[e], Manyu Li[f], Yuxi Cheng[g], Shan Li[a], Ying Wang[a], Chengai Wu[a,*]

[a] *Department of Molecular Orthopaedics, National Center for Orthopaedics, Beijing Research Institute of Traumatology and Orthopaedics, Beijing Jishuitan Hospital, Capital Medical University, Beijing, 100035, China*
[b] *Department of Anesthesiology, National Center for Orthopaedics, Beijing Jishuitan Hospital, Capital Medical University, Beijing, 100035, China*
[c] *Department of Medical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China*
[d] *Department of Hepatobiliary Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China*
[e] *Department of Orthopaedic Surgery, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, Shandong, 250012, China*
[f] *Department of Gastroenterology, Qilu Hospital of Shandong University, Jinan, Shandong, 250012, China*
[g] *Xiangya Stomatological Hospital & Xiangya School of Stomatology, Central South University, Changsha, Hunan, 410008, China*

ARTICLE INFO

Keywords:
Osteoarthritis
Early diagnosis
Chondrocyte
Biomarker
Machine learning
B3GALNT1
GRB10
KLF9
SCRG1
ZNF423

ABSTRACT

Background: Osteoarthritis (OA) is a common chronic joint disease. This study aimed to investigate possible OA diagnostic biomarkers and to verify their significance in clinical samples.
Methods: We exploited three datasets from the Gene Expression Omnibus (GEO) database, serving as the training set. We first determined differentially expressed genes and screened candidate diagnostic biomarkers by applying three machine learning algorithms (Random Forest, Least Absolute Shrinkage and Selection Operator logistic regression, Support Vector Machine-Recursive Feature Elimination). Another GEO dataset was used as the validation set. The test set consisted of RNA-sequenced peripheral blood samples collected from patients and healthy donors. Blood samples and chondrocytes were collected for quantitative real-time PCR to confirm expression levels. Receiver operating characteristic curves were generated for individual and combined biomarkers.
Results: In total, 251 DEGs were screened, where *B3GALNT1*, *SCRG1* and *ZNF423* were screened by all three algorithms. The area under the curve (AUC) of various biomarkers in our test set did not reach as high as that in public datasets. *GRB10* exhibited highest AUC of 0.947 in the training set but 0.691 in our test set, while the favorable combined model comprising *B3GALNT1*, *GRB10*, *KLF9* and *SCRG1* demonstrated an AUC of 0.986 in the training set, 1.000 in the validation set and 0.836 in our test set.
Conclusion: We identified a combined model for early diagnosis of OA that includes *B3GALNT1*, *GRB10*, *KLF9* and *SCRG1*. This finding offers new avenues for further exploration of mechanisms underlying OA.

[*] Corresponding author. Department of Molecular Orthopaedics, Beijing Research Institute of Traumatology and Orthopaedics, Beijing Jishuitan Hospital, Capital Medical University, Beijing, 100035, China.

E-mail address: wuchengai05@163.com (C. Wu).

[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.heliyon.2024.e35121
Received 12 April 2024; Received in revised form 22 July 2024; Accepted 23 July 2024
Available online 23 July 2024

1. Introduction

Osteoarthritis (OA) is a common chronic joint disease that results from a dysregulation of joint homeostasis. In 2019, 528 million people worldwide were affected by OA, a 113 % increase from 1990 [1]. Approximately 73 % of people with osteoarthritis were aged 55 years or older, and 60 % were female [2,3]. The knee is the most typically afflicted joint, followed by the hip and hand [4]. Roughly half of those over 45 with knee OA have other associated symptoms, and about 30 % of them have radiographic evidence of the condition [5]. Approximately 10 % of people have symptomatic and radiographic hip OA [5]. At present, there are no curative treatments available for OA, and management strategies mainly focus on alleviating symptoms through pain relief and total joint replacement [6].

The absence of curative treatments is explained by the complexity of the disease's etiology and pathogenesis, as well as the diversity of its phenotypes and endotypes. The exact cause of OA is still unknown, though several variables such as advancing age, gender, joint malalignment, obesity, and genetic susceptibility have been linked to the development of the disease. The pathophysiology of OA is complex and influences a variety of joints and various joint tissues such as ligaments, menisci, cartilage, the subchondral bone, the synovial membrane, and the joint capsule [7]. The prevailing pathogenic hypothesis suggests that OA begins with an initial injury, often biomechanical, to any of these structures, which then initiates the release of the medium that activates different inflammatory pathways, ultimately leading to cartilage damage [8]. Yet given that the etiology and pathogenesis of OA are so poorly understood and the phenotypes and endotypes of OA are diverse, the development of disease-modifying therapeutic approaches is challenging [9].

Currently, clinical symptoms are used to diagnose OA, which in some countries is supplemented by joint imaging techniques [10]. But an absence of alternative diagnostic procedures and a linked propensity for OA patients to present significantly after the onset of the condition has led OA to be considered a "silent" disease, in which symptoms and imaging changes often only become apparent once significant damage to the articular cartilage has occurred. Unfortunately, any long-term subclinical phase can result in permanent damage, and without an established diagnosis at an early stage, the disease tends to advance quickly in most patients, leading to a poor prognosis [10]. Therefore, an early and accurate diagnosis of OA is particularly important.

Biomarkers that indicate articular cartilage degeneration offer great potential in enabling such diagnoses and more effective therapeutic interventions for OA. Finding these biomarkers has become easier with recent large-scale, high-throughput gene chip analyses of diseased and healthy human specimens. Such analyses facilitate the study of diseases across epigenetic variants, at transcriptome levels, and through somatic mutations [11]. Based on these analyses, recent studies have indicated that the development of OA is influenced by several distinct genes. For instance, Liang et al. found that *APOLD1* (apolipoprotein L domain containing 1) and *EPYC* (epiphycan) are significant OA diagnostic genes [12]. By analyzing the Genome Expression Omnibus (GEO) database, Deng et al. also demonstrated that *GRB10* (growth factor receptor bound protein 10) and *E2F3* (E2F transcription factor 3) could serve as diagnostic markers for OA [13]. While preliminary studies have indicated that a few functioning genes are crucial to the development of OA, they have not been fully explored as a diagnostic tool. In addition, most of the datasets used in the above studies were mRNA expression datasets derived from articular cartilage or synovial tissue, which limits the relevance of the findings to clinical

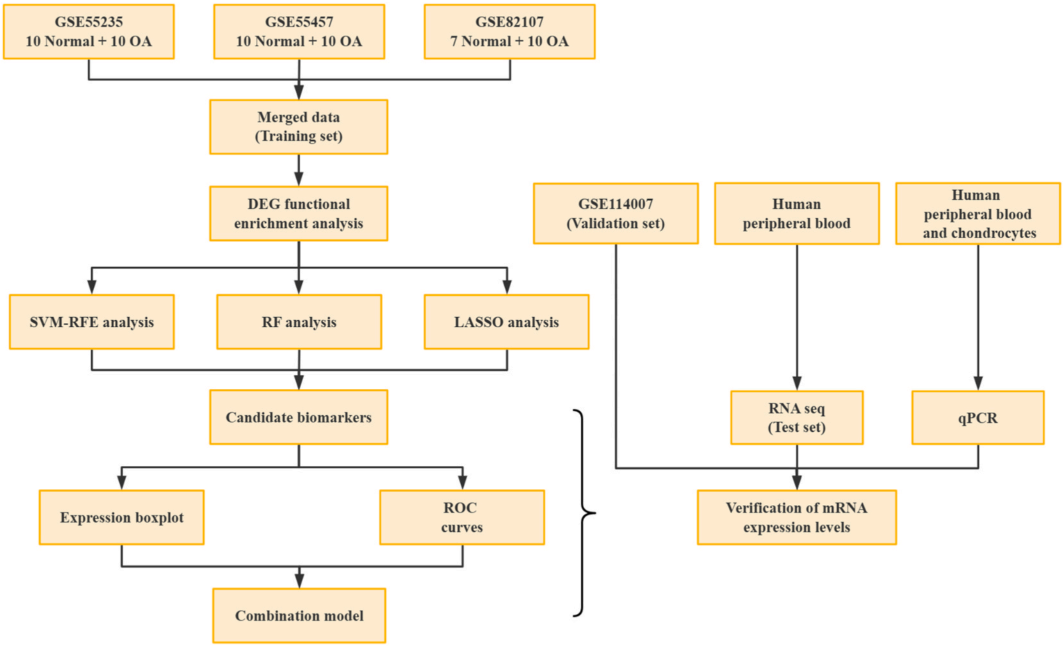


Fig. 1. General study flow chart. This schematic representation outlines the research procedure, consisting of three main steps: (1) screening for candidate biomarkers in the training set; (2) evaluation of the model performance in the validation set (GSE114007) and the test set (real-world patients from our hospital); and (3) verification of candidate biomarkers by qPCR.

applications.

In this work, we employed three machine learning algorithms, random forest (RF), least absolute shrinkage and selection operator (LASSO) logistic regression, and support vector machine-recursive feature elimination (SVM-RFE), to determine biomarkers that could predict OA diagnoses. Specifically, RF was selected for its ability to handle high-dimensional data and its robustness against overfitting. LASSO logistic regression was chosen for its capability to perform variable selection and regularization to enhance predictive accuracy. SVM-RFE was included for its efficiency in feature selection, particularly in identifying the most relevant features for classification tasks. These algorithms complement each other, allowing for a comprehensive and robust identification of candidate biomarkers. We first extracted OA microarray datasets from the GEO database, conducted differentially expressed gene (DEG) analysis on these datasets, and subsequently performed functional enrichment analyses on the DEGs before submitting the data for machine learning algorithmic analysis. We plotted receiver operating characteristic (ROC) curves and calculated areas under the curve (AUCs) with a 95 % confidence interval (CI) to evaluate the accuracy and reliability of the candidate biomarkers. For corroboration and to maximize the diagnostic value of differential candidate biomarkers, we collected clinical peripheral blood samples from both healthy donors and OA patients, as well as chondrocytes from knee OA patients (Fig. 1). The peripheral blood samples enabled us to better detect gene expression levels than in damaged cartilage and will be easier to diagnose by combining multiple significant candidate biomarkers into a favorable combined model.

In clinical practice, the absence of early diagnostic biomarkers often leads to a lost opportunity for timely treatment in OA patients, consequently resulting in a suboptimal prognosis [14]. Our identification of diagnostic biomarkers in this study therefore has significant importance in facilitating early diagnosis and enhancing OA prognoses.

2. Material and methods

2.1. Data acquisition and processing

The gene expression profiling of a total of twenty-seven healthy samples together with thirty OA samples derived from three microarray datasets were downloaded from the GEO database [15,16], including the GSE55235 [17], GSE55457 [17] and GSE82107 [18] datasets. Ten synovial samples from OA patients and ten from healthy subjects were enrolled from both GSE55235 and GSE55457 datasets, respectively. Seven healthy samples and ten OA samples were acquired from the GSE82107 dataset. Both GSE55235 and GSE55457 datasets were annotated using the GPL96 platform and the GSE82107 dataset was annotated using the GPL570 platform. When more than one probe was assigned to a single gene, the mean expression value was calculated. After normalization, the three datasets were merged together to form the training set, and then batch effects were removed for further analyses. The normalized expression data of another dataset, GSE114007 [19], with twenty OA samples and eighteen healthy samples overall, was downloaded as well to serve as the validation set. R software (version 4.2.2) was used for data processing, where the “GEOquery” [16], “stringr”, “limma” [20], “dplyr”, and “sva” packages were applied during this step.

2.2. DEG analysis and exploration of candidate diagnostic biomarker

The “limma” package was used to determine DEGs in the training set between healthy and OA samples [20]. Genes with an adjusted P (adj.P) value < 0.05 and a $|\log2$ fold change (FC)$| > 1$ were classified as DEGs. The “ggplot2” package was used to draw a volcano plot displaying the up-regulated or down-regulated genes. The “clusterProfiler” [21] and “ggplot2” packages were used to conduct and visualize the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of up- and down-regulated DEGs. The RF, LASSO logistic regression, and SVM-RFE algorithms [22] were applied for the identification of OA biomarkers. The RF algorithm was conducted with the “randomForest” package, selecting the number of tree nodes that minimized training error after training a random forest model containing 500 trees. We used grid search to optimize the number of trees and the maximum depth. LASSO regression was executed with the “glmnet” package, employing a ten-fold cross-validation to select the optimal lambda value that minimized cross-validated error [23]. The SVM-RFE algorithm was implemented with the “caret” and “e1071” packages, employing a ten-fold cross-validation strategy with grid search to select the best set of features. During cross-validation, we recorded the average RMSE (Root Mean Squared Error) for various sets of potential genes, and the set with the minimum RMSE was selected as the ultimate biomarker set. Additionally, we highlighted the use of k-fold cross-validation to prevent overfitting and to ensure robust model performance. The overlapping genes selected by the three algorithms were considered promising candidate biomarkers and depicted in a Venn diagram generated using the “venn” and “VennDiagram” packages. ROC curves were plotted for each candidate biomarker and combinations thereof, with respective AUCs calculated using the “pROC” package to estimate the biomarkers’ predictive properties. The expression of candidate biomarkers in healthy and OA samples was visualized using boxplots.

2.3. Collection of blood samples

The study subjects for this study were patients with OA aged between 52 and 70 years awaiting artificial joint replacement in Jishuitan Hospital. Inclusion criteria were as follows: (1) the diagnosis was in accordance with the Orthopaedic Branch of the Chinese Medical Association; and (2) the patient was not yet being treated with nonsteroidal anti-inflammatory drugs (NSAIDs). Exclusion criteria included: (1) those with a history of severe infection, surgery, or tumor; (2) those with severe cardiac, pulmonary, hepatic, or renal dysfunction; and (3) those with a history of autoimmune diseases such as rheumatoid arthritis and ankylosing spondylitis. The study also included a control group consisting of healthy volunteers between the ages of 28 and 62 who had not received medical treatment in the last 6 months, and who were without any other chronic disease (such as coronary artery disease, diabetes mellitus, or hypertension) or osteoarticular diseases (including skeletal fluorosis, gout, or rheumatoid arthritis). All participants were of Chinese Han ancestry, detailed in Table 1. Written informed consent forms were signed by all participants. Our trials complied with the Declaration of Helsinki and were authorized through the Ethics Committee of Beijing Jishuitan Hospital (ethics code: 201611–03).

2.4. Cell isolation and culture

Human chondrocytes were obtained from the OA patients' knee joints at the time of total joint replacement procedures. The chondrocytes were separated from the cartilage fragments of severely damaged tissues on the inner side of the tibial plateau and the typically undamaged tissues on the outer side of the tibial plateau, which were separated into normal controls and OA samples based on the degree of surface cartilage degradation and relative smoothness. Within an hour of the surgical procedure, all articular cartilage samples, namely, subchondral bone together with all cartilage zones (including calcified ones), were taken from the lateral tibial plateau. The procedures applied to extract chondrocytes were as below [24]: articular cartilage specimens were cut into pieces (1 mm^3), twice-cleaned in pre-cooled DMEM media, and later enzyme-digested with 0.25 % trypsin in a 5 % CO_2 environment at 37 °C for a maximum of 30 min. Following a 5-min centrifugation at 1000 × *g*, the supernatant was fully discarded, and the cell suspensions were digested in basal medium with 0.2 % type II collagenase added for 3–6 h at 37 °C. After separating the chondrocytes and filtering them using 70 μm nylon filters, we twice-cleaned the cells through sterile phosphate-buffered saline (PBS), and grew them in culture media with 1 % penicillin-streptomycin and 10 % fetal bovine serum in an incubator set at 37 °C with 5 % CO_2. The cells were digested utilizing trypsin (Gibco, USA) and then passaged at a 1:3 ratio when they reached 80–90 % confluency. We only used the first passage in our in vitro experiments.

2.5. RNA sequencing (RNA-seq) assay

All the collected blood samples were processed on the HiSeq platform, using an Illumina TruseqTM RNA sample preparation Kit. After library building, SeqPrep (https://github.com/jstjohn/SeqPrep) was used to analyze the raw data and acquire qualified mRNA and lncRNA sequences for subsequent analysis. The candidate biomarkers identified from the training set were verified in our own transcriptome sequencing data, which served as the test set.

Table 1
The age, gender, and living habits of OA patients and healthy volunteers in the test set.

Sample	Age (years)	Gender	Alcohol Consumption	Smoking Status
		RNA sequencing assay		
Normal 1	49	Female	Non-drinker	Never smoked
Normal 2	62	Female	Non-drinker	Never smoked
Normal 3	53	Female	Non-drinker	Never smoked
Normal 4	53	Male	Moderate drinker	Current smoker
Normal 5	58	Female	Non-drinker	Never smoked
OA 1	69	Female	Non-drinker	Never smoked
OA 2	68	Female	Non-drinker	Never smoked
OA 3	58	Female	Moderate drinker	Never smoked
OA 4	63	Female	Moderate drinker	Former smoker
OA 5	70	Female	Non-drinker	Current smoker
OA 6	61	Female	Non-drinker	Former smoker
OA 7	68	Male	Heavy drinker	Former smoker
OA 8	69	Female	Non-drinker	Never smoked
OA 9	63	Female	Non-drinker	Never smoked
OA 10	55	Female	Non-drinker	Former smoker
OA 11	69	Male	Moderate drinker	Current smoker
		qPCR		
Normal 1	27	Female	Non-drinker	Never smoked
Normal 2	25	Female	Non-drinker	Never smoked
Normal 3	26	Female	Non-drinker	Never smoked
Normal 4	27	Male	Moderate drinker	Current smoker
Normal 5	33	Male	Non-drinker	Former smoker
OA 1	60	Male	Moderate drinker	Current smoker
OA 2	56	Male	Non-drinker	Former smoker
OA 3	52	Female	Non-drinker	Never smoked
OA 4	54	Female	Non-drinker	Never smoked
OA 5	55	Female	Non-drinker	Never smoked

Moderate drinker: up to 1 drink per day for women, up to 2 drinks per day for men.
Heavy drinker: more than moderate drinking.

2.6. Quantitative real-time PCR (qPCR) analysis of cells

Six-well plates were seeded with chondrocytes, which were subsequently cultured in the full culture media. As directed by the manufacturer, the total RNA was extracted from chondrocytes with the MiniBEST Universal RNA Extraction Kit (TaKaRa, China). To create cDNA and prevent DNA contamination, the PrimeScript™ RT reagent kit with gDNA Eraser (TaKaRa, China) was used. A TB Green Premix Ex Taq (TaKaRa, China) along with an Applied Biosystems 7500 Real-Time PCR System was employed for real-time PCR in accordance with the manufacturers' instructions. The comparative Ct ($2^{-\Delta\Delta Ct}$) approach was used to examine the data and determine the relative gene expression. Table 2 includes a list of primer sequences.

2.7. Quantitative real-time PCR analysis of blood

RNA was extracted from the blood of both OA patients and healthy donors with the TRIZOL reagent (Invitrogen) following the supplier's instructions. Using a HyperScript III RT SuperMix for qPCR with gDNA Remover (NovaBio R202), RNA totaling (1 μg) was reverse transcribed into cDNA in a final amount of 20 μl by random primers. In accordance with the supplier's instructions, the reverse transcription procedure was finished for 15 min at 37 °C and later for 5 s at 85 °C. The qPCR analysis was accomplished using 2 × S6 Universal SYBR qPCR Mix (NovaBio Q204) following the supplier's instructions. Data collection and qPCR analysis were carried out with an ABI 7900HT device. The comparative Ct ($2^{-\Delta\Delta Ct}$) approach was again used to examine these data and determine the relative gene expression. Table 2 presents a list of primer sequences. For accuracy, each specimen underwent three separate analyses.

2.8. Statistical analysis

Prism 8 (GraphPad, San Diego, USA) was employed to evaluate all data, which were then exhibited as the mean ± standard deviation (SD). To examine differences based on experimental design, a two-tailed *t*-test with Tukey post hoc testing was employed for multiple comparisons. To assure dependability, all experiments were conducted at least three times. Unless otherwise noted, data were deemed statistically significant at a significance level of $P < 0.05$.

3. Results

3.1. DEGs between OA and normal samples and enrichment analyses

Using the criterion set out in the method, 251 genes in total were found to have differential expression. Specifically, 147 genes were found to be up-regulated and 104 down-regulated in OA samples in comparison with healthy counterparts. As shown in the volcano plot, *SCRG1* (stimulator of chondrogenesis 1), *B3GALNT1* (beta-1,3-N-acetylgalactosaminyltransferase 1), *TNFSF11* (TNF superfamily member 11), *KDELR3* (KDEL endoplasmic reticulum protein retention receptor 3) and *CX3CR1* (C-X3-C motif chemokine receptor 1) were identified as the top five up-regulated genes with the minimum *P* value while *GRB10, KLF9* (KLF transcription factor 9), *BCL6* (BCL6 transcription repressor), *SIK1* (salt inducible kinase 1) and *APOLD1* were identified as the top five down-regulated genes with the minimum *P* value (Fig. 2A). After unmapped genes were removed, it was observed that more than a tenth of the up-regulated genes were enriched to leukocyte migration (16/121), followed by collagen-containing extracellular matrix (14/121) and signaling receptor activator activity (14/121), etc. (Fig. 2B). The KEGG enrichment results demonstrated that up-regulated genes were mostly associated with osteoclast differentiation and rheumatoid arthritis, etc. (Fig. 2C). Conversely, GO enrichment analyses suggested the down-regulated DEGs were related to response to nutrient levels (Fig. 2D), while the tumor necrosis factor (TNF) signaling pathway, the interleukin-17 (IL-17) signaling pathway, the NF-κB signaling pathway and the NOD-like receptor signaling pathway, etc. were enriched by down-regulated DEGs according to KEGG analysis results (Fig. 2E).

Table 2
The primers used in quantitative real-time PCR studies.

Target	Sequence (5′-3′)
GAPDH-F	AAGGGTCATCATCTCTGCCC
GAPDH-R	GTGAGTGCATGGACTGTGGT
SCRG1-F	TCACCATTGGGCTAACTTTGC
SCRG1-R	GAAGGTTGTGACAGTTGTGATCT
B3GALNT1-F	CTTCACACTTCGAGAGCATTCA
B3GALNT1-R	CCCCAAGTAACTCTAATGGCCT
ZNF423-F	GGAACAGCGTGACAAGTCAAG
ZNF423-R	ACAGTGATCGCAGGTGTAAATTG
SIK1–F	GCTTCTGAACCATCCACACAT
SIK1-R	GTGCCCGTTGGAAGTCAAATA
GRB10-F	CTCGTGGCAATGGATTTTTCTG
GRB10-R	TCACTGTACTTAGGGTAGAAGGG
KLF9-F	GCCGCCTACATGGACTTCG
KLF9-R	GGATGGGTCGGTACTTGTTCA

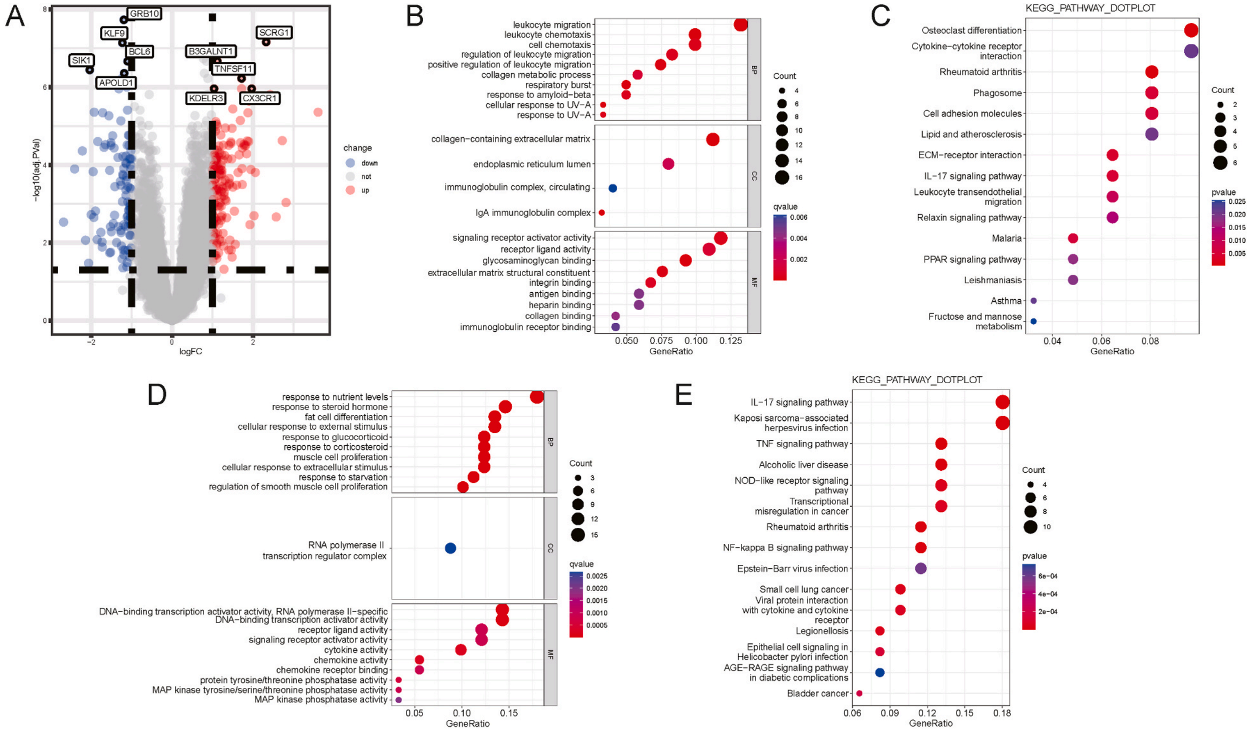


Fig. 2. DEGs between normal and OA samples and results of enrichment analyses. (A) Volcano plot of DEGs between normal and OA samples. Healthy samples serve as the control group. Up-regulated and down-regulated DEGs in the OA samples are indicated by red and blue dots, separately. The top five up-regulated and down-regulated DEGs are labelled. (B) Bubble plots describing GO enrichment analysis and (C) KEGG enrichment analysis of up-regulated DEGs in OA samples. (D) Bubble plots presenting the GO enrichment analysis and (E) KEGG enrichment analysis of down-regulated DEGs in OA samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.2. Identification of candidate diagnostic biomarkers

The SVM-RFE algorithm characterized 22 DEGs as promising diagnostic predictive biomarkers, as verified by the ten-fold cross-validation to have the lowest RMSE (Fig. 3A). These 22 candidate biomarkers included *GRB10*, *B3GALNT1*, *SCRG1*, *ZNF423* (zinc finger protein 423), *KLF9*, and *SIK1*, etc. The LASSO method filtered out 21 DEGs that could predict the diagnosis of OA with the lowest binomial deviance (Fig. 3B). This set of 21 biomarkers included *GRB10*, *KLF9*, *B3GALNT1* and *ZNF423*, etc. The RF algorithm discovered that 25 biomarkers could predict the clinical diagnosis with high accuracy (Fig. 3C). Among these biomarkers, *KLF9* may have the highest importance due to its greatest mean decrease in Gini value. *SDC1* (syndecan 1), *LTC4S* (leukotriene C4 synthase) and *LST1* (leukocyte specific transcript 1) followed closely in terms of importance (Fig. 3D). Collectively, the Venn plot indicated three overlapping identified biomarkers, namely *B3GALNT1*, *ZNF423* and *KLF9*. These three biomarkers were consistently identified by the SVM-RFE, LASSO and RF three algorithms, highlighting their significance in predicting clinical diagnosis (Fig. 3E). We then checked the expression levels and AUCs in the training set for the three identified biomarkers, as well as other DEGs that were identified by at least two of the algorithms. The results demonstrated that *B3GALNT1*, *SCRG1* and *ZNF423* were all significantly up-regulated in OA samples (Fig. 4A–C) while *GRB10*, *KLF9* and *SIK1* were all significantly down-regulated (Fig. 4D–F). Among these, *GRB10* exhibited the largest AUC of 0.947 (95 % CI: 0.883–0.991), indicating its strong diagnostic potential (Fig. 4G) whereas *ZNF423* had the lowest AUC of 0.886 (95 % CI: 0.789–0.962) (Fig. 4H). The AUCs for the other four factors were all above 0.9, indicating their value for high diagnostic accuracy (Fig. 4I–L) (Table 3).

3.3. Preliminary evaluation of candidate biomarkers in the validation set

B3GALNT1, SCRG1 and *ZNF423* expression levels in the validation set (Fig. 5A–C) were remarkably higher in OA samples versus the healthy control samples. However, the *GRB10* expression showed no significant difference between the two groups (Fig. 5D). On the other hand, *KLF9* and *SIK1* were remarkably higher in healthy control samples (Fig. 5E–F). Overall, other than *GRB10*, the remaining candidate biomarkers had consistent expression levels between the validation and training sets (Table 3). The performance of *GRB10* alone in the validation set, was not as good as that in training set, with an AUC of only 0.558 (95 % CI: 0.361–0.744) (Fig. 5G). A slight drop in the AUCs of *ZNF423* and *SCRG1* was also observed, but both AUCs were above 0.8 (0.808 for *SCRG1* and 0.828 for *ZNF423*) (Fig. 5H and J). Similar to that of *GRB10*, the AUC of *B3GALNT1* was sharply dropped to 0.728 (95 % CI:

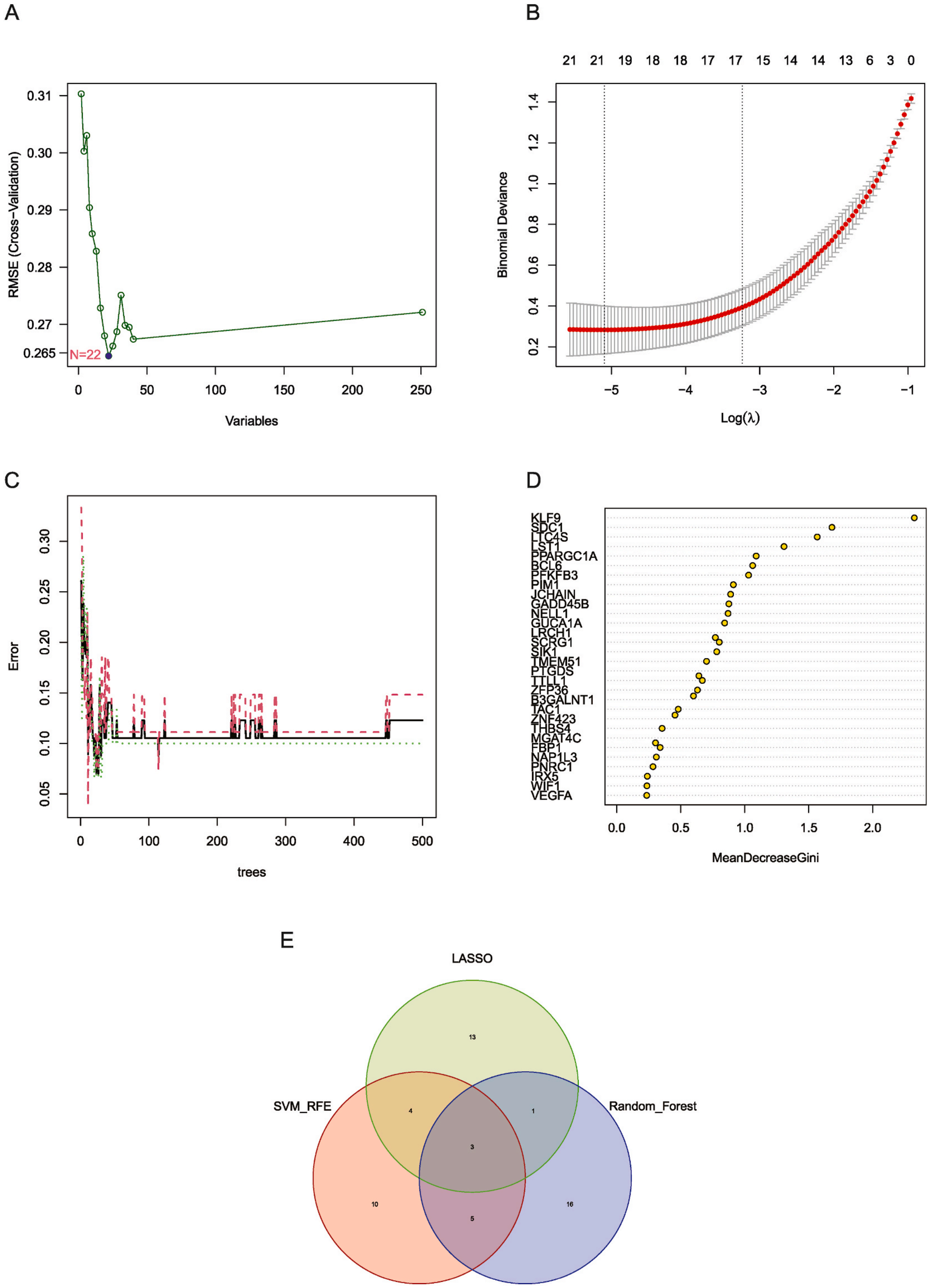


Fig. 3. Identification of candidate diagnostic biomarkers. (A) Candidate biomarkers identified by the SVM-RFE algorithm. All 251 DEGs are screened by SVM-RFE algorithm to find the feature set with the lowest RMSE (Root Mean Squared Error). (B) Candidate biomarkers filtered out by the LASSO method. The optimal lambda value represented by the dashed line on the left minimizes cross-validated error. (C) Candidate biomarkers discovered by the RF algorithm. The red dashed line represents the optimal tree number with the maximum depth. (D) Candidate biomarkers ranked by importance according to the mean decrease in Gini value as determined by the RF algorithm. (E) Venn diagram displaying the overlap of candidate diagnostic biomarkers derived from three algorithms. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

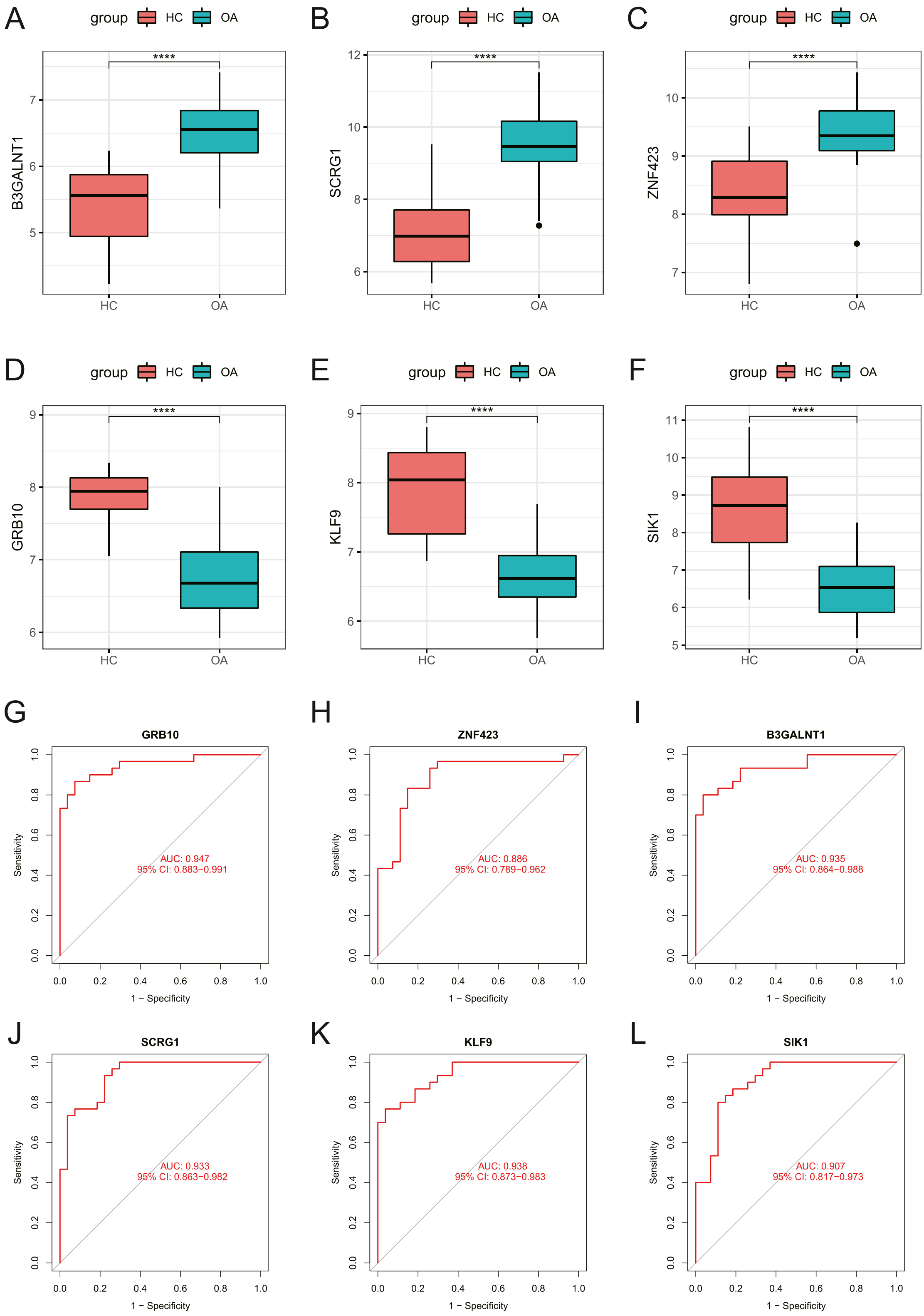


Fig. 4. Determination of candidate diagnostic biomarkers. (A–F) Expression levels of B3GALNT1, SCRG1, ZNF423, GRB10, KLF9 and SIK1 from the training set. The y axis indicates the expression level. (G–L) ROC curves of B3GALNT1, SCRG1, ZNF423, GRB10, KLF9 and SIK1. The x axis indicates false positive rate, calculated by 1-specificity, while the y axis indicates the specificity.

Table 3
The results of expression levels and AUCs of six candidate biomarkers.

	Training set		Validation set		Test set			
	Expression	AUC	Expression	AUC	Expression in blood by RNA-seq	AUC	Expression in chondrocytes by qPCR	Expression in blood by qPCR
B3GALNT1	High	0.935	High	0.728	Ns	0.636	High	High
GRB10	Low	0.947	Ns	0.558	Ns	0.691	High	High
KLF9	Low	0.938	Low	0.928	Ns	0.527	High	Low
SCRG1	High	0.933	High	0.808	Ns	0.636	High	High
SIK1	Low	0.907	Low	0.961	Ns	0.618	High	Low
ZNF423	High	0.886	High	0.828	Ns	0.582	High	High

Ns: not significant.

0.550–0.881) (Fig. 5I). The performance of *KLF 9* (AUC: 0.928; 95 % CI: 0.828–0.997) remained almost unchanged in the validation set (Fig. 5K). Conversely, *SIK1* had the largest AUC of 0.961 (95 % CI: 0.900–0.997) among all candidate biomarkers in the validation set (Fig. 5L) (Table 3).

3.4. *Verification of candidate biomarkers in the test set by RNA-seq assay*

The results of the peripheral blood RNA-seq assay suggested that there appeared to be no obvious differences in the expression levels of the six candidate biomarkers. However, subtle differences could still be observed. It appeared that the expression of *B3GALNT1*, *SCRG1*, *GRB10* and *ZNF423* showed a slight, albeit not significant, upregulation in OA samples (Fig. 6A–D). The other two biomarkers, *KLF9* and *SIK1*, demonstrated no significant difference (Fig. 6E–F). However, except for *KLF9*, the expression levels of the remaining biomarkers were generally low by the RNA-seq assay. The ROC curves indicated that the AUC of *B3GALNT1* was 0.636 (95 % CI: 0.291–0.927) (Fig. 6G), which was similar to *SCRG1* (95 % CI: 0.345–0.891) (Fig. 6H). *GRB10* exhibited the largest AUC of 0.691 (95 % CI: 0.364–1.000) (Fig. 6I). *ZNF423* possessed an AUC of 0.528 (95 % CI: 0.327–0.773) (Fig. 6J), which was slightly higher than that of *KLF9* (0.527; 95 % CI: 0.255–0.800) (Fig. 6K). In addition, *SIK1* exhibited an AUC of 0.618 (95 % CI: 0.273–0.927) (Fig. 6L), ranking the second highest among the six biomarkers. These biomarkers have shown some diagnostic efficacy in real-world patients to some extent.

3.5. *Intracellular and peripheral levels of candidate biomarkers*

We performed qPCR to detect the expression of *B3GALNT1*, *SCRG1*, *ZNF423*, *KLF9*, *SIK1*, and *GRB10* in knee chondrocytes from OA patients. The findings suggested that *ZNF423* ($P < 0.0001$), *SCRG1* ($P < 0.0001$), *B3GALNT1* ($P = 0.0002$), *GRB10* ($P < 0.0001$), *SIK1* ($P < 0.0001$) and *KLF9* ($P < 0.0001$) expression levels were evidently higher in OA samples versus normal samples (Fig. 7A–F). The results demonstrated that the relative expression of these three key biomarkers, namely *B3GALNT1*, *SCRG1*, and *ZNF423*, in OA cells were consistent with the findings from the bioinformatic analyses. We next performed qPCR analysis to examine the expression of the six candidate biomarkers in the peripheral blood of both healthy individuals and those with OA. Our findings revealed that *SCRG1* ($P < 0.0001$), *B3GALNT1* ($P < 0.0001$), *GRB10* ($P = 0.0001$), and *ZNF423* ($P = 0.0014$) expression levels were significantly higher in OA samples compared to healthy samples (Fig. 8A–C, 8F). However, *KLF9* ($P < 0.0001$) and *SIK1* ($P < 0.0001$) were significantly lower in OA samples compared to healthy samples (Fig. 8D–E).

3.6. *Efficacy evaluation of different diagnostic models with combined biomarkers*

Given these results, especially the performance of single candidate biomarkers, we tested several combinations of different candidate biomarkers. The results obtained from the training set demonstrated that the AUCs of a combination of two biomarkers, for example, B3GALNT1+SCRG1, GRB10+SCRG1, and GRB10+KLF9, were all above 0.95 (Fig. 9A–C), showing a favorable diagnostic ability. However, the combination of three or four candidate biomarkers could not significantly improve the diagnostic ability of the model in the training set (Fig. 9D–G). In the validation set, B3GALNT1+SCRG1 (AUC: 0.886; 95 % CI: 0.769–0.970) (Fig. 9H) performed relatively poorly, though most combination models could well verify the diagnostic capabilities of candidate biomarkers (Fig. 9I–N). Surprisingly, the combination of GRB10+KLF9+SIK1 exhibited an AUC of 1.000 (95 % CI: 1.000–1.000) (Fig. 9L). Finally, in our test set, the combination of two candidate biomarkers seemed ineffective. The AUCs of B3GALNT1+SCRG1 and GRB10+SCRG1 were 0.764 (95 % CI: 0.491–0.982) (Figs. 9O) and 0.782 (95 % CI: 0.509–1.000) (Fig. 9P), separately. Nevertheless, the addition of one more biomarker could not change the situation (Fig. 9Q–R). Therefore, four candidate biomarkers were therefore incorporated into the final model. Among all, the comparatively favorable model included B3GALNT1+GRB10+SCRG1+ZNF423. The AUC of the aforementioned model was 0.836 (95 % CI: 0.600–1.000) (Fig. 9S), and could be raised to 0.855 (95 % CI: 0.600–1.000) when *ZNF423* was replaced by *KLF9* (Fig. 9T), which owned the greatest AUC.

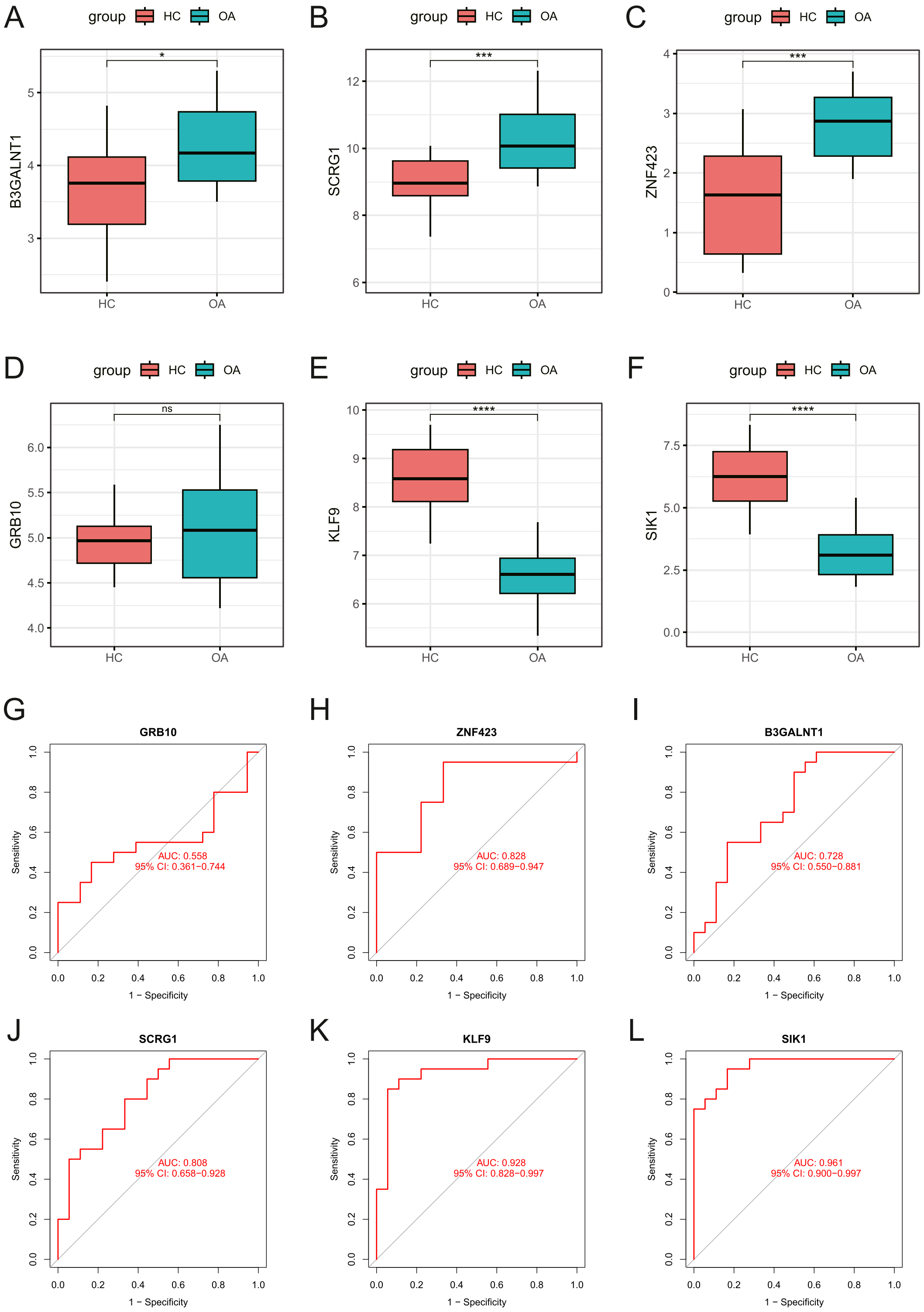


Fig. 5. Preliminary validation of candidate biomarkers in the validation set. (A–F) Expression levels of B3GALNT1, SCRG1, ZNF423, GRB10, KLF9 and SIK1 in the validation set. The y axis indicates the expression level. (G–L) ROC curves of B3GALNT1, SCRG1, ZNF423, GRB10, KLF9 and SIK1. The x axis indicates false positive rate, calculated by 1-specificity, while the y axis indicates the specificity.

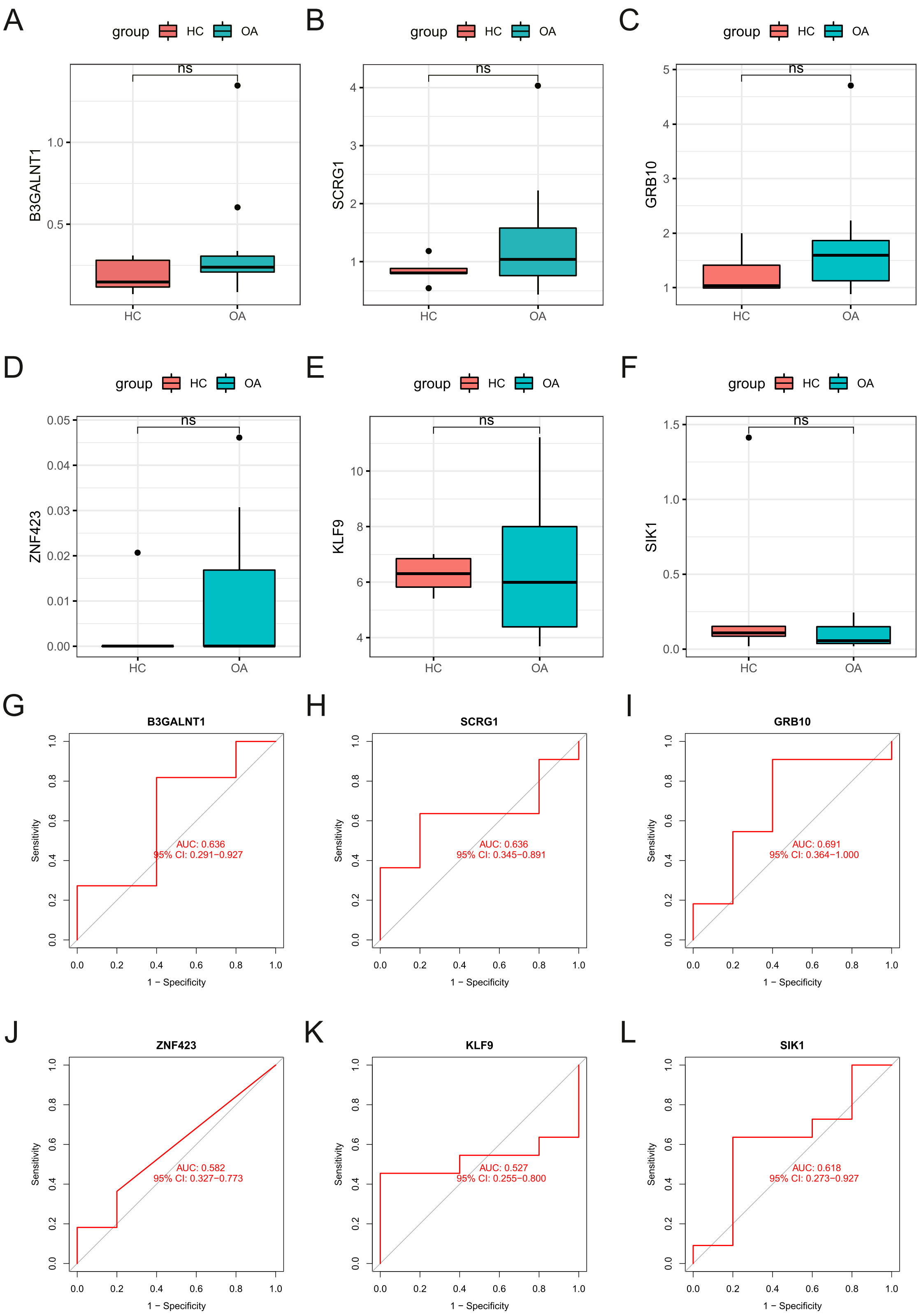


Fig. 6. Verification of candidate biomarkers in the test set. (A–F) Expression levels of B3GALNT1, SCRG1, ZNF423, GRB10, KLF9 and SIK1 in peripheral blood samples by RNA-seq assay. The y axis indicates the expression level. (G–L) ROC curves of B3GALNT1, SCRG1, ZNF423, GRB10, KLF9 and SIK1. The x axis indicates false positive rate, calculated by 1-specificity, while the y axis indicates the specificity.

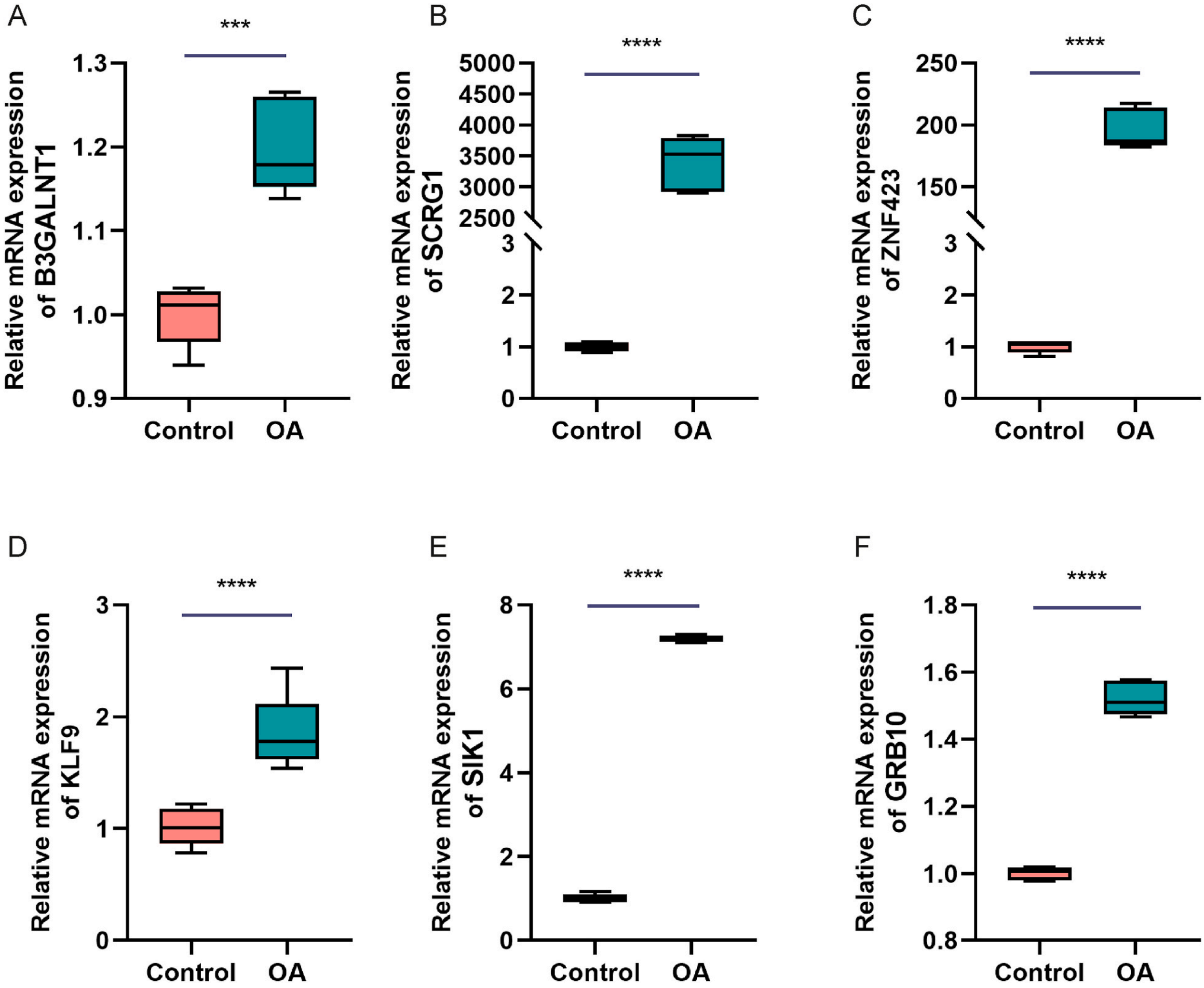


Fig. 7. Intracellular levels of candidate biomarkers in OA samples compared to the control group. (A–F) Quantitative real-time PCR analyses of B3GALNT1, SCRG1, ZNF423, KLF9, SIK1 and GRB10. ***P < 0.001; ****P < 0.0001.

4. Discussion

OA is a major health burden globally that results in severe pain, disability, and financial hardship for sufferers. The epidemiology of the disease is complex and influenced by various factors such as genetic, biological and biomechanical factors. OA was long considered an "abrasive" disease, but is now considered to have a complex pathophysiological process that impacts multiple joints and joint structures [5]. The International Association for the Study of Osteoarthritis finds that OA initially presents as a molecular disorder caused by abnormalities in the metabolism of joint tissue. This is subsequently followed by anatomical and/or physiological disorders, which are characterized by degeneration of cartilage, joint inflammation, bone redundancy, remodeling of bone, and the loss of normal joint function. Ultimately all of these factors contribute to the development of the disease [25]. Despite the fact that total joint arthroplasty is a valid treatment for end-stage OA, it does not always yield satisfactory outcomes, and the longevity of the prosthesis is limited [10]. In short, the current clinical diagnostic process falls short in adequately addressing the patients' need to minimize the risk of worsening of the disease, and in facilitating the development of innovative disease-modifying drugs for OA [26]. As a result, researches on OA have shifted the focus towards early diagnosis and treatment for the disease. In traditional treatment approaches, clinicians primarily rely on imaging and clinical symptoms to make a diagnosis. However, conventional imaging techniques are limited in their ability to detect early stages of the disease, although they are more effective in identifying advanced diseases. Additionally, there is not always a strong correlation between pain and pathologic changes in joint structures [27,28], which hinders detection of early stages of the disease. And in the early stages, the timely initiation of treatment with disease-relieving drugs may yield better therapeutic results. Thus, advancements in the early diagnosis of OA hold promise for the development of new and improved treatment methods.

This study is one such effort, aiming to identity diagnostic biomarkers for OA that could be used as a non-invasive tool for early

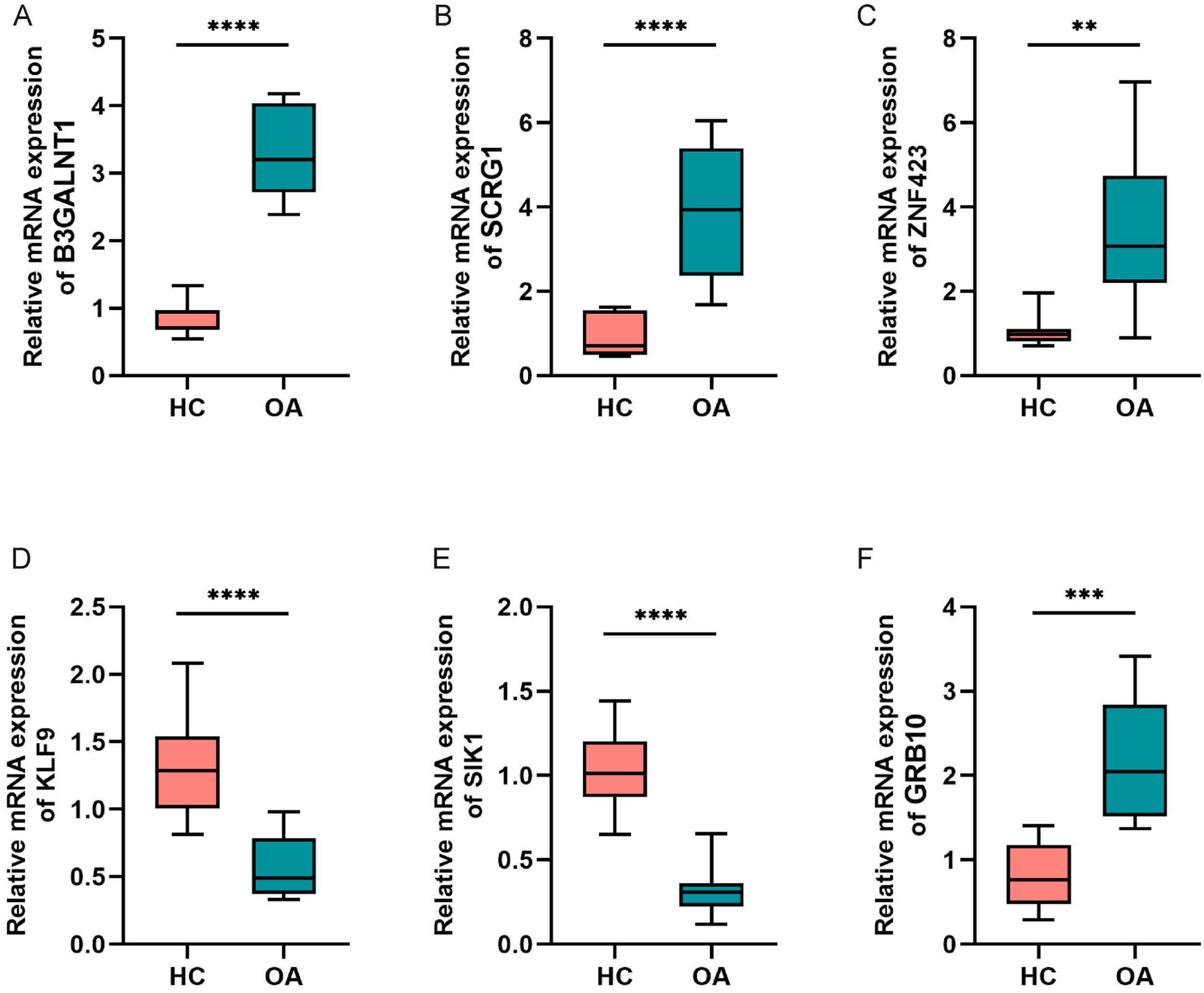
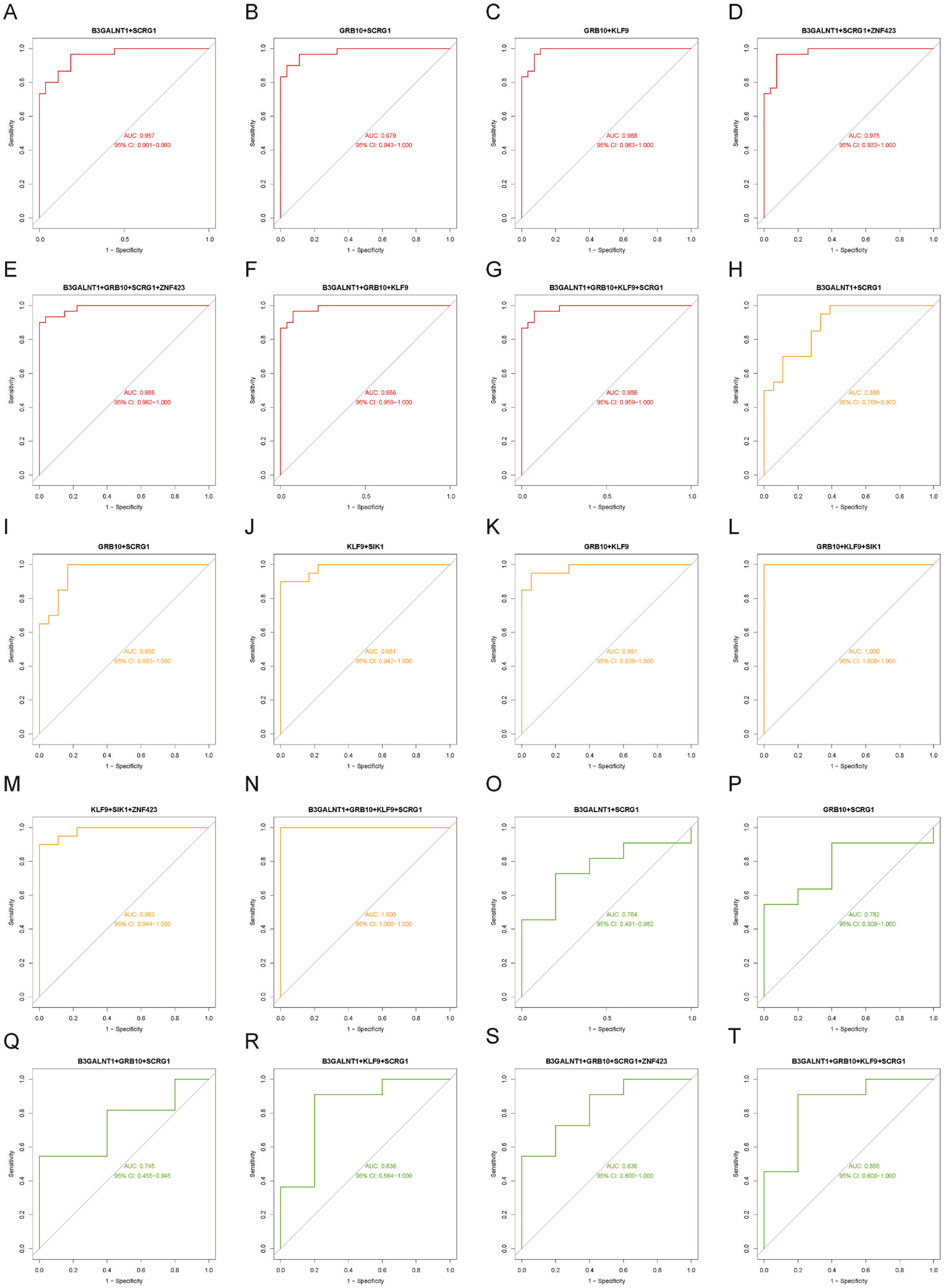


Fig. 8. Peripheral levels of candidate biomarkers in OA patients compared to the control group. (A–F) Quantitative real-time PCR analyses of B3GALNT1, SCRG1, ZNF423, KLF9, SIK1 and GRB10. **P < 0.01; ***P < 0.001; ****P < 0.0001.

diagnosis of OA, along with ongoing disease surveillance. To achieve this, we initially searched the relative datasets of GEO public database to determine the DEGs associated with OA in a training set. Then, the identified DEGs were feature selected using three independent algorithms, where LASSO tends to select sparse models, Random Forest is able to deal with non-linear relationships, and SVM-RFE is suitable for feature ranking and selecting the most relevant set of features. Thus, a total of six genes, *B3GALNT1*, *SCRG1*, *ZNF423*, *KLF9*, *SIK1*, and *GRB10*, were recognized as promising candidate diagnostic biomarkers for OA. Subsequently, boxplots and ROC curves were generated to assess the expression level and efficacy of these candidate biomarkers. Then another independent GEO dataset, GSE114007, was used as a validation set for further confirmation of the diagnostic capability of the candidate biomarkers. To confirm the above findings from the public database, the RNA sequencing profiles of blood samples of OA patients and healthy donors in our hospital were used as a test set. We found the combination of B3GALNT1+GRB10+KLF9+SCRG1 showed the best diagnostic potential across the test, validation and training sets.

The qPCR expression verification was divided into two parts: intracellular and peripheral blood validation. In the intracellular qPCR validation, we observed that the differences in expression levels of *B3GALNT1*, *SCRG1* and *ZNF423* between normal and OA samples were consistent with the bioinformatics analysis. In addition, the qPCR verification using peripheral blood samples showed the expression levels of *B3GALNT1*, *SCRG1*, *ZNF423*, *KLF9*, and *SIK1*, were consistent with the bioinformatic analysis.

Due to the potential impact of various factors on expression results, our qPCR verification using chondrocytes yielded conclusions that were inconsistent with results of RNA-seq assay. For example, the properties of chondrocytes can be influenced by differences in cell sources, varying from donor to donor. Furthermore, cell characteristics can undergo changes under diverse in vitro culture conditions, potentially influencing the outcomes of expression verification. Additionally, the intricate biological processes of chondrocytes involve numerous cell signaling pathways and molecular mechanisms, making it challenging to precisely replicate them during expression verification, which could affect the verification results. In our verification experiments, the limited availability of



(caption on next page)

Fig. 9. Evaluation of the efficacy of different diagnostic models using combined biomarkers. (A–G) ROC curves of B3GALNT1+SCRG1, GRB10+SCRG1, GRB10+KLF9, B3GALNT1+SCRG1+ZNF423, B3GALNT1+GRB10+SCRG1+ZNF423, B3GALNT1+GRB10+KLF9, B3GALNT1+GRB10+KLF9+SCRG1 in the training set. (H–N) ROC curves of B3GALNT1+SCRG1, GRB10+SCRG1, KLF9+SIK1, GRB10+KLF9, GRB10+KLF9+SIK1, KLF9+SIK1+ZNF423, B3GALNT1+GRB10+KLF9+SCRG1 in the validation set. (O–T) ROC curves of B3GALNT1+SCRG1, GRB10+SCRG1, B3GALNT1+GRB10+SCRG1, B3GALNT1+KLF9+SCRG1, B3GALNT1+GRB10+SCRG1+ZNF423, B3GALNT1+GRB10+KLF9+SCRG1 in the test set. The x axis indicates false positive rate, calculated by 1-specificity, while the y axis indicates the specificity.

chondrocyte samples may have had an impact on the expression verification results. Nevertheless, the qPCR results from blood samples in the test set remained consistent with both the training and validation sets, thereby providing support for our conclusions.

In summary, *GRB10* showed a consistent trend in the qPCR expression verification, albeit opposite to the bioinformatic analysis. *SIK1* and *KLF9* exhibited expression levels in blood that were consistent with the bioinformatic analysis. Only three biomarkers, *B3GALNT1*, *SCRG1* and *ZNF423*, consistently showed increased expression in OA samples, suggesting their potential as reliable biomarkers of OA and possible therapeutic targets for OA disease. Considering all the results, the combination of *B3GALNT1*, *GRB10*, *KLF9*, and *SCRG1* biomarkers appeared to be a preferable diagnostic model. This conclusion was supported by both RNA-seq assay and qPCR of blood samples.

For the four identified biomarkers, we discovered that *GRB10* and *SCRG1* are reported to be correlated with synovial and cartilage inflammation. For instance, it has been discovered that human articular cartilage expresses *SCRG1* exclusively during cartilage degeneration. *SCRG1* overexpression has been exhibited to promote cartilage production in C3H10T1/2 cells, which are mesenchymal cells derived from mouse embryos, possessing the ability to differentiate into osteogenic, adipogenic, and chondrogenic lineages. Additionally, *SCRG1* overexpression has been found to suppress the proliferation of human mesenchymal stem cells (hMSCs). These results in combination with our findings imply that *SCRG1* could be very significant for cartilage growth [29]. On the other hand, *GRB10* has been identified to have an accurate diagnostic role in several spinal disorders, particularly in degenerative lumbar disc disease [30]. Additionally, both *SCRG1* and *GRB10* have been reported as potential therapeutic targets for human synovitis [13,31]. What's more, a previous bioinformatics study concluded that increased KLF9 expression was associated with the development of osteoarthritis, considering KLF9 as a signature gene of osteoarthritis [32]. B3GALNT1 is a gene that encodes an enzyme involved in glycosylation, which is the process of adding glycan molecules to proteins and lipids. The B3GALNT1 enzyme catalyzes the addition of N-acetylgalactosamine (GalNAc) to glycan chains, which is an important step in the synthesis of complex glycans. This glycosylation process can affect the structure and function of various glycoproteins and glycolipids. Altered B3GALNT1 expression or activity has been associated with cancer [33]. These results provide support for the significance of those biomarkers in the context of cartilage and synovial inflammation. One study released earlier this year discovered four macrophage polarization-related biomarkers, including *CSF1R*, *CEBPB*, *CX3CR1*, and *TLR7*, for the diagnosis of OA based on machine learning. The AUCs of these four were all above 0.980 while three achieved 1.000, suggesting an overfitting problem. In their validation dataset, the AUCs were still greater than 0.75 [34]. Data from real world patients were not available in their study, whereas in our study, samples from healthy people and OA patients were used, making our conclusions more reliable.

Although our findings support that *B3GALNT1*, *GRB10*, *KLF9* and *SCRG1* have high potential for clinical application, this work has its limitations. Firstly, the expression level of these biomarkers were not significantly different in our validation cohort, possibly due to the limited number of samples. We acknowledge the demographic limitations and sample size constraints of the datasets used, as well as the variability in disease stages. Furthermore, we note that the datasets mainly include samples from specific populations, which may affect the generalizability of our findings to other demographic groups. A larger sample is required to verify these findings more robustly. In addition, potential biases may be introduced during model training and validation, such as the effects of overfitting and inherent biases in the dataset. Some limitations do exist for each algorithm while there are some strengths though. RF's strength lies in handling high-dimensional data and being less sensitive to overfitting, while its limitation is its sensitivity to outliers. LASSO is proficient in variable selection but may struggle with high-dimensionality. SVM-RFE is great for feature selection but is computationally intense. Further research is required to investigate the distinct biological roles of these four biomarkers, and to establish the efficacy of the four-biomarker-combined diagnostic model in larger independent cohorts.

5. Conclusion

In this study, with the help of bioinformatics analyses and experimental validation, we successfully identified four diagnostic biomarkers of OA, namely *B3GALNT1*, *GRB10*, *KLF9* and *SCRG1*. Together, these biomarkers hold great potential as a diagnostic model for predicting osteoarthritis, providing novel insights for OA diagnosis and thus its treatment in clinical practice. Identification of these biomarkers promises a deeper understanding of this debilitating disease and with further research, much-improved therapies and clinical outcomes could be achieved.

Data availability statement

Data from public databases used in the study have been mentioned in the method section. The corresponding author may be contacted for more information.

Funding

This study was funded by the Beijing Municipal Health Commission (Grant No. BJRITO-RDP-2023) and the Beijing Natural Science Foundation - Haidian Original Innovation Joint Fund (Grant No. L222089).

CRediT authorship contribution statement

Xinyu Wang: Writing – original draft, Validation, Data curation, Conceptualization. **Tianyi Liu:** Writing – original draft, Visualization, Data curation, Conceptualization. **Yueyang Sheng:** Data curation. **Yanzhuo Zhang:** Data curation. **Cheng Qiu:** Writing – review & editing. **Manyu Li:** Data curation. **Yuxi Cheng:** Data curation. **Shan Li:** Data curation. **Ying Wang:** Data curation. **Chengai Wu:** Writing – review & editing, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Not applicable.

References

[1] D.J. Hunter, L. March, M. Chew, Osteoarthritis in 2020 and beyond - authors' reply, Lancet 397 (10279) (2021) 1060.

[2] A. Cieza, K. Causey, K. Kamenov, S.W. Hanson, S. Chatterji, T. Vos, Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019, Lancet 396 (10267) (2021) 2006–2017.

[3] Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet 396 (10258) (2020) 1204–1222.

[4] H. Long, Q. Liu, H. Yin, K. Wang, N. Diao, Y. Zhang, J. Lin, A. Guo, Prevalence trends of site-specific osteoarthritis from 1990 to 2019: findings from the global burden of disease study 2019, Arthritis Rheumatol. 74 (7) (2022) 1172–1183.

[5] J.N. Katz, K.R. Arant, R.F. Loeser, Diagnosis and treatment of hip and knee osteoarthritis: a review, JAMA 325 (6) (2021) 568–578.

[6] C.G. Boer, K. Hatzikotoulas, L. Southam, L. Stefánsdóttir, Y. Zhang, R. Coutinho de Almeida, T.T. Wu, J. Zheng, A. Hartley, M. Teder-Laving, et al., Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations, Cell 184 (18) (2021) 4784–4818.e4717.

[7] R.F. Loeser, J.A. Collins, B.O. Diekman, Ageing and the pathogenesis of osteoarthritis, Nat. Rev. Rheumatol. 12 (7) (2016) 412–420.

[8] E. Sanchez-Lopez, R. Coras, A. Torres, N.E. Lane, M. Guma, Synovial inflammation in osteoarthritis progression, Nat. Rev. Rheumatol. 18 (5) (2022) 258–275.

[9] M. Englund, Osteoarthritis, part of life or a curable disease? A bird's-eye view, J. Intern. Med. 293 (6) (2023) 681–693.

[10] S. Glyn-Jones, A.J. Palmer, R. Agricola, A.J. Price, T.L. Vincent, H. Weinans, A.J. Carr, Osteoarthritis. Lancet 386 (9991) (2015) 376–387.

[11] M.A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, et al., A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, Briefings Bioinf. 14 (6) (2013) 671–683.

[12] Y. Liang, F. Lin, Y. Huang, Identification of biomarkers associated with diagnosis of osteoarthritis patients based on bioinformatics and machine learning, J Immunol Res 2022 (2022) 5600190.

[13] Y.J. Deng, E.H. Ren, W.H. Yuan, G.Z. Zhang, Z.L. Wu, Q.Q. Xie, GRB10 and E2F3 as diagnostic markers of osteoarthritis and their correlation with immune infiltration, Diagnostics 10 (3) (2020).

[14] S. Xiao, L. Chen, The emerging landscape of nanotheranostic-based diagnosis and therapy for osteoarthritis, J. Contr. Release 328 (2020) 817–833.

[15] R. Edgar, M. Domrachev, A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (1) (2002) 207–210.

[16] S. Davis, P.S. Meltzer, GEOquery: a bridge between the gene expression Omnibus (GEO) and BioConductor, Bioinformatics 23 (14) (2007) 1846–1847.

[17] D. Woetzel, R. Huber, P. Kupfer, D. Pohlers, M. Pfaff, D. Driesch, T. Häupl, D. Koczan, P. Stiehl, R. Guthke, et al., Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation, Arthritis Res. Ther. 16 (2) (2014) R84.

[18] M.G. Broeren, M. de Vries, M.B. Bennink, P.L. van Lent, P.M. van der Kraan, M.I. Koenders, R.M. Thurlings, F.A. van de Loo, Functional tissue analysis reveals successful cryopreservation of human osteoarthritic synovium, PLoS One 11 (11) (2016) e0167076.

[19] K.M. Fisch, R. Gamini, O. Alvarez-Garcia, R. Akagi, M. Saito, Y. Muramatsu, T. Sasho, J.A. Koziol, A.I. Su, M.K. Lotz, Identification of transcription factors responsible for dysregulated networks in human osteoarthritis cartilage by global gene expression analysis, Osteoarthritis Cartilage 26 (11) (2018) 1531–1538.

[20] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.

[21] T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, et al., clusterProfiler 4.0: a universal enrichment tool for interpreting omics data, Innovation 2 (3) (2021) 100141.

[22] J. Tang, Y. Wang, Y. Luo, J. Fu, Y. Zhang, Y. Li, Z. Xiao, Y. Lou, Y. Qiu, F. Zhu, Computational advances of tumor marker selection and sample classification in cancer proteomics, Comput. Struct. Biotechnol. J. 18 (2020) 2012–2025.

[23] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Software 33 (1) (2010) 1–22.

[24] X. Wang, Y. Ning, P. Zhang, B. Poulet, R. Huang, Y. Gong, M. Hu, C. Li, R. Zhou, M.J. Lammi, et al., Comparison of the major cell populations among osteoarthritis, Kashin-Beck disease and healthy chondrocytes by single-cell RNA-seq analysis, Cell Death Dis. 12 (6) (2021) 551.

[25] R. Altman, E. Asch, D. Bloch, G. Bole, D. Borenstein, K. Brandt, W. Christy, T.D. Cooke, R. Greenwald, M. Hochberg, et al., Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association, Arthritis Rheum. 29 (8) (1986) 1039–1049.

[26] A. Jamshidi, J.P. Pelletier, J. Martel-Pelletier, Machine-learning-based patient-specific prediction models for knee osteoarthritis, Nat. Rev. Rheumatol. 15 (1) (2019) 49–60.

[27] J.S. Lawrence, J.M. Bremner, F. Bier, Osteo arthrosis, Prevalence in the population and relationship between symptoms and x-ray changes, Ann. Rheum. Dis. 25 (1) (1966) 1–24.

[28] P.A. Dieppe, Relationship between symptoms and structural change in osteoarthritis. what are the important targets for osteoarthritis therapy? J. Rheumatol. 70 (Suppl 2004) 50–53.

[29] K. Ochi, A. Derfoul, R.S. Tuan, A predominantly articular cartilage-associated gene, SCRG1, is induced by glucocorticoid and stimulates chondrogenesis in vitro, Osteoarthritis Cartilage 14 (1) (2006) 30–38.

[30] L. Wei, J. Guo, W. Zhai, Y. Xie, Y. Jia, CircRNA GRB10 is a novel biomarker for the accurate diagnosis of lumbar degenerative disc disease, Mol. Biotechnol. 65 (5) (2023) 816–821.

[31] G. Liu, G. He, J. Zhang, Z. Zhang, L. Wang, Identification of SCRG1 as a potential therapeutic target for human synovial inflammation, Front. Immunol. 13 (2022) 893301.

[32] J. Zhang, S. Zhang, Y. Zhou, Y. Qu, T. Hou, W. Ge, S. Zhang, KLF9 and EPYC acting as feature genes for osteoarthritis and their association with immune infiltration, J. Orthop. Surg. Res. 17 (1) (2022) 365.

[33] A. Jiang, X. Chen, H. Zheng, N. Liu, Q. Ding, Y. Li, C. Fan, X. Fu, X. Liang, T. Tian, et al., Lipid metabolism-related gene prognostic index (LMRGPI) reveals distinct prognosis and treatment patterns for patients with early-stage pulmonary adenocarcinoma, Int. J. Med. Sci. 19 (4) (2022) 711–728.

[34] P. Hu, B. Li, Z. Yin, P. Peng, J. Cao, W. Xie, L. Liu, F. Cao, B. Zhang, Multi-omics characterization of macrophage polarization-related features in osteoarthritis based on a machine learning computational framework, Heliyon 10 (9) (2024) e30335.