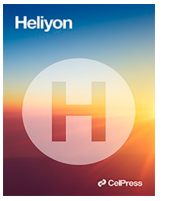
Heliyon 10 (2024) e35511

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Identification of diagnostic biomarkers of rheumatoid arthritis based on machine learning-assisted comprehensive bioinformatics and its correlation with immune cells

Kai-lang Mu [1], Fei Ran [1], Le-qiang Peng, Ling-li Zhou, Yu-tong Wu, Ming-hui Shao, Xiang-gui Chen, Chang-mao Guo, Qiu-mei Luo, Tian-jian Wang, Yu-chen Liu [*], Gang Liu [**]

Guizhou University of Traditional Chinese Medicine, Guiyang, 550025, Guizhou, China

ARTICLE INFO

Keywords:
Rheumatoid arthritis
Biomarkers
Diagnostic genes
Immune cells
Machine learning
Bioinformatics

ABSTRACT

Background: Rheumatoid arthritis (RA) is a chronic systemic autoimmune disease characterized by inflammatory cell infiltration, which can lead to chronic disability, joint destruction and loss of function. At present, the pathogenesis of RA is still unclear. The purpose of this study is to explore the potential biomarkers and immune molecular mechanisms of rheumatoid arthritis through machine learning-assisted bioinformatics analysis, in order to provide reference for the early diagnosis and treatment of RA disease.

Methods: RA gene chips were screened from the public gene GEO database, and batch correction of different groups of RA gene chips was performed using Strawberry Perl. DEGs were obtained using the limma package of R software, and functional enrichment analysis such as gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), disease ontology (DO), and gene set (GSEA) were performed. Three machine learning methods, least absolute shrinkage and selection operator regression (LASSO), support vector machine recursive feature elimination (SVM-RFE) and random forest tree (Random Forest), were used to identify potential biomarkers of RA. The validation group data set was used to verify and further confirm its expression and diagnostic value. In addition, CIBERSORT algorithm was used to evaluate the infiltration of immune cells in RA and control samples, and the correlation between confirmed RA diagnostic biomarkers and immune cells was analyzed.

Results: Through feature screening, 79 key DEGs were obtained, mainly involving virus response, Parkinson's pathway, dermatitis and cell junction components. A total of 29 hub genes were screened by LASSO regression, 34 hub genes were screened by SVM-RFE, and 39 hub genes were screened by Random Forest. Combined with the three algorithms, a total of 12 hub genes were obtained. Through the expression and diagnostic value verification in the validation group data set, 7 genes that can be used as diagnostic biomarkers for RA were preliminarily confirmed. At the same time, the correlation analysis of immune cells found that γδT cells, $CD4^{+}$ memory activated T cells, activated dendritic cells and other immune cells were positively correlated with multiple

* Corresponding author.
** Corresponding author.
E-mail addresses: lyc8564732@163.com (Y.-c. Liu), liugang888_2000@163.com (G. Liu).
[1] Joint first authors.

https://doi.org/10.1016/j.heliyon.2024.e35511
Received 11 January 2024; Received in revised form 29 July 2024; Accepted 30 July 2024
Available online 5 August 2024

RA diagnostic biomarkers, $CD4^+$ naive T cells, regulatory T cells and other immune cells were negatively correlated with multiple RA diagnostic biomarkers.
Conclusions: The results of novel characteristic gene analysis of RA showed that *KYNU*, *EVI2A*, *CD52*, *C1QB*, *BATF*, *AIM2* and *NDC80* had good diagnostic and clinical value for the diagnosis of RA, and were closely related to immune cells. Therefore, these seven DEGs may become new diagnostic markers and immunotherapy markers for RA.

1. Introduction

Rheumatoid arthritis is an autoimmune disease dominated by inflammatory arthritis [1]. It is characterized by multi-joint, symmetrical, and invasive arthritis of the small joints of the hand and foot. In the course of rheumatoid arthritis, the immune system attacks the joint tissue, often accompanied by extra-articular organ involvement and serum rheumatoid factor positive, which can eventually lead to chronic disability, joint destruction and loss of function [2–4]. According to the 2017 Global Burden of Disease (GBD) epidemiological report, the global prevalence of RA is 0.27 % [5]. The actual cause of RA is still unknown, which has a serious impact on individuals and society. However, early diagnosis of RA disease provides an opportunity for effective therapeutic response. At present, rheumatoid factor (RF) and anti-cyclic citrullinated peptide antibody (ACPA) are used as serum biomarkers for early diagnosis of RA [6]. However, these two serum markers are not positive in all early RA patients. Therefore, it is of great significance to find novel and feasible biomarkers for early diagnosis and effective treatment of RA [7,8].

Recent studies have shown that immune cells interfere with the information transmission between cells in the immune response of RA, and stimulate cells to move to inflammation, infection and trauma sites, participate in the whole process of RA occurrence, development and repair, and play a crucial role in the pathological changes caused by RA [9,10]. Among them, B cells secrete rheumatoid factors that recognize immunoglobulin Fe segments and form immune complexes, and can release chemokines and fix complements, so that inflammatory cells accumulate in the joints of patients [11]. Macrophages are abundant in cartilage tissue and synovium of RA patients. When activated, they can overexpress major histocompatibility complex (MHC) II molecules, chemokines and inflammatory factors, which are closely related to the degree of joint injury and clinical symptoms of patients [12]. The number of neutrophils increases significantly during RA activity and is always activated [13]. T cells bind to MHC II and antigen peptides, activate macrophages, release pro-inflammatory cytokines, activate synovial fibroblasts and chondrocytes, and secrete a variety of enzymes that degrade glycoproteins and collagen, thereby destroying tissues [14]. Th17 cells can induce synovial matrix and innate lymphocytes to secrete granulocyte-macrophage colony stimulating factor (MG-CSF), thereby triggering and enhancing RA [15]. Therefore, exploring the correlation between RA and immune cells from the perspective of the immune system is of great significance for elucidating the molecular system of RA and finding new diagnostic biomarkers and immunotherapy targets.

Based on the gene chip data in the GEO public database, this study used multiple sets of GEO data chips to obtain RA differentially expressed genes, and performed GO, KEGG, DO, and GSEA functional enrichment analysis. Three machine learning methods, such as LASSO regression, SVM-RFE and Random Forest, were used to identify potential biomarkers of RA, and their expression and diagnostic value were verified. Finally, the potential biomarkers of RA were analyzed for immune cell infiltration and immune cell correlation. This study will help to further understand the pathogenesis of RA and provide reference for its diagnosis, prevention and treatment as well as new biomarkers.

2. Materials and methods

2.1. Data acquisition and preprocessing

2.1.1. Data acquisition

Five datasets were obtained from the GEO database (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo), as detailed in Table 1. The four data sets of GSE1919, GSE93272, GSE10500 and GSE29746 were used as the training group, and the GSE55235 data set was used as the verification group for subsequent analysis.

2.1.2. Data processing

The gene expression matrix was extracted using the GEOquery package of R 4.3.1 software, and the gene probes and non-specific probes containing missing values were removed and annotated according to the official website information. DEGs were screened with

Table 1
Information on RA-related datasets obtained from the GEO database.

GEO data set name	Country	Platform file name	Number of RA samples	Number of control samples
GSE1919	Germany	GPL91	5	5
GSE93272	Japan	GPL570	232	43
GSE10500	America	GPL8300	5	3
GSE29746	Spain	GPL4133	9	11
GSE55235	Germany	GPL96	10	10

|log2FC|≥0.5, P < 0.05 as the screening criteria, and heat maps and volcano maps were drawn. The data with large values in the gene expression matrices of GSE1919, GSE93272, GSE10500 and GSE29746 were processed by log2, and the R 4.3.1 software 'limma' and 'sva' packages were used. After merging the data, the data set was corrected to eliminate the differences caused by batch effects.

2.2. Enrichment analysis

Gene ontology (GO) enrichment analysis was performed using R 4.3.1 software 'clusterProfiler', 'org.Hs.eg.db', 'enrichplot', 'ggplot2', 'stringi', 'limma' packages with p-value and q-value less than 0.05 as screening conditions. At the same time, based on the KEGG database (www.kegg.jp/kegg/kegg1.html) [16–18], the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed using the above R 4.3.1 software toolkit, and the results with significant P values were selected for visual analysis. With p-value and q-value less than 0.05 as the filtering conditions, disease ontology enrichment analysis (DO) was performed, and the top 30 diseases with q-value significance were selected for visual analysis. Gene set enrichment analysis (GSEA) was performed with p-adjust<0.05 as the filtering condition, and the enrichment of the top five biological functions or pathways in the normal group samples and RA disease group samples was selected for visual analysis.

2.3. Machine learning to identify potential biomarkers

The LASSO regression analysis was performed using the 'glmnet' package in R 4.3.1 software. The point with the smallest cross-validation error was used as the filtering condition. The number of genes corresponding to the point with the smallest cross-validation error was selected as the number of potential biomarkers identified by LASSO machine learning. The genes corresponding to the average ranking of LASSO regression analysis were used as RA potential biomarkers. SVM-RFE analysis was performed using R software 'e1071', 'kernlab', and 'caret' packages. The number of genes corresponding to the minimum cross-validation error in the analysis results was used as the number of potential biomarkers identified by SVM-RFE machine learning. The genes corresponding to the average ranking of SVM-RFE analysis were used as RA potential biomarkers. Random Forest analysis was performed using the 'randomForest' package of R software. The importance score of each differential gene was obtained at the minimum error of the cross-validation curve. Genes with an importance score greater than 1 were used as potential biomarkers for RA. The venn package was used to intersect the intersection genes obtained by LASSO, SVM-RFE, and Random Forest to obtain the final RA potential biomarkers obtained by the three machine learning methods.

2.4. Expression and diagnostic value verification of potential biomarkers for RA

The Extra Trees classifier model was used to screen out biomarkers with discriminative ability in the training test set and the verification test set. Meanwhile, The GSE55235 data set was used as the verification group to analyze the differences of RA potential biomarkers under '2.3′ and verify their expression differences in other RA sample data. Using the gene expression matrix of the validation group, the survival analysis of the RA potential markers with significant differences in expression was performed to verify its diagnostic value.

2.5. To obtain the correlation analysis of immune cell matrix and immune cell infiltration

Perl 5.32.1.1 software was used to organize the gene expression data into a gene matrix with row name as gene name and column name as sample name. The 'limma' package in R 4.3.1 language BioManager was used to correct the gene expression matrix of RA group and control group. The CIBERSORT method was used to perform deconvolution analysis on the expression matrix of human immune cell subtypes. The relative proportions of 22 immune cells were calculated, and a P value was obtained for each sample. The samples were screened according to $P < 0.05$ to obtain the immune cell composition matrix. The barplot function of R language 'graphics' package was used to draw the histogram of each immune cell composition ratio in the two groups of samples. The 'corrplot' package in R language was used to analyze the correlation of immune cells in the samples. The 'vioplot' package in R language was used to compare the proportion of immune cells in the samples of RA patients and control groups, and a violin map was drawn.

2.6. Correlation analysis of immune cells

The CIBERSORT algorithm was used to evaluate the immune cell infiltration of normal samples and RA samples in multiple sets of GEO data, and $P < 0.05$ was used as the screening standard. The immune cells evaluated include: B cells naive, B cells memory, Plasma cells, T cells CD8, T cells CD4 naive, T cells CD4 memory resting, T cells CD4 memory activated, T cells CD4 memory activated, T cells follicular helper, T cells regulatory, T cells gamma delta, NK resting, NK activated, Monocytes, Macrophages M0, Macrophages M1, Macrophages M2, Macrophages M2, Dendritic mast cells activated, Monocytes, Macrophages M0, Macrophages M1, Macrophages M2, Dendritic mast cells activated. Correlation analysis was used to study the correlation between RA potential biomarkers and immune infiltrating cells to explore the effect of RA potential biomarkers on immune microenvironment. Finally, the correlation between RA potential biomarkers and immune cells was analyzed.

3. Result

3.1. DEGs identification

Four RA gene chip data, including 251 RA and 62 normal samples, were used to obtain 79 DEGs by GEOquery package of R 4.3.1 software, including 77 up-regulated genes and 2 down-regulated genes. The results of DEGS identification were visualized by volcano map and heat map, as shown in Fig. 1A and. B. It can be seen from the figure that among the four genes with significant differences in RA gene chip data, up-regulated genes account for the majority, and DEGs have obvious expression differences in normal samples and RA samples.

3.2. Biological function analysis of DEGs

GO analysis obtained 227 BP items. It mainly involves the response to virus (GO:0009615), defense response to virus (GO:0051607), defense response to symbiont (GO:0140546), antimicrobial humoral response (GO:0019730), mucosal innate immune response (GO:0002227), antimicrobial peptide-mediated antimicrobial humoral immune response (GO:0061844), etc. MF entry 27. It mainly involves vitamin D-dependent calcium-binding protein (GO:0048306), RAGE receptor binding (GO:0050786), structural composition of ribosome (GO:0003735), toll-like receptor binding (GO:0035325), etc. CC item 40. It mainly involves cytochrome complex (GO:0070069) respiratory chain complex (GO:0098803), mitochondrial respiratory complex (GO:0005746), secretory granule lumen (GO:0034774), cytoplasmic vesicle lumen (GO:0060205), vesicle lumen (GO:0031983), etc (Fig. 2A).

DO enrichment analysis showed that RA differential genes were enriched in 84 diseases, including Dermatitis, Prostate cancer, Male reproductive organ cancer, Intracranial hypertension, Atopic dermatitis, etc (Fig. 2B). A total of 38 KEGG pathways were significantly enriched, including Parkinson disease, Prion disease, Huntington disease, Non-alcoholic fatty liver disease, Alzheimer disease, etc (Fig. 2D).

GSEA enrichment analysis is shown in Fig. 2C. Biological processes such as cell junction assembly, cell junction organization, response to epidermal growth factor and molecular functions such as actin binding and chromatin binding were significantly active in the control group. It is highly expressed in KEGG pathways such as Alzheimers disease, Huntington disease, Oxidative phosphorylation, Parkinson disease and Ribosome in RA samples.

3.3. Potential biomarkers of RA

Using LASSO regression analysis from DEGs expression matrix data, when the cross validation error is the smallest. LASSO regression analysis obtained 29 potential RA biomarkers: *EVI2A*, *COX7A2*, *CGRRF1*, *KYNU*, *RPL39*, *FCGBP*, *CD52*, *PDCD10*, *AIM2*, *ZNF267*, *BATF*, *SPAG1*, *PSMA4*, *TNFSF10*, *TSPAN2*, *NDC80*, *ENTPD1*, *S100A12*, *RPL34*, *LRRN3*, *C1QB*, *KLRB1*, *BMX*, *IFI6*, *CXCL10*, *TNFRSF17*, *CRISP3*, *IFI27* and *IFI44L* (Fig. 3A). SVM-RFE machine learning method is used to minimize the cross-validation error. *CKS2*, *PSMA2*, *CD58*, *S100A8*, *UQCRQ*, *C14orf2*, *CLEC2B*, *CAPZA2*, *CD52*, *FCGBP*, *EVI2A*, *IL15*, *CGRRF1*, *LY96*, *KYNU*, *COX7C*, *DBI*, *C1QB*, *CASP3*, *NDUFB3*, *SNRPG*, *COX7A2*, *ANXA3*, *GPR65*, *FAS*, *SUB1*, *IFI6*, *BATF*, *NDC80, RPL39*, *TMCO1*, *AIM2*, *SPAG1* and *LRRN3* can be used as potential biomarkers of RA obtained by SVM-RFE method (Fig. 3B). Random Forest machine learning method was used to obtain *CKS2*, *LY96*, *S100A8*, *COX7C*, *IL15*, *EVI2A*, *NDC80*, *LRRN3*, *C1QB*, *PSMA2*, *CD58*, *CASP3*, *CGRRF1*, *LMNB1*, *NAT1*, *AIM2*, *C14orf2*, *BMX*, *SPAG1*, *KLRB1*, *BATF*, *CD86*, *DBI*, *TSPAN2*, *UQCRQ*, *ENTPD1*, *COX7A2*, *FCGBP*, *TFEC*, *POLR2K*, *IFI27*, *ANXA3*, *BCL2A1*, *RNASE2*, *CD52*, *RSAD2*, *RPS7*, *HAT1* and *KYNU* 39 RA potential biomarkers obtained by Random Forest method (Fig. 3C). A total of 12 RA potential biomarkers such as *EVI2A*, *COX7A2*, *CGRRF1*, *KYNU*, *FCGBP*, *CD52*, *AIM2*, *BATF*, *SPAG1*, *NDC80*, *LRRN3*, and *C1QB* were obtained by intersecting the results obtained by the three machine learning methods (Fig. 3D).

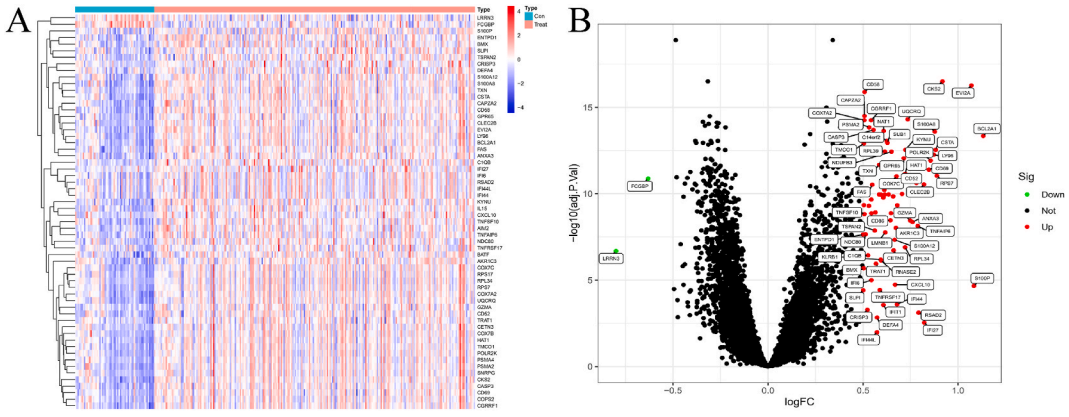


Fig. 1. Differential gene expression in RA. A: Heat map, B: Volcano map.

ID	Description
GO:0009615	Response to virus
GO:0051607	Defense response to virus
GO:0140546	Defense response to symbiont
GO:0019730	Antimicrobial humoral response
GO:0002227	Innate immune response in mucosa
GO:0061844	Antimicrobial humoral immune response mediated by antimicrobial peptide
GO:0048306	Calcium-dependent protein binding
GO:0050786	Rage receptor binding
GO:0003735	Structural constituent of ribosome
GO:0035325	Toll-like receptor binding
GO:0070069	Cytochrome complex
GO:0098803	Respiratory chain complex
GO:0005746	Mitochondrial respirasome
GO:0034774	Secretory granule lumen
GO:0060205	Cytoplasmic vesicle lumen
GO:0031983	Vesicle lumen

Fig. 2. Diagram of biological function analysis of differential genes. A represents GO enrichment analysis map, B represents DO enrichment analysis map, C represents GSEA enrichment analysis map, and D represents KEGG enrichment analysis map.

3.4. Diagnostic value of biomarkers for rheumatoid arthritis

The GSE55235 data set was used to verify the expression differences of RA potential biomarkers in control samples and RA samples (Fig. 4). $P < 0.05$ was considered to have significant differences. From Fig. 4, it can be seen that 8 RA biomarkers such as *KYNU*, *FCGBP*, *EVI2A*, *CD52*, *C1QB*, *BATF*, *AIM2* and *NDC80* have significant differences in the expression of control samples and RA samples in the validation dataset, while 4 RA biomarkers such as *COX7A2*, *CGRRF1*, *SPAG1* and *LRRN3* have no significant differences in the expression of control samples and RA samples in the validation dataset. The Extra Trees classifier model was used to construct multiple decision trees, and their prediction results were integrated. At the same time, the ROC analysis of RA markers was performed using the validation group data set to verify their diagnostic efficacy for RA (Fig. 5A and. B). AUC>0.800 was considered to have diagnostic ability. It can be seen from Fig. 5 that the ROC curve of the Extra Trees classifier model shows that the AUC of all RA feature genes is greater than 0.800, and the main diagonal of the confusion matrix shows higher prediction accuracy, indicating that this model has higher generalization ability and prediction accuracy(Fig. 5C). The ROC of single RA characteristic gene showed that the AUC of seven RA biomarkers, such as *KYNU*, *EVI2A*, *CD52*, *C1QB*, *BATF*, *AIM2* and *NDC80*, were all greater than 0.800, which could be preliminarily determined as the diagnostic biomarkers of RA. The R software 'dplyr', 'pROC', 'ggplot2', 'survival', 'regplot', 'rms', 'ggsci', 'survminer', 'timeROC', 'ggDCA' and 'limma' packages were used to identify seven RA diagnostic biomarkers in the training group data expression matrix using the Norman logistic regression model to analyze their weight coefficients. The nomogram model is further established (Fig. 6A). The R software 'pROC' package was used to verify the diagnostic nomogram model in the training group data set (Fig. 6B). When the AUC value in nomoscore was greater than 0.6, it could be considered that the established diagnostic nomogram model had good diagnostic value. It can be seen from Fig. 6B that the AUC value in nomoscore was 0.601, indicating that the established RA diagnostic nomogram model had good diagnostic value. The AUC values of each potential biomarker were all greater than 0.7, and the AUC value of *EVI2A* even reached 0.836, further indicating that the established RA diagnostic nomogram model had good diagnostic value.

3.5. Results of immune cell infiltration

The histogram of the proportion of immune cells showed the infiltration ratio of 22 immune cells in RA samples and control samples (Fig. 7A). Compared with the control group, the T cells gamma delta and Macrophages M0 infiltration levels in the RA group were higher. The infiltration level of Mast cells activated was low. The correlation heat map (Fig. 7B) showed that Neutrophils was positively correlated with Macrophages M0 ($r = 0.31$, $P < 0.05$), that is, when Neutrophils decreased, Macrophages M0 also decreased relatively; there was a negative correlation between T cells CD8 and Neutrophils ($r = -0.560$, $P < 0.05$). The difference of immune cell infiltration between RA group and control group was analyzed by violin diagram (Fig. 7C). We found that T cells CD4 naive ($P = 0.009$), T cells CD4 memory activated ($P < 0.001$), T cells regulatory ($P < 0.001$), T cells gamma delta ($P = 0.001$), Dendritic cells resting ($P < 0.001$), Mast cells activated ($P = 0.012$) and Neutrophils ($P = 0.043$) were significantly different between RA samples and control samples.

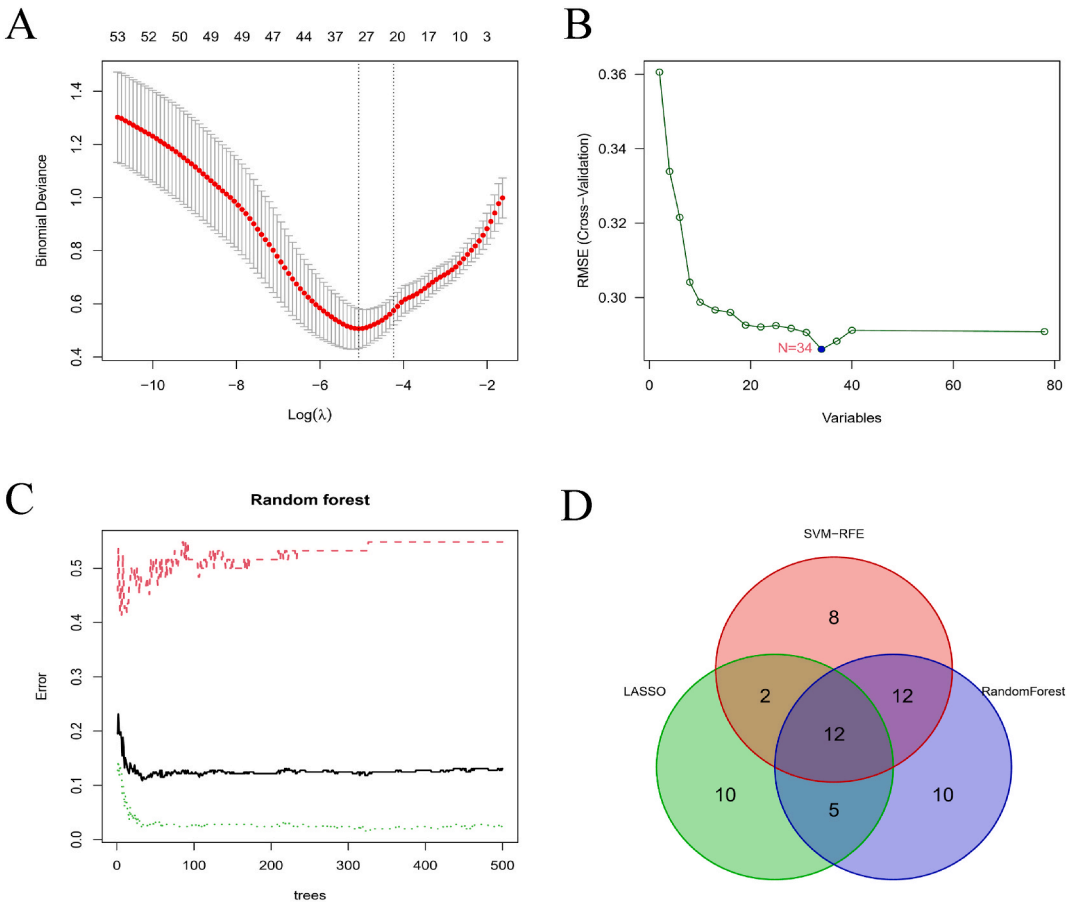


Fig. 3. Machine learning screens potential biomarker maps. a represents the LASSO regression analysis chart, showing 29 potential biomarkers when the cross-validation error is minimal. b represents the SVM-RFE analysis plot, and there are 34 potential biomarkers when the cross-validation error is the smallest. c represents a random forest tree map, which is screened with an importance score greater than 1, showing 39 potential biomarkers. d represents the intersection graph of veen machine learning results, showing that the three machine learning algorithms have 12 common potential biomarkers.

3.6. Correlation between potential biomarkers and RA infiltrated immune cells

The CIBERSORT algorithm evaluates the immune cell infiltration of normal samples and RA samples in multiple sets of GEO data. After calculating the score of each immune cell, the correlation between seven RA diagnostic biomarkers and immune cells was analyzed (Fig. 8A–G). Dendritic cells resting, T cells gamma delta, Neutrophils, T cells CD4 memory activated were positively correlated with AIM2, T cells CD4 naive, B cells naive, T cells regulatory, NK cells resting were negatively correlated with AIM2. T cells gamma delta, T cells CD4 memory activated, T cells CD8 were positively correlated with BATF, T cells CD4 naive, Neutrophils were negatively correlated with BATF. T cells CD4 memory activated and dendritic cells activated were positively correlated with C1QB, while T cells CD4 naive was negatively correlated with C1QB. T cells gamma delta, T cells CD4 memory activated, T cells CD8 were positively correlated with CD52, while Neutrophils, Macrophages M0, T cells regulatory were negatively correlated with CD52.

4. Discussion

In this study, four RA disease-related gene data chips were used to analyze 79 DEGs.These DEGs have significant expression differences in 62 control samples and 251 RA samples, which can be used for further study of RA diagnostic biomarkers. GO functional enrichment analysis showed that 79 DEGs mainly played a role in response to viruses, defense response to viruses and symbionts, antimicrobial humoral response, mucosal innate immune response, antimicrobial peptide-mediated antimicrobial humoral immune response, vitamin D-dependent calcium binding protein, RAGE receptor binding, ribosomal structure composition, toll-like receptor binding, cytochrome complex, respiratory chain complex, mitochondrial respiratory complex, secretory granule lumen, cytoplasmic vesicle cavity and vesicle cavity. KEGG pathway enrichment analysis showed that 79 DEGs were mainly involved in Parkinson's pathway, prion pathway, Huntington pathway, nonalcoholic fatty liver pathway and Alzheimer 's disease pathway. DO enrichment analysis showed that 79 DEGs were highly expressed in the occurrence and development of prostate cancer, dermatitis, male

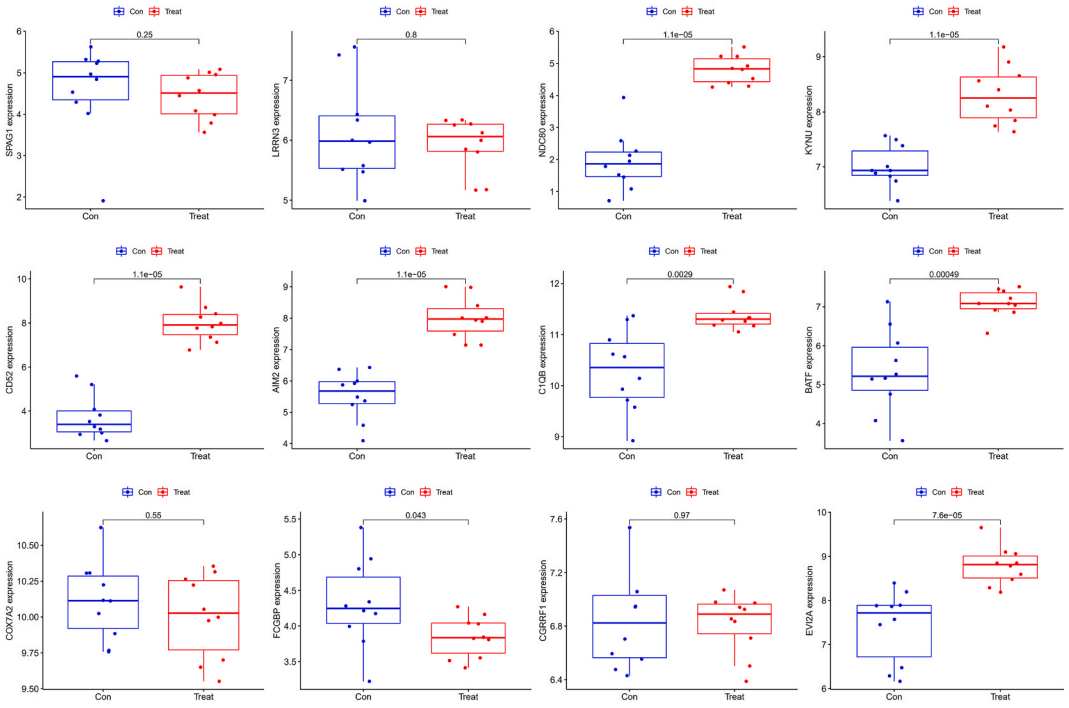


Fig. 4. Differential expression of potential biomarkers of RA in the validation dataset. When $P < 0.05$ was considered as statistically significant for the potential biomarker in the validation dataset.

reproductive organ cancer, intracranial hypertension and atopic dermatitis. GSEA enrichment analysis of 79 DEGs showed that these DEGs were significantly active in biological processes such as cell junction components, cell junction tissues, response to epidermal growth factor, and molecular functions such as actin binding and chromatin binding in the control samples. It is highly expressed in KEGG pathways such as Alzheimer's disease, Huntington 's disease, oxidative phosphorylation, Parkinson's disease and ribosome in RA samples.

The LASSO regression algorithm is a biased estimation method for processing multi-collinearity data. It is suitable for processing high-dimensional data or large amounts of data, and can scientifically and objectively screen characteristic indicators related to research purposes [19–22]. SVM-RFE is a sequential backward selection algorithm based on the maximum margin principle of SVM [23–25]. It trains samples through the model, then sorts the scores of each feature, removes the features with the smallest feature score, and then uses the remaining features to train the model again for the next iteration. Finally, the required number of features is selected, and the feature index can be obtained objectively according to the eigenvalue [26]. Random Forest is an algorithm that integrates multiple trees through the idea of ensemble learning, and can accurately screen out characteristic indicators with importance scores as thresholds [27]. At present, three machine learning algorithms, LASSO, SVM-RFE and Random Forest, are widely used in disease characteristic biomarkers and prognosis model construction, which are of great value in promoting the development of life science [28–31]. In this study, 29,34 and 39 characteristic genes were obtained by using LASSO regression analysis, SVM-RFE and Random Forest machine learning methods, respectively. The intersection of the three machine learning screening results was obtained, and a total of 12 RA potential biomarkers were obtained. These potential biomarkers have excellent performance in terms of sample difference, multicollinearity, sequence backward selection, and importance score, and can be further studied for RA diagnostic biomarkers.

The expression differences of 12 RA potential biomarkers in control samples and RA samples were analyzed by using the validation group data set ($P < 0.05$). The results showed that KYNU, FCGBP, EVI2A, CD52, C1QB, BATF, AIM2 and NDC80 had significant expression differences between the control group and the experimental group (AUC > 0.800), showing high sensitivity and low misjudgment rate. Preliminary analysis showed that KYNU, EVI2A, CD52, C1QB, BATF, AIM2 and NDC80 could be used as biomarkers with diagnostic value for RA. Modern pharmacological studies have shown that KYNU can be specifically highly expressed under the induction of inflammation in the body [32]. CD52 can activate NF-κB by inhibiting Toll-like receptors and trigger apoptosis in the process of inflammatory response [33]. C1QB can be used as a prognostic or predictive marker for neuropathic pain and can stimulate body pain in RA diseases [34]. BATF can regulate the destruction of osteoarthritis cartilage, and its expression is up-regulated in synovial tissue caused by CIA or K/BxN serum metastasis [35,36]. AIM2 can regulate the activation of caspase-1, promote the cleavage of caspase-1 and activate proinflammatory cytokines (such as IL-1β and IL-18), which is highly expressed in the process of inflammatory response [37]. EVI2 A and NDC80 have not been reported to be associated with RA disease, or can be used as potential new diagnostic biomarkers for RA.

By studying the correlation between seven RA diagnostic biomarkers and immune cells, it was found that γδ T cells, $CD4^+$ memory

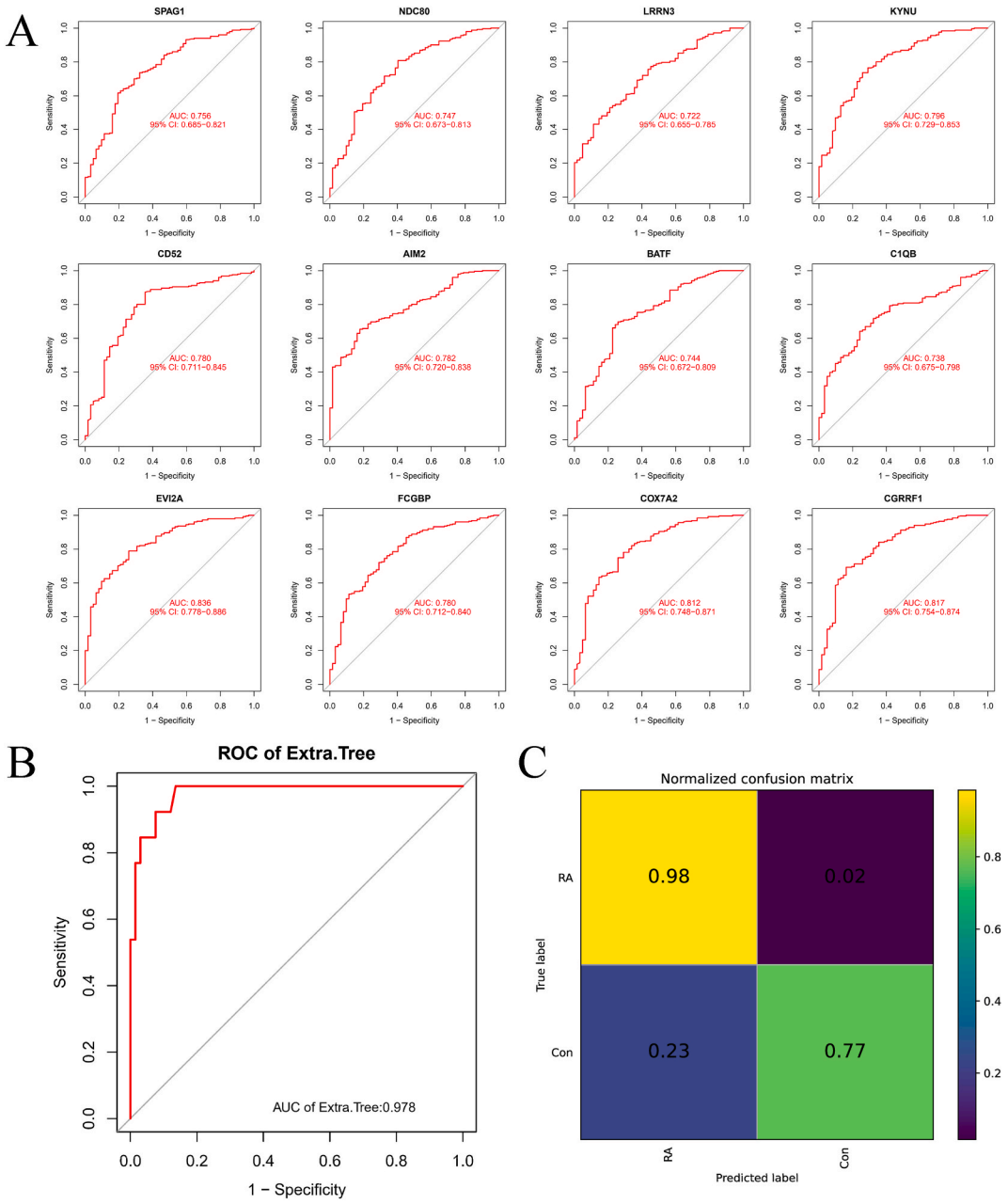


Fig. 5. Analysis of RA biomarkers in the validation group data set. A: ROC analysis of RA biomarkers with significant differences in the validation group data set; B: Common ROC analysis of RA biomarkers under the Extra Tree classifier model; C: Confusion matrix of training set and test set. AUC greater than 0.800 was considered to be a potential biomarker for RA with diagnostic ability.

activated T cells, activated dendritic cells and other immune cells were positively correlated with multiple RA diagnostic biomarkers, and $CD4^+$ naive T cells, regulatory T cells and other immune cells were negatively correlated with multiple RA diagnostic biomarkers. Studies have shown that the increase of γδ T cells can promote the production of IL-17 inflammatory factors, thereby accelerating the body's inflammatory response [38]. $CD4^+$ memory activated T cells are involved in the pathogenesis of RA autoimmune diseases, which can penetrate into the joints of RA patients and produce cytokines including tumor necrosis factor-α (TNF-α), leading to joint inflammation and bone destruction [39]; activation of dendritic cells can promote the development of arthritis [40]; $CD4^+$ naive T cells are easy to differentiate into regulatory T cells [41], and the depletion of regulatory T cells can lead to the occurrence of various autoimmune diseases, including arthritis. The supplement of regulatory T cells can alleviate the symptoms of arthritis [42].

RA is a chronic, autoimmune disease that mainly affects joints, but may also cause systemic symptoms and organ damage [1]. Its clinical significance mainly includes the following aspects: (1) Joint destruction and loss of function. RA can lead to joint

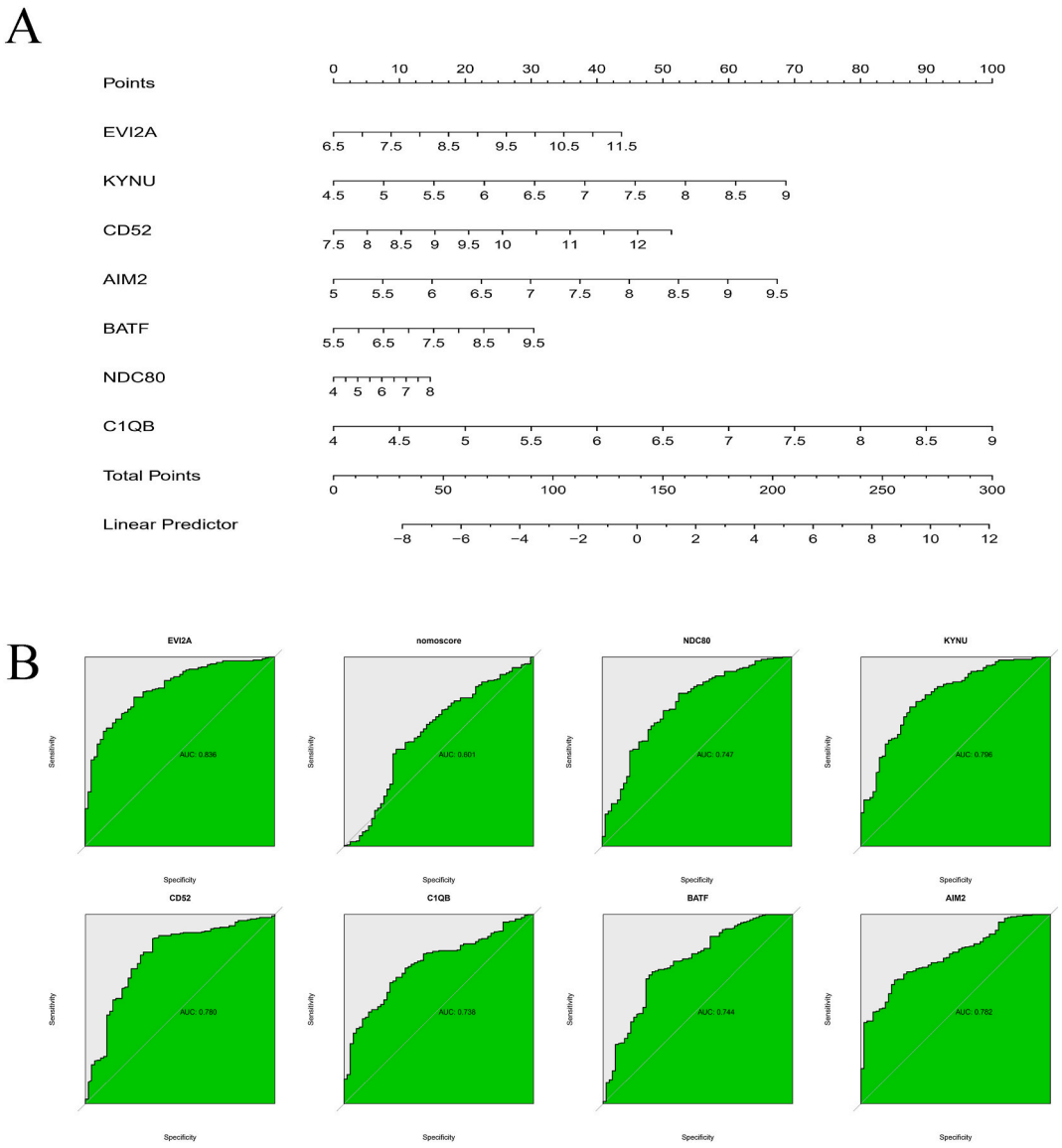


Fig. 6. Nomogram model for RA diagnosis. According to the sum of the test results of each index, the possibility of RA was judged. A: Nomogram model. B: Validation of ROC curve.

inflammation, articular cartilage and bone destruction, and ultimately lead to impaired joint function, or even loss of normal function. This will greatly affect the quality of life and daily activities of patients. (2) Systemic symptoms. In addition to joint symptoms, RA may also cause systemic symptoms, such as fatigue, anorexia, weight loss and anemia. These symptoms may affect the patient 's overall health. (3) May cause cardiovascular and other organ damage. There is a correlation between RA and cardiovascular disease, and patients have a higher risk of cardiovascular disease. In addition, RA may also cause inflammation and damage to organs such as eyes, skin, and lungs. (4) It may have an impact on psychosocial interaction. The nature and symptoms of chronic diseases may lead to depression, anxiety, and depression in patients, affecting their social activities and mental health. (5) Produce economic burden. The treatment and management of RA requires long-term drug therapy, physical therapy, rehabilitation and regular medical monitoring, which may bring economic burden, especially for patients who cannot get enough medical care [6,7]. Therefore, understanding and effective management of RA is essential to alleviate the suffering of patients and improve their quality of life. In this study, the clinical sample data of RA were used to analyze and explore the diagnostic biomarkers of RA and their correlation with immune cells. The results of this study are helpful for the early diagnosis and treatment of RA disease, which can help slow down the progression of disease, reduce the occurrence of complications and improve the overall health level of patients.

There are some limitations in this study. First, the study lacks clinical information, including changes in gene expression during disease progression, and is limited by the small sample size of the validated data set. Secondly, this study is only carried out from the perspective of gene transcriptome, and lacks multi-omics experiments. In addition, the results of this study are not completely

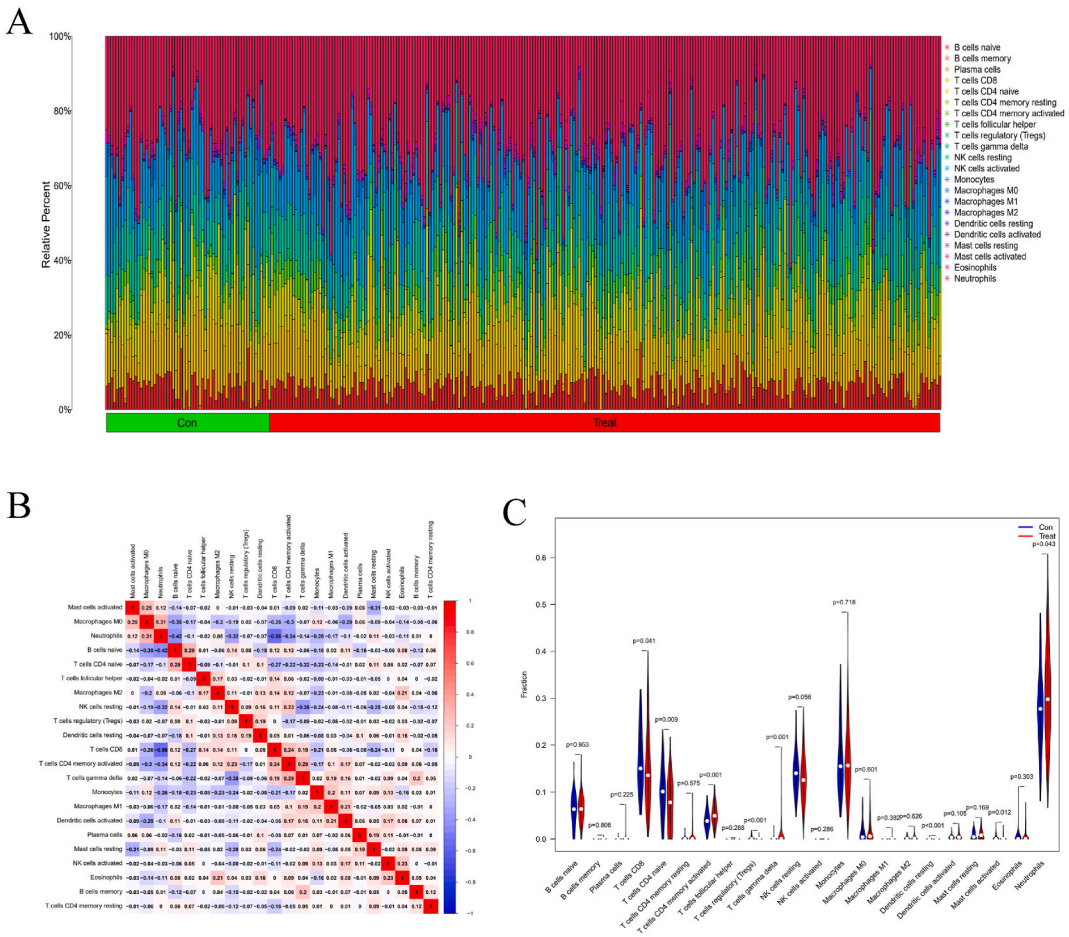


Fig. 7. Results of immune cell infiltration. A represents bars of 22 immune cells. B represents the heat map of immune cell correlation, with a reder color indicating a greater positive correlation and a bluer color indicating a greater negative correlation. C represents violin plot, showing the difference in the expression of 22 immune cells between RA samples and control samples, and $P < 0.05$ was considered as significant difference. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

consistent with the results of several similar studies [43–46]. Therefore, GEO chip data is crucial for further analysis in future studies to minimize the error range. At the same time, while providing a reference for the diagnosis, prevention and treatment of RA diseases and new biomarkers, it is still necessary to further verify the reliability of the results through large sample experiments.

5. Conclusion

In summary, this study combines RA gene expression profile chip data, bioinformatics, GO function, KEGG pathway, DO disease ontology, GSEA gene set analysis, machine learning (including LASSO, SVM-RFE, Random Forest), GEO verification and immune-related research methods are combined to reveal seven differentially expressed genes of RA (KYNU, EVI2 A, CD52, C1QB, BATF, AIM2 and NDC80), and their expression and diagnostic value are verified to identify new RA diagnostic biomarkers. It provides a new idea for the accurate diagnosis and immunotherapy of RA disease.

Data availability statement

All data in this paper can be obtained through Baidu network disk (https://pan.baidu.com/pcloud/home), copy link: (pwd = 8888)." title = "https://pan.baidu.com/s/1zZmkmTs223njhThXA7ANyg?(pwd = 8888).">https://pan.baidu.com/s/1zZmkmTs223njhThXA7ANyg?(pwd = 8888). At the same time, the data that support the findings of this study are available from the corresponding author, Y–C Liu, upon reasonable request.

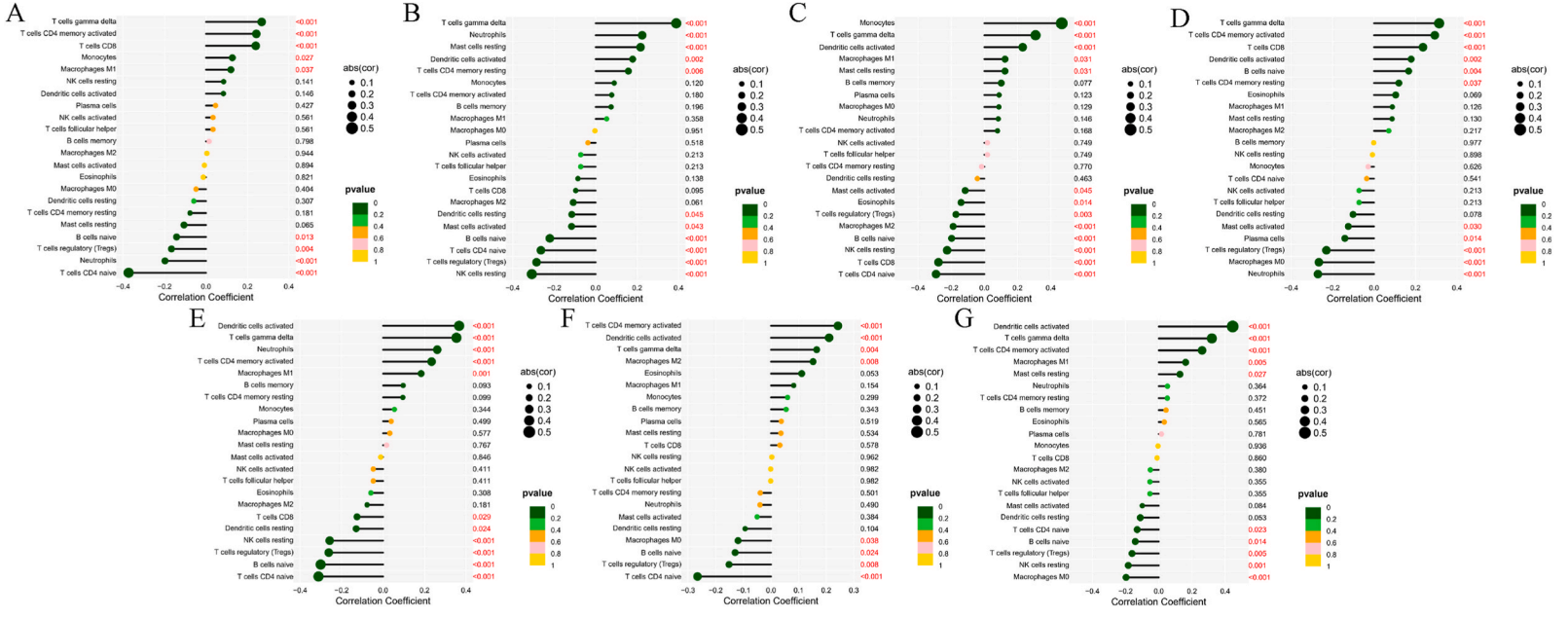


Fig. 8. Seven potential biomarkers were associated with RA infiltrating immune cells. (A: BATF. B: EVI2A. C: KYNU. D: CD52. E: AIM2. F: C1QB. G: NDC80.)

CRediT authorship contribution statement

Kai-lang Mu: Writing – review & editing, Writing – original draft, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Fei Ran:** Data curation, Conceptualization. **Le-qiang Peng:** Software, Data curation. **Ling-li Zhou:** Software, Conceptualization. **Yu-tong Wu:** Software, Data curation. **Ming-hui Shao:** Software, Resources. **Xiang-gui Chen:** Software, Resources, Data curation. **Chang-mao Guo:** Formal analysis, Data curation. **Qiu-mei Luo:** Software, Resources. **Tian-jian Wang:** Software, Conceptualization. **Yu-chen Liu:** Investigation, Funding acquisition. **Gang Liu:** Software, Formal analysis.

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.

Acknowledgements

Thanks to all the staff who contributed to the GEO databases. Thanks to the reviewers and editors for their sincere comments. This study was supported by the Science and Technology Top Talent Program of Guizhou Education Department (Qian teaching skill[2022] 083), Research Innovation and Exploration Project of Guizhou University of Traditional Chinese Medicine in 2021 (2019YFC1712500302).

References

[1] L. Grange, F. Alliot-Launois, [Rheumatoid arthritis], La Revue du praticien 72 (1) (2022) 68–70.
[2] C. Sunderhaus, Living with rheumatoid arthritis, Prof. Case Manag. 27 (1) (2022) 30–32.
[3] A.M. Ortiz, S. Ataman, Established rheumatoid arthritis. Best practice & research, Clin. Rheumatol. 33 (5) (2019) 101483.
[4] J.T. Giles, Extra-articular manifestations and comorbidity in rheumatoid arthritis: potential impact of pre-rheumatoid arthritis prevention, Clin. Therapeut. 41 (7) (2019) 1246–1255.
[5] A. Finckh, et al., Global epidemiology of rheumatoid arthritis, Nat. Rev. Rheumatol. 18 (10) (2022) 591–602.
[6] H. Allard-Chamard, G. Boire, Serologic diagnosis of rheumatoid arthritis, Clin. Lab. Med. 39 (4) (2019) 525–537.
[7] J.J. Cush, Rheumatoid arthritis: early diagnosis and treatment, Med. Clin. 105 (2) (2021) 355–365.
[8] V. Badot, [Early diagnosis of rheumatoid arthritis], Rev. Med. Brux. 35 (4) (2014) 215–222.
[9] G.S. Firestein, I.B. McInnes, Immunopathogenesis of rheumatoid arthritis, Immunity 46 (2) (2017) 183–196.
[10] F.M. Meier, M. Frerix, W. Hermann, U. Müller-Ladner, Current immunotherapy in rheumatoid arthritis, Immunotherapy 5 (9) (2013) 955–974.
[11] M. Volkov, K.A. van Schie, D. van der Woude, Autoantibodies and B Cells: the ABC of rheumatoid arthritis pathophysiology, Immunol. Rev. 294 (1) (2020) 148–163.
[12] E. Siouti, E. Andreakos, The many facets of macrophages in rheumatoid arthritis, Biochem. Pharmacol. 165 (2019) 152–169.
[13] C. Barranco, Rheumatoid arthritis: neutrophils play the right CARD, Nat. Rev. Rheumatol. 12 (6) (2016) 314–315.
[14] C.J. Malemud, Defective T-cell apoptosis and T-regulatory cell dysfunction in rheumatoid arthritis, Cells 7 (12) (2018) 223.
[15] L. Himer, et al., [Role of Th17 cells in rheumatoid arthritis], Orv. Hetil. 151 (25) (2010) 1003–1010.
[16] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[17] M. Kanehisa, Toward understanding the origin and evolution of cellular organisms, Protein Sci. 28 (11) (2019) 1947–1951.
[18] M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, M. Ishiguro-Watanabe, KEGG for taxonomy-based analysis of pathways and genomes, Nucleic Acids Res. 51 (D1) (2023) D587–D592.
[19] H.R. Frost, C.I. Amos, Gene set selection via LASSO penalized regression (SLPR), Nucleic Acids Res. 45 (12) (2017) e114.
[20] C. Cheng, Z.C. Hua, Lasso peptides: heterologous production and potential medical application, Front. Bioeng. Biotechnol. 28 (8) (2020) 5571165.
[21] Z. Li, M.J. Sillanpää, Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection, Theor. Appl. Genet. 125 (3) (2012) 419–435.
[22] S. Han, N. Wang, Y. Guo, F. Tang, L. Xu, Y. Ju, L. Shi, Application of sparse representation in bioinformatics, Front. Genet. 15 (12) (2021) 810875.
[23] H. Sanz, C. Valim, E. Vegas, J.M. Oller, F. Reverter, SVM-RFE: selection and visualization of the most relevant features through non-linear kernels, BMC Bioinf. 19 (1) (2018) 432.
[24] S. Sahran, D. Albashish, A. Abdullah, N.A. Shukor, Md Hayati, S. Pauzi, Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading, Artif. Intell. Med. 87 (2018) 78–90.
[25] P.A. Mundra, J.C. Rajapakse, SVM-RFE with MRMR filter for gene selection, IEEE Trans. NanoBioscience 9 (1) (2010) 31–37.
[26] M.L. Huang, Y.H. Hung, W.M. Lee, R.K. Li, B.R. Jiang, SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier, Sci. World J. 14 (2014) 795624.
[27] J.M. Nguyen, et al., Random forest of perfect trees: concept, performance, applications, and perspectives, Bioinformatics 37 (15) (2021) 2165–2174.
[28] M. Forsting, Machine learning will change medicine, Journal of nuclear medicine : official publication, Society of Nuclear Medicine 58 (3) (2017) 357–358.
[29] B. Van Calster, L. Wynants, Machine learning in medicine, N. Engl. J. Med. 380 (26) (2019) 2588.
[30] Ascent of machine learning in medicine, Nat. Mater. 18 (5) (2019) 407.
[31] R.C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.
[32] M. Huhn, et al., Inflammation-Induced mucosal KYNU expression identifies human ileal crohn's disease, J. Clin. Med. 9 (5) (2020) 1360.
[33] M. Rashidi, et al., CD52 inhibits Toll-like receptor activation of NF-κB and triggers apoptosis to suppress inflammation, Cell Death Differ. 25 (2) (2018) 392–405.
[34] J.A. Yang, J.M. He, J.M. Lu, L.J. Jie, Jun, Gal, Cd74, and C1qb as potential indicator for neuropathic pain, J. Cell. Biochem. 119 (6) (2018) 4792–4798.
[35] S.H. Park, et al., BATF regulates collagen-induced arthritis by regulating T helper cell differentiation, Arthritis Res. Ther. 20 (1) (2018) 161.
[36] J. Rhee, et al., Inhibition of BATF/JUN transcriptional activity protects against osteoarthritic cartilage destruction, Ann. Rheum. Dis. 76 (2) (2017) 427–434.
[37] P. Kumari, A.J. Russo, S. Shivcharan, V.A. Rathinam, AIM2 in health and disease: inflammasome and beyond, Immunol. Rev. 297 (1) (2020) 83–95.
[38] P. Gaur, R. Misra, A. Aggarwal, Natural killer cell and gamma delta T cell alterations in enthesitis related arthritis category of juvenile idiopathic arthritis, Clin. Immunol. 161 (2) (2015) 163–169.
[39] R. Giwa, J.R. Brestoff, Mitochondria transfer to CD4(+) T cells may alleviate rheumatoid arthritis by suppressing pro-inflammatory cytokine production, Immunometabolism 4 (2) (2022) e220009.
[40] R. Yabe, et al., TARM1 contributes to development of arthritis by activating dendritic cells through recognition of collagens, Nat. Commun. 12 (1) (2021) 94.
[41] B. Martin, et al., Highly self-reactive naive CD4 T cells are prone to differentiate into regulatory T cells, Nat. Commun. 4 (2013) 2209.
[42] N. Komatsu, H. Takayanagi, Regulatory T cells in arthritis, Progress in molecular biology and translational science 136 (2015) 207–215.

[43] H. Chen, et al., Identification of diagnostic biomarkers, immune infiltration characteristics, and potential compounds in rheumatoid arthritis, BioMed Res. Int. 2022 (2022) 1926661.
[44] S. Zhou, H. Lu, M. Xiong, Identifying immune cell infiltration and effective diagnostic biomarkers in rheumatoid arthritis by bioinformatics analysis, Front. Immunol. 13 (12) (2021) 726747.
[45] R. Yu, et al., Identification of diagnostic signatures and immune cell infiltration characteristics in rheumatoid arthritis by integrating bioinformatic analysis and machine-learning strategies, Front. Immunol. 6 (12) (2021) 724934.
[46] H. Hu, et al., Identification and validation of ATF3 serving as a potential biomarker and correlating with pharmacotherapy response and immune infiltration characteristics in rheumatoid arthritis, Front. Mol. Biosci. 13 (8) (2021) 761841.