KLF9 and EPYC function as characteristic genes of osteoarthritis and are associated with immune infiltration

Jiayin Zhang
The second hospital of Jilin University
Shengjie Zhang
The second hospital of Jilin University
Yu Zhou
Changchun University of Chinese Medicine
Yuan Qu
The second hospital of Jilin University
Tingting Hou
The second hospital of Jilin University
Wanbao Ge
The second hospital of Jilin University
Shanyong Zhang (jzhangshanyong@yeah.net)
The second hospital of Jilin University

Research Article

Keywords: Osteoarthritis, KLF9, EPYC, Machine Learning, Bioinformatics, Immune Infiltration

Posted Date: May 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1653926/v1

Abstract

Background: Osteoarthritis is a common degenerative disease of articular cartilage. Its typical features include articular cartilage degeneration, subchondral bone structural changes and osteophyte formation. The main clinical manifestations are increasingly severe knee joint swelling, stiffness, deformity and limited mobility. With the advent of the era of big data, the processing of massive data has become a hot topic, and the continuous development and improvement of machine learning algorithms have laid the foundation for the era of big data. This paper uses machine learning methods to identify the real key characteristic genes of osteoarthritis and explore the relationship between them and immune infiltration, so as to reveal the pathogenic mechanism of osteoarthritis at the molecular biology level, aiming to provide osteoarthritis diagnosis and treatment.

Method: Download the GSE55235 and GSE55457 datasets from the GEO database, merge the two as the training set, and download the GSE98918 data as the validation set. Gene differential expression analysis was performed on the training set, lasso regression model and support vector machine model were constructed by machine learning algorithm, and the intersection genes were taken as feature genes and receiver operating characteristic curves were drawn. Finally, the expression profile of osteoarthritis was analyzed by immune cell infiltration and the correlation between the co-expression of characteristic genes and immune cells was analyzed.

Conclusion: EPYC and KLF9 can be used as the characteristic genes of osteoarthritis.The silencing of EPYC and the overexpression of KLF9 are related to the occurrence of osteoarthritis and the infiltration of immune cells.

1. Introduction

Osteoarthritis (OA) is a common chronic joint disease in middle-aged and elderly people, characterized by degenerative changes of articular cartilage, destruction of rings and progressive bone hyperplasia. Genetic factors have long been confirmed to be related to the occurrence of OA. Currently, approximately 250 million people worldwide are affected by it. OA often affects multiple joints throughout the body, of which the knee joint is the most common site, followed by the wrist and hip joints [1]. The etiology of OA has not yet been elucidated, but its risk factors are numerous, including genetics, gender, joint damage, age, and obesity [2]. Joint injuries are becoming more common as the global population ages and obesity increases. Some scholars believe that the mechanical damage of the joints is dominant in the occurrence and progression of OA [3], while other scholars believe that genetic factors play an important role in OA [4]. The main clinical manifestations of OA are joint pain, swelling, stiffness, dysfunction and gradually aggravating, severe cases can lead to disability. Diagnosis is also based on imaging changes. First-line therapies such as non-steroidal anti-inflammatory drugs, drugs such as acetaminophen and glucocorticoids, and even joint replacement surgery are mainly focused on the relief of symptoms such as pain and limited mobility. However, there is currently no treatment for OA [5–7]. Therefore, the search for genetic biomarkers of OA is of great significance for the diagnosis and treatment of the disease.

Machine Learning (ML) is a multi-domain interdisciplinary subject. As the core of artificial intelligence and data science, machine learning mainly studies how computers simulate or realize human learning behaviors to acquire new knowledge or skills, and to reorganize existing knowledge structures to continuously improve their performance. With the advent of the era of big data, machine learning has been widely used in various fields in biomedicine, such as genomics, proteomics, microarrays, systems biology, evolution, and text mining[8–10]. Tibshirani[11] proposed the Lasso (The Least Absolute Shrinkage and Selectionator Operator) algorithm in 1996. The algorithm obtains a refined model by constructing a penalty function. The basic idea is to minimize the residual sum of squares under the constraint that the sum of the absolute valuesof the regression coefficients is less than a constant, so as to generate some regression coefficients strictly equal to 0, and obtain a model with strong explanatory power. Support vector machine (SVM) is widely used in pattern recognition, machine learning and other fields. Support vector machine recursive feature elimination (SVM-RFE) is a sequence backward selection algorithm based on the maximum interval principle of SVM. It passes the model training samples, sorts the score of each feature, removes the feature with the smallest feature score, then retrains the model with the remaining features, performs the next iteration, and finally selects the required number of features[12–13]. SVM-RFE can better screen out the characteristic genes of the disease for the diagnosis and treatment of the disease. The disease signature genes screened by the Lasso regression model and the SVM-REF model will be more reliable.

In recent years, immune infiltration has become more widely used in bioinformatics analysis, and there is evidence that chondrocytes in osteoarthritis patients release specific antigens to trigger the activation of immune responses. There are many immune cells involved in OA, including innate immunity and acquired immunity, which makes anti-cytokine less effective in OA[14]. Therefore, to clarify the infiltration of immune cells in the synovium of OA patients and the genes related to their regulation appear particularly important.

Based on the algorithm of machine learning in bioinformatics, this study screened the characteristic genes related to the pathogenesis of OA through the R-project, and established the relationship between them and immune cell infiltration, aiming to reveal the complex pathogenesis of OA and Provide a reference for the development of more novel markers for the diagnosis of OA.

2. Materials And Methods

2.1 Data download and integration

We downloaded three datasets GSE55235(10 normal synovial tissues and 10 OA synovial tissues), GSE55457(10 normal synovial tissues and 10 OA synovial tissues) and GSE98918(12 normal synovial tissues and 12 OA synovial tissues) from the GEO database respectively[15–16], and the first two datasets were integrated and batch corrected through the ‘sva’ package[17] in R-project, and the integrated results

were used as training sets (20 OA synovial tissue samples and 20 normal synovial tissue samples), while GSE98918 was used as a validation set.

2.2 Screening of Differentially Expressed Genes (DEGs).

Read the combined training set data through R language, and divide it into OA group and normal control group, extract gene expression in the two groups, and set the filtering threshold: |Log2FC|>1.5, adjust.Pvalue < 0.05, and both are considered to be significantly different. Load the 'limma' package[18], perform difference analysis according to the filter conditions, output the difference analysis results, and load the 'pheatmap' package[19] and the 'ggplot2' package[20] to draw the heatmap and the volcano map respectively.

2.3 Enrichment analysis of genes

We performed GO, KEGG, DO enrichment analysis on the differential genes, respectively. Set the filter conditions for enrichment analysis: pValue < 0.05, qValue < 0.05, if both are satisfied, the enrichment result is considered meaningful. The 'org.Hs.eg.db' package[21] is loaded for gene ID conversion, and the 'clusterProfiler' package[22–23] is used for enrichment analysis. After saving the analysis results, the 'enrichplot'[24] and ggplot2 packages are used to visualize the enrichment results and draw bubble charts and histograms. Finally, we performed GSEA enrichment analysis on all genes, saved the top five enriched pathways and made graphs to visualize them.

2.4 Screening of signature genes

We load the 'glmnet' package[25] to build the Lasso regression model using differential genes, load the 'e1071' package[26], the 'caret' package[27] and the 'kernlab' package[28] to build the SVM-RFE model, save and output the genes screened by the two, take the intersection genes as the feature genes and use the 'venn' package[29] to draw the Wei Entu for visualization.

2.5 Drawing of receiver operating characteristic(ROC) curve

We use the 'pROC' package[30] to draw the ROC curve of the eigengenes in the training set. AUC > 0.9 means that the gene has a high accuracy in diagnosing diseases.

2.6 Verification of Characteristic Genes

We used the 'limma' package to analyze the expression of the selected characteristic genes in the validation set. P < 0.05 considered that there was a difference in gene expression between the disease group and the normal group. At the same time, draw eigengenes in the validation set to get the ROC curve and compare it with the results obtained in the training set.

2.7 Analysis of immune cell infiltration

We used CIBERSORT algorithm to calculate the abundance of various immune cell types in the samples[31]. The CIBERSORT source code was first created, and the expression levels of 22 kinds of

marker genes of immune cells were prepared. The immune cell infiltration analysis was performed on the data of the training set. PValue < 0.05 was set as the filtering condition to filter the immune infiltration results, and the results were saved. Bar diagrams drawing visualized the content of immune cells in each sample, and “corrplot” package[32] was used to draw the correlation heat map. Violin plots were drawn using “vioplot” package[33] to show differences in immune cells between OA group and normal group.

2.8 Correlation analysis of characteristic genes and immune cells

We use the ‘reshape2’ package[34] to organize the gene expression data, obtain the expression levels of the characteristic genes, cycle the immune cells, set p < 0.05 as the correlation filter condition, and use the ‘ggpubr’ and ‘ggExtra’ packages[35–36] to draw correlation scatterplots and correlation bars Lollipop chart for visualization of correlation results.

3. Result

3.1 Screening of differentially expressed genes

Through the screening of differential genes, we found that a total of 122 genes (48 up-regulated genes and 74 down-regulated genes) were differentially expressed between the OA group and the normal group and the difference fold was more than 3 times (Fig. 1(a), (b)).

3.2 Results of Gene Enrichment

3.2.1 Differential gene enrichment analysis results

The Gene oncology(GO) enrichment results suggested that the biological processes that the differential genes were mainly involved in included the response to lipopolysaccharide, the response to steroid hormones, and the response to bacteria-derived molecules. The cellular components(CC) where the functions of differential gene products are located mainly include extracellular matrix, membrane rafts and plasma membrane microdomains. The molecular functions(MF) of differential gene products include receptor ligand activity, signaling receptor activator activity, and cytokine activity (Fig. 2a). The KEGG pathway enrichment analysis indicated that the differential genes were mainly involved in the interleukin 17 (IL-17) signaling pathway, followed by Kaposi sarcoma-associated herpesvirus infection, rheumatoid arthritis and tumor necrosis factor (TNF) signaling pathways (Fig. 2b). Disease ontology(DO)enrichment analysis revealed that the differential genes were mainly enriched in benign cellular tumors, preeclampsia, lymphocytic leukemia and OA (Fig. 2c).

3.2.2 Analysis Results of Gene Set Enrichment

The results of GSEA-GO analysis showed that the normal synovial tissue gene set has DNA-binding transcription activator activity, and the products function at the nuclear plaques, mainly involved in the regulation of RNA splicing (Fig. 3a). The main product of the OA synovial tissue gene set functions at the nucleolar plaques and participates in the antigen-antibody binding reaction (Fig. 3b).

The results of GSEA-KEGG analysis showed that the normal synovial tissue gene set was mainly involved in adipocytokine signaling pathway, MAPK signaling pathway and NOD-like receptor signaling pathway (Fig. 3c). While the OA synovial tissue gene set is mainly involved in signaling pathways such as allograft rejection, lysosomal pathway and phosphorylation (Fig. 3d).

3.3 Screening of Characteristic Genes

By constructing a lasso regression model on the training set, we screened out 14 eigengenes with diagnostic significance (KLF9, APOLD1, TIPARP, EPYC, JUN, PPP1R15A, FKBP5, RND1, CCZ1B, ZIC1, MGC12488, TAC1, WIF1, ERAP2) (Fig. 4a). Further constructing the SVM-RFE model, we screened out 2 characteristic genes (KLF9, EPYC) with diagnostic significance (Fig. 4b). Taking the intersection of the two regression models (Fig. 4c), so far, we believe that KLF9 and EPYC can be used as characteristic genes of OA.

3.4 ROC Curve Drawing

By plotting the ROC curve of KLF9 and EPYC genes, we know that KLF9 (AUC = 0.992, CI = 0.97-1.00) and EPYC (AUC = 0.990, CI = 0.96-1.00) are sensitive to diagnosis OA (Fig. 5).

3.5 The result of validating the model

In the validation set, the expression levels of KLF9 and EPYC were differentially analyzed, and the results showed that KLF9 was lowly expressed in OA synovial tissue, and EPYC was highly expressed in OA synovial tissue, and the difference in expression between the two was statistically significant ($P < 0.05$) ($P < 0.05$).(Fig. 6). At the same time, by drawing the ROC curves of the two genes in the validation set, we found that KLF9 and EPYC also had high sensitivity (AUC > 0.9) in diagnosing OA (Fig. 7), which was consistent with the results of the training set.

3.6 The results of Immune Cell Infiltration

The expression of all immune cells in each sample has been represented by histograms (Fig. 8a). The results of correlation analysis between immune cells suggested that resting mast cells and regulatory T cells, plasma cells and memory B cells, T cells γδ and activated CD4 + memory T cells, naive CD4 T cells and activated CD4 + memory T cells, resting Resting NK cells and naive CD4 + T cells, eosinophils and activated NK cells, resting memory CD4 + T cells and activated NK cells were positively correlated (correlation coefficient R > 0.5), naive B cells And memory B cells, the expression of activated mast cells and resting mast cells was negatively correlated (correlation coefficient R<-0.5), and the correlation

between other immune cells has been shown by correlation heat map (Fig. 8b). A total of 5 kinds of immune cells have different expressions in OA and normal human synovial tissue. Resting state memory CD4 + T cells, activated NK cells, and activated mast cells are lowly expressed in OA synovial tissue ($p < 0.01$)., while regulatory T cells, resting-state mast cells were highly expressed in OA synovial tissue (Fig. 8c). The co-expression correlation analysis of characteristic genes and immune cells indicated that KLF9 was associated with resting memory CD4 + T cells ($R = 0.67$, $p < 0.01$), activated mast cells ($R = 0.67$, $p < 0.01$) and activated The expression of NK cells ($R = 0.39$, $p = 0.012$) was positively correlated with CD8 + T cells ($R=-0.32$, $p = 0.041$), plasma cells ($R=-0.38$, $p = 0.016$), resting mast cells ($R=-0.51$, $p < 0.01$) and the expression of regulatory T cells ($R=-0.56$, $p < 0.01$) were negatively correlated (Fig. 9)(Fig. 11(a)). EPYC versus resting mast cells ($R = 0.66$, $p < 0.01$), plasma cells ($R = 0.45$, $p < 0.01$), memory B cells ($R = 0.45$, $p = 0.01$) and regulatory T cells ($R = 0.37$, $p = 0.019$) was positively correlated with activated NK cells ($R=-0.46$, $p < 0.01$), resting CD4 + memory T cells ($R=-0.53$, $p < 0.01$) and activated mast cells ($R=-0.53$, $p < 0.01$) $=-0.57$, $p < 0.01$) was negatively correlated (Fig. 10) (Fig. 11(b)).

4. Conclusion

Through the analysis of the gene expression chip of OA, we found that the silencing of KLF9 and the overexpression of EPYC are related to the occurrence of OA, and we can know that the two can be used as diagnostic features of OA through machine learning algorithm Gene. Resting memory CD4 + T cells, activated NK cells and activated mast cells were suppressed in OA synovial tissue while regulatory T cells were overexpressed in OA synovial tissue. KLF9 was positively correlated with the expression of resting memory CD4 + T cells, activated NK cells and activated mast cells, and negatively correlated with the expression of CD8 + T cells, plasma cells, resting mast cells and regulatory T cells. EPYC was positively correlated with the expression of plasma cells, resting-state mast cells and regulatory T cells, and negatively correlated with the expression of resting-state memory CD4 + T cells and activated mast cells.

5. Discussion

OA is one of the most common joint diseases in the elderly, with a high public health burden and no cure. Articular cartilage has the function of buffering and reducing friction, and it is also the most severely degenerated part of OA. Therefore, rebuilding the integrity of articular cartilage is expected to replace joint replacement as a new method for radical OA.

Oxidative stress and reactive oxygen species (ROS) have been shown to be strongly associated with the occurrence of OA. When chondrocytes, synoviocytes, and osteoblasts are continuously subjected to external mechanical stress, they can produce excessive pro-inflammatory mediators to break down the pro-inflammatory mediators. Oxidative/antioxidant balance, which in turn degrades the extracellular matrix[37]. As a member of the KLFs family, Kruppel-like factor 9 (KLF9) plays an important role in the oxidative stress response. Studies have shown that Nrf2 can stimulate the expression of KLF9, and KLF9 inhibits the expression of several important antioxidant enzymes such as thioredoxin reductase 2, resulting in a further increase in Klf9-dependent ROS and ultimately the degradation of cartilage[38].

Through the KEGG pathway enrichment analysis, we learned that the differentially expressed genes in OA are mainly involved in the IL-17 signaling pathway, and IL-17 can also promote the process of oxidative stress. Whether there is a potential connection between the two needs further research.

Epiphycan (EPYC) is a protein-coding gene. It is a member of the leucine-rich small repeat proteoglycan (SLRP) family. This gene consists of seven exons and regulates fibrillation by interacting with collagen fibrils and other extracellular matrix proteins. EPYC is involved in cartilage formation in normal synovial tissue. EPYC knockout mice develop osteoarthritis with age[39–40]. In the present study, EPYC is overexpressed in OA, and we speculate that this is most likely because the destruction of articular cartilage causes chondrocytes to increase EPYC production in an attempt to repair the damaged extracellular matrix (ECM). Since EPYC belongs to the SLRPS family, the effects of the SLRP family on cartilage and the mechanisms involved in the occurrence of OA are numerous and complex, including changes in the extracellular collagen network and TGF-b signaling pathway. Therefore, the mechanism of EPYC regulation in OA needs to be further elucidated. In addition, studies have confirmed that NSAIDS drugs can reduce the expression of EPYC in prostate cancer cells[41]. As the first-line treatment of OA, NSAIDS drugs need to be further explored.

The CIBERSORT score is widely used in gene expression profiling to quantify immune cell scores with high accuracy. The infiltration of immune cells in OA synovial tissue has become the consensus of many scholars. Among them, CD4 + T cells, mast cells, and macrophages play an important role in synovial inflammation. IgE-dependent mast cell activation and the pathogenic role of mast cell-mediated tryptase in osteoarthritis have been demonstrated, but mast cells themselves are not differentially expressed in OA synovial tissue[42]. In this study, the immune infiltration analysis showed that resting mast cells were highly expressed in OA synovial tissue, while activated mast cells were lowly expressed. We speculate that this is probably because mast cells are not directly involved in the pathogenic process of OA, but mediate Other proteases or histamine indirectly lead to the occurrence of OA, but the specific mechanism of its role in OA needs to be further elucidated. Regulatory T cells (Tregs) play an important immunomodulatory role in many inflammatory and autoimmune diseases, but they are more inhibiting osteoclasts and helper T cells to protect local articular cartilage from destruction[43–45]. Our experimental results suggest that regulatory T cells infiltrate the OA synovium, which is likely related to synovial tissue destruction leading to reactive proliferation of regulatory T cells to suppress local inflammatory responses. Of course, this requires further verification by experiments.

It is not the first time that machine learning algorithms have been used in gene screening for OA. In past experiments, we believed that the threshold of difference was too low, which resulted in too many differential genes after filtering, which affected the accuracy of enrichment analysis and machine learning algorithms. On this basis, we increased the threshold of the difference analysis to 3 times, that is, set the LogFC to 1.5. We believe that the results of this analysis are more accurate and meaningful.

In short, we processed the chip expression data by computer, screened the characteristic diagnostic genes of OA by using machine learning algorithm, and explored the relationship between them and immune

cells, in order to provide reference direction for the early diagnosis and treatment of OA.

Declarations

Ethics approval and consent to participate: Not applicable**.**

Consent for publication: Not applicable.

Conpeting Interests: The authors declare no conflicts of interest.

Author contribution: All authors contributed to the study conception and design. Jiayin Zhang and Shanyong Zhang designed the study. Material preparation and data collection were performed by Yu Zhou, Yuan Qu, Tingting Hou, Wanbao Ge. Data analysis was performed by Jiayin Zhang. The first draft of the manuscript was written by Jiayin Zhang and Shengjie Zhang. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding: This research received no external funding.

Availability of data and material: All data generated or analysed during this study are included in this published article.

References

1. David, J, Hunter, Sita, & Bierma-Zeinstra. (2019). Osteoarthritis.Lancet393(10182):1745-1759. https://doi.org/10.1016/S0140-6736(19)30417-9
2. Wieland, H. A., Michaelis, M., Kirschbaum, B. J., & Rudolphi, K. A. (2005). Osteoarthritis—an untreatable disease?. Nature reviews Drug discovery, 4(4), 331-344.https://doi.org/10.1038/nrd1693
3. Felson, D. T. (2013). Osteoarthritis as a disease of mechanics. Osteoarthritis and cartilage, 21(1), 10-15.https://doi.org/10.1016/j.joca.2012.09.012
4. Spector, T. D., & MacGregor, A. J. (2004). Risk factors for osteoarthritis: genetics. Osteoarthritis and cartilage, 12, 39-44.https://doi.org/10.1016/j.joca.2003.09.005
5. Sinusas, K. (2012). Osteoarthritis: diagnosis and treatment. American family physician, 85(1), 49-56.https://doi.org/10.1136/bmj.1.5222.355-a
6. Martel-Pelletier, J. (2004). Pathophysiology of osteoarthritis. Osteoarthritis and cartilage, 12, 31-33.https://doi.org/10.1016/j.joca.2003.10.002
7. Abramoff, B., & Caldera, F. E. (2020). Osteoarthritis: pathology, diagnosis, and treatment options. Medical Clinics,104(2), 293-311.https://doi.org/10.1016/j.mcna.2019.10.007
8. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415
9. Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larranaga, P., & Lozano, J. A. (2010). Machine learning: an indispensable tool in bioinformatics. In Bioinformatics methods in clinical research (pp.

25-48). Humana Press. https://doi.org/10.1007/978-1-60327-194-3_2

10. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006). Machinelearninginbioinformatics.Briefingsin bioinformatics,7(1),86-112. https://doi.org/10.1093/bib/bbk007
11. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
12. Mundra, P. A., & Rajapakse, J. C. (2009). SVM-RFE with MRMR filter for gene selection. IEEE transactions on nanobioscience, 9(1), 31-37.https://doi.org/10.1109/TNB.2009.2035284
13. Duan, K. B., Rajapakse, J. C., Wang, H., & Azuaje, F. (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE transactions on nanobioscience, 4(3), 228-234. https://doi.org/10.1109/TNB.2005.853657
14. Haseeb, A., & Haqqi, T. M. (2013). Immunopathogenesis of osteoarthritis. Clinical immunology, 146(3), 185-196. https://doi.org/10.1016/j.clim.2012.12.011
15. Woetzel, D., Huber, R., Kupfer, P., Pohlers, D., Pfaff, M., Driesch, D., ... & Kinne, R. W. (2014). Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. Arthritis research & therapy, 16(2), 1-22.https://doi.org/10.1186/ar4526
16. Brophy RH, Zhang B, Cai L, Wright RW et al. Transcriptome comparison of meniscus from patients with and without osteoarthritis. Osteoarthritis Cartilage 2018 Mar;26(3):422-432. https://doi.org/10.1016/j.joca.2017.12.004
17. Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Elana J. Fertig, Andrew E. Jaffe, Yuqing Zhang, John D. Storey and Leonardo Collado Torres (2021).sva: Surrogate Variable Analysis. R package version 3.40.0.
18. Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47. https://doi.org/10.1093/nar/gkv007
19. Raivo Kolde (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. s://CRAN.R-project.org/package=pheatmap
20. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
21. Marc Carlson (2021). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.13.0.
22. T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo, and G Yu. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021, 2(3):100141
23. Guangchuang Yu, Li-Gen Wang, Yanyan Han and Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 2012, 16(5):284-287

24. Guangchuang Yu (2021). enrichplot: Visualization of Functional Enrichment Result. R package version 1.12.3. https://yulab-smu.top/biomedical-knowledge-mining-book/
25. Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. https://www.jstatsoft.org/v33/i01/.
26. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2021). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-9. https://CRAN.R-project.org/package=e1071
27. Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-90. https://CRAN.R-project.org/package=caret
28. Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. http://www.jstatsoft.org/v11/i09/
29. Adrian Dusa (2021). venn: Draw Venn Diagrams. R package version 1.10.https://CRAN.R-project.org/package=venn
30. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. https://doi.org/10.1186/1471-2105-12-77
31. Newman, Aaron M; Liu, Chih Long; Green, Michael R; Gentles, Andrew J; Feng, Weiguo; Xu, Yue; Hoang, Chuong D; Diehn, Maximilian; Alizadeh, Ash A (2015). Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12(5), 453–457. https://doi.org/10.1038/nmeth.3337
32. Taiyun Wei and Viliam Simko (2021). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.90). Available from https://github.com/taiyun/corrplot.
33. Daniel Adler and S. Thomas Kelly (2021). vioplot: violin plot. R package version 0.3.7 https://github.com/TomKellyGenetics/vioplot
34. Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. http://www.jstatsoft.org/v21/i12/.
35. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
36. Dean Attali and Christopher Baker (2019). ggExtra: Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements. R package version 0.9. https://CRAN.R-project.org/package=ggExtra
37. Zahan, O. M., Serban, O., Gherman, C., & Fodor, D. (2020). The evaluation of oxidative stress in osteoarthritis. Medicine and pharmacy reports, 93(1), 12. https://doi.org/10.15386/mpr-1422
38. Lepetsos, P., & Papavassiliou, A. G. (2016). ROS/oxidative stress signaling in osteoarthritis. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1862(4), 576-591. https://doi.org/10.1016/j.bbadis.2016.01.003

39. Farran, A., Valverde-Franco, G., Paré, F., Tío, L., Monfort, J., Pelletier, J. P., & Martel-Pelletier, J. (2017). In vivo effect of opticin deficiency in a surgically induced model of osteoarthritis. Osteoarthritis and Cartilage, 25, S61. https://doi.org/10.1016/j.joca.2017.02.111
40. Ni, G. X., Li, Z., & Zhou, Y. Z. (2014). The role of small leucine-rich proteoglycans in osteoarthritis pathogenesis. Osteoarthritis and cartilage, 22(7), 896-903. https://doi.org/10.1016/j.joca.2014.04.026
41. Reyes, N., Rebollo, J., & Geliebter, J. (2019). Effects of NSAIDs on gene expression of small leucine-rich proteoglycans in prostate cancer cells. https://doi.org/10.1158/1538-7445.AM2019-128
42. Wang, Q., Lepus, C. M., Raghu, H., Reber, L. L., Tsai, M. M., Wong, H. H., ... & Robinson, W. H. (2019). IgE-mediated mast cell activation promotes inflammation and cartilage destruction in osteoarthritis. Elife, 8, e39905. https://doi.org/10.7554/eLife.39905
43. Li, Y. S., Luo, W., Zhu, S. A., & Lei, G. H. (2017). T cells in osteoarthritis: alterations and beyond. Frontiers in immunology, 8, 356. https://doi.org/10.3389/fimmu.2017.00356
44. Gol-Ara, M., Jadidi-Niaragh, F., Sadria, R., Azizi, G., & Mirshafiey, A. (2012). The role of different subsets of regulatory T cells in immunopathogenesis of rheumatoid arthritis. Arthritis, 2012. https://doi.org/10.1155/2012/805875
45. Zaiss, M. M., Frey, B., Hess, A., Zwerina, J., Luther, J., Nimmerjahn, F., ... & David, J. P. (2010). Regulatory T cells protect from local and systemic bone destruction in arthritis. The Journal of Immunology, 184(12), 7238-7246. https://doi.org/10.4049/jimmunol.0903841

Figures

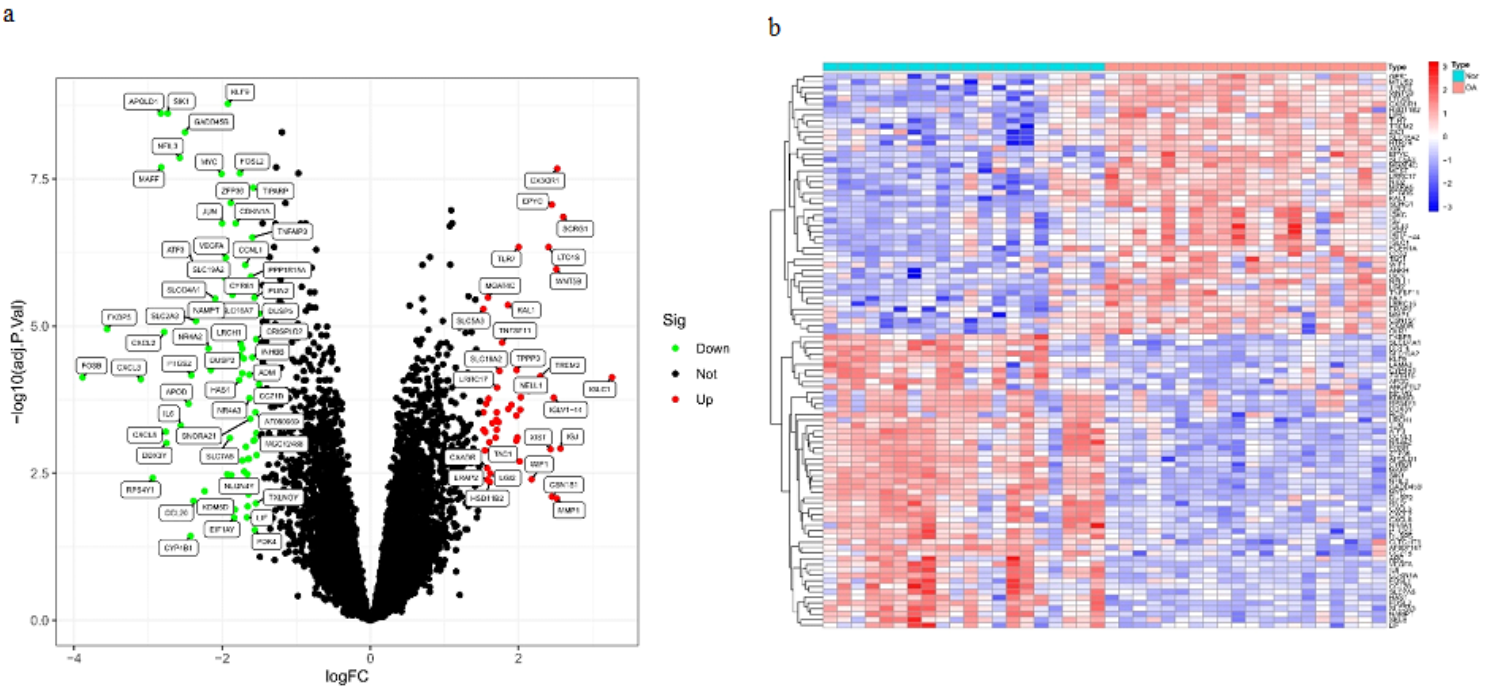
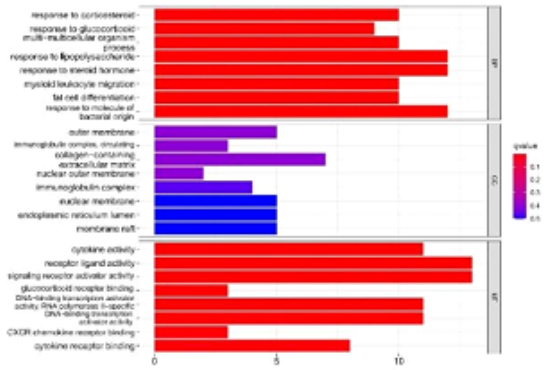


Figure 1

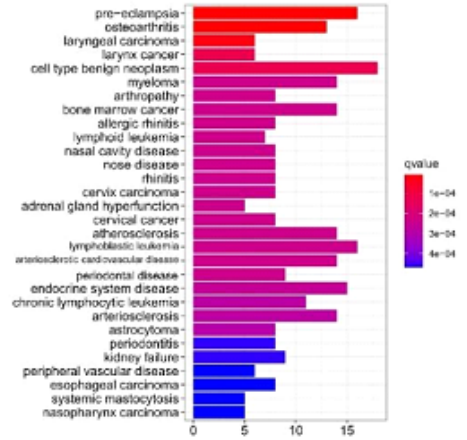
Screening of differentially expressed genes. (a) Volcano map of DEGs; red represents up-regulated differential genes, black represents no significant difference genes, and green represents down-regulated differential genes.

(b)The thermal map of expression level of different genes in every synovial tissue sample, the redder the color, the higher the expression, the bluer the color, the lower the expression.

a



b



c

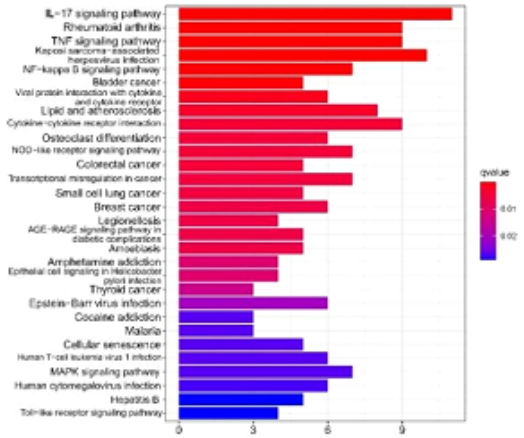


Figure 2

Gene Ontology (GO), Disease Ontology (DO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of DEGs. (a) GO enrichment analysis, where the horizontal axis represents the number of DEGs under the GO term. (b) DO enrichment analysis, where the horizontal axis represents the number of DEGs under the DO term. (c) KEGG enrichment analysis, where the horizontal axis represents the number of DEGs under the KEGG term.

Figure 3

Gene GO and KEGG enrichment analysis of all normal genes and all OA genes.

(a)GSEA-GO enrichment analysis on all normal genes, saved the top five enriched pathways.

(b) GSEA-GO enrichment analysis on all OA genes, saved the top five enriched pathways.

(c) GSEA-KEGG enrichment analysis on all normal genes, saved the top five enriched pathways.

(d) GSEA-KEGG enrichment analysis on all OA genes, saved the top five enriched pathways.

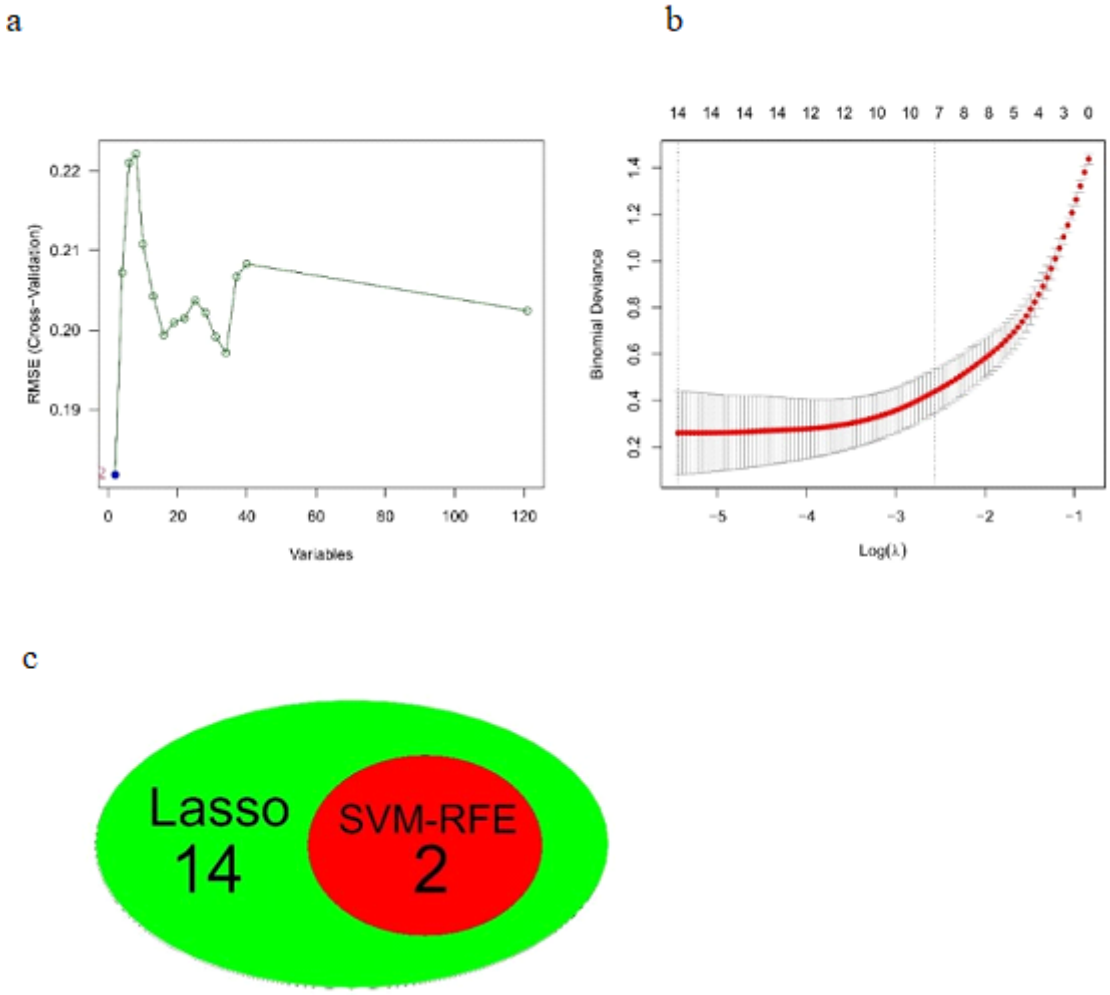


Figure 4

Screening of diagnostic markers. (a) Least absolute shrinkage and selection operator (LASSO) logistic regression algorithm to screen diagnostic markers. (b) Support vector machine-recursive feature elimination (SVM-RFE) algorithm to screen diagnostic markers. (c) Venn diagram shows the intersection of diagnostic markers obtained by the two algorithms.

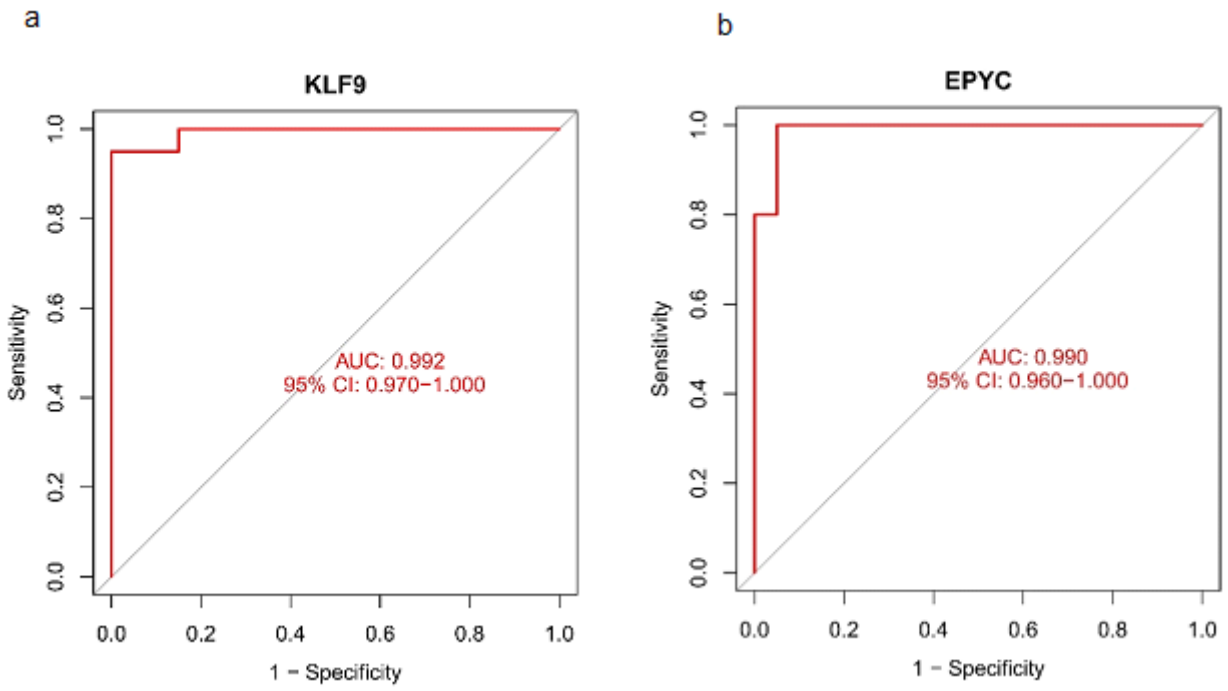


Figure 5

ROC curve of KLF9(a) and EPYC(b) genes in the training set.

Figure 6

Box diagram of difference analysis of the expression levels of KLF9(a) and EPYC(b) in the validation set. The blue marks represent the normal, The red marks represent the OA.

Figure 7

ROC curve of KLF9(a) and EPYC(b) genes in the validation set.

Figure 8

Evaluation and visualization of immune cell infiltration. (a) Content of different immune cells in each sample. (b)Correlation heat map of 22 types of immune cells. red represents a positive correlation, blue represents a negative correlation. The darker the color, the stronger the correlation. (c) Violin diagram of the proportion of 22 types of immune cells. The red marks represent the difference in infiltration between the two groups of samples

Figure 9

Correlation between KLF9 gene expression and different immune cells infiltrating.

Figure 10

Correlation between EPYC gene expression and different immune cells infiltrating.

Figure 11

Correlation between KLF9, EPYC, and infiltrating immune cells. (a) Correlation between KLF9 and infiltrating immune cells. (b) Correlation between EPYC and infiltrating immune cells. The size of the dots represents the strength of the correlation between genes and immune cells; the larger the dots, the stronger the correlation, and the smaller the dots, the weaker the correlation. The color of the dots represents the p-value, the greener the color, the lower the p-value, and the yellower the color, the larger the p-value. $p < 0.05$ was considered statistically significant.