TLAR
1947

ORIGINAL ARTICLE

Identification of the candidate genes of diagnosing rheumatoid arthritis using the single-cell sequencing technology and T cell subclusters analysis of patients with rheumatoid arthritis

Yajing Liu[1] , Shaoguang Fan[2] , Shan Meng[1]

[1]Department of Rheumatology and Immunology, Affiliated Hospital of Weifang Medical University, Weifang, China
[2]Department of Rheumatology and Immunology, Institute of Traumatic Orthopedics, The 80th Army Hospital of the Chinese People's Liberation Army, Weifang, China

ABSTRACT

Objectives: This study aims to analyze the heterogeneity among different cell types in peripheral blood mononuclear cells (PBMC) in rheumatoid arthritis (RA) patients and to analyze T cell subsets to obtain key genes that may lead to RA.

Materials and methods: The sequencing data of 10,483 cells were obtained from the GEO data platform. The data were filtered and normalized initially and, then, principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (TSNE) cluster analysis were performed using the Seurat package in R language to group the cells, thereby obtaining the T cells. The T cells were subjected to subcluster analysis. The differentially expressed genes (DEGs) in T cell subclusters were obtained, and the hub genes were determined by Gene Ontology (GO) functional enrichment analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis and protein-protein interaction (PPI) network construction. Finally, the hub genes were validated using other datasets in the GEO data platform.

Results: The PBMC of RA patients were mainly divided into T cells, natural killer (NK) cells, B cells, and monocyte cells. The number of T cells was 4,483, which were further divided into seven clusters. The pseudotime trajectory analysis showed that the differentiation of T cells developed from cluster 0 and cluster 1 to cluster 5 and cluster 6. Through GO, KEGG and PPI analysis, the hub genes were identified. After validation by external data sets, nine genes were identified as candidate genes highly associated with the occurrence of RA, including CD8A, CCL5, GZMB, NKG7, PRF1, GZMH, CCR7, GZMK, and GZMA.

Conclusion: Based on single-cell sequencing analysis, we identified nine candidate genes for diagnosing RA, and validated their diagnostic value for RA patients. Our findings may provide new sights for the diagnosis and treatment of RA.

Keywords: Diagnostic biomarkers, rheumatoid arthritis, single-cell RNA-seq, T cells.

Rheumatoid arthritis (RA) is a chronic autoimmune disease and has the characteristics of a progressive symmetric of the involved joints, which can result in both cartilage destruction and bone erosion.[1] The incidence of RA ranges from 0.4 to 1.3% all over the world, its incidence in women are usually three times higher than that in men, and frequency of new RA diagnoses peaks at the sixth decade of life.[2] The pathological characteristics of RA include synovial lining cell proliferation, infiltration of a large number of inflammatory cells in the interstitium, the new generation of microvessels, the formation of pannus, and the destruction of cartilage and bone tissue.[3,4] The pathogenesis of RA may be related to heredity, infection, or sex hormones.[5,6]

However, to date, the specific etiology and pathogenesis of RA are still unknown.

It is known that a variety of immune cells represented by T cells, B cells, macrophages, and natural killer (NK) cells participate in and mediate the autoimmune response of RA, and the autoimmune response induce the secretion of chemokines and proteases, which disrupts immune homeostasis, thereby further accelerating cartilage and bone erosion.[7-10] T lymphocytes are derived from lymphoid stem cells in the bone marrow, mature in the thymus, and mainly mediate adaptive cellular immune responses. According to whether CD4 is expressed in the T cells, CD4+ T cells are identified, and Th1 and Th17, as the main subtypes of the CD4+ T cells are currently considered to be involved in the pathogenesis of RA.[11,12] Wang et al.[13] suggested that Th17/Treg balance was broken in peripheral blood from RA patients, which may play an important role in the development of RA. At present, although targeted drugs for immune cells or the inflammatory factors produced by them have been put into clinical practice, and good curative effects have been obtained, the specific mechanism of T cells involved in the occurrence of RA is still unclear. Therefore, in-depth study of T cells in the RA patients may provide new strategies for the diagnosis and management of RA.

With the development of bioinformatics technology, microarrays based on high-throughput platforms have emerged as an efficient tool to investigate gene expression profiles and explore molecular mechanisms of various diseases. Rheumatoid arthritis is a chronic systemic disease characterized by inflammatory synovitis with unknown etiology. Synovial tissues are the important tissue in gene analysis. Bioinformatics analysis can screen out promising biomarkers and analysis pathways and immune infiltration patterns in RA, which may shed light on the understanding of the pathogenesis, occurrence, development, and consequently provide new strategies for its diagnosis, prevention, and treatment.[14]

Until now, as a powerful tool, single-cell ribonucleic acid (RNA) sequencing (scRNA-seq) analysis has been used to reveal cell diversity and cell-to-cell communication at single-cell level. This scRNA-seq analysis approach makes an improvement in understanding cellular heterogeneity and easy to screen out promising biomarkers for diagnosis, prognosis, and therapeutic response of RA. Although several previous studies[15,16] have performed scRNA-seq analysis on RA cells. It is rarely known the heterogeneity between T cells subtypes. Therefore, in the present study, a comprehensive analysis of scRNA-seq data was performed to reveal the development map of T cells, and develop a novel gene signature to accurately predict the prognosis of RA.

MATERIALS AND METHODS

Data collection and preparation

We downloaded the single-cell sequencing data in the GSE159117 dataset from the GEO data platform (https://www.ncbi.nlm.nih.gov/geo), which utilizes the 10× Genomics scRNA-seq method to analyze peripheral blood mononuclear cell (PBMC) samples isolated from peripheral blood of RA patients. A total of 10,483 single-cell sequencing data from RA patients were obtained. Then, we used the Seurat package in R language to process the data, and used the "CreateSeuratObject" function to build a Seurat object.

Quality control, data filtering, standardization

We used the Percentage FeatureSet function to calculate the percentage of mitochondrial genes, and used the "FilterCells" function to filter the Seurat object, with the retention criteria of 200<features<2,500 and the percentage of mitochondrial genes <5%. We used the "NormalizeData" function to normalize the data.

Identification of hypervariable genes between cells

The expression of some genes varies greatly between different cells, which are called hypervariable genes. We used the "FindVariableGenes" function to detect hypervariable genes, and used the "vst" method to select the top 2,000 hypervariable genes. The data is further scaled using the "ScaleData" function.

Principal component analysis (PCA)

We used the "RunPCA" function to perform PCA on hypervariable genes, used the "JackStraw" and "ScoreJackStraw" functions to score each principal component, and used the "JackStrawPlot" and "EbowPlot" functions to visualize the score of each principal component. The dimension of the principal component for cluster analysis is selected accordingly.

T-Distributed Stochastic Neighbor Embedding (TSNE) cluster analysis

We used the "RunTSNE" function to perform TSNE cluster analysis, and used the "TSNEPlot" function to visualize cell clusters.

Cell cluster annotation

Cell clustering was annotated using the SingleR package in the R language. We extracted T cells as the follow-up research objects.

T cell subset grouping and pseudotime trajectory analysis

The TSNE cluster analysis was performed again using the Seurat package to determine the T cell subclusters. We used "FindMarkers" to determine the differentially expressed genes (DEGs) between T cell subgroup and other cells, with the screening criteria of logFC>0.25 and adjPval<0.05.

The pseudotemporal ordering of single cells using Monocle R package was performed to reveal the dynamic changes in the transcriptome of developing T cells.

Functional annotation of T cell subclusters

To reveal the biological functions of the DEGs, enrichment analysis was performed on the obtained DEGs. Gene Ontology (GO) functional enrichment analysis, including three aspects of cell composition (CC), biological process (BP), and molecular function (MF), and Kyoto Encyclopedia of Genes and Genomes (KEGG) signal pathway enrichment analysis were performed by Metascape online analysis tool (https://metascape.org) to investigate functional relationships between subgroups.

Construction of protein-protein interaction (PPI) network and identification of hub genes

The top 10 DEGs in each T cell subcluster were selected to construct the PPI network.

The STRING online database (http://string-db.org) was used to do the PPI network analysis for the DEGs, and then the Cytoscape 3.6.1 (http://www.cytoscape.org/) was used to visualize the PPI network, and the CytoHubba plug-in was used to identify the hub genes.

Validation of the hub genes

The GSE12021, GSE55457, and GSE55584 were used as the validation datasets to analyze and evaluate the value of the hub genes in the diagnosis of RA.

RESULTS

Quality control of single-cell sequencing data

The quality control chart is shown in Figure 1a, where the range of detected gene numbers and the sequencing count of each cell are illustrated. We accordingly filtered cells with a percentage of mitochondrial sequencing count <5% and 200<features<2,500 to obtain 8,564 cells with high quality.

In addition, there was no correlation between counts and the percentage of mitochondrion (Pearson r=-0.08, Figure 1b), while there was a significant correlation between features and counts (Pearson r=0.94, Figure 1c).

Identification of hypervariable genes

Top 2,000 hypervariable genes were selected and, among them, the top 10 genes were S100A9, LYZ, S100A8, GNLY, CCL3, IGLL5, IFI27, CCL4, CDKN1C and LST1 (Figure 1d).

PCA analysis

Furthermore, we used the PCA method to screen the significantly correlated genes in each component. The top 30 significantly correlated genes are shown in the dot plots (Figure 1e and 1f) and heat map (Figure 1g).

Determination of the number of dimensions

The PCA, a linear dimensionality reduction method, was utilized to identify the significantly available dimensions of data sets with estimated *p* value. JackStraw Plot (Figure 1h) and Elbow plot

(Figure 1i) showed that 10 principal components were adopted.

TSNE cluster analysis

The TSNE cluster analysis showed the cells could be divided into 10 clusters, as shown in Figure 2a, indicating that there was an obvious heterogeneity among PBMC cells in RA, and the SingleR package was used to annotate each cluster as various types of cells. The cell type of each cluster is shown in Figure 2b. The top 10 DEGs in each cluster are shown in Figure 2c.

Analysis of T cell subclusters

The T cells obtained in the above analysis were further grouped, and a total of 4,483 cell data were obtained. The Seurat package was used again for grouping analysis, and the T cells were divided into seven subclusters, which are shown in Figure 3a. The top 10 DEGs in each subcluster are shown in Figure 3b.

Pseudotime trajectory analysis of T cell subclusters

The differentiation trajectory and pseudotime analysis was investigated using the
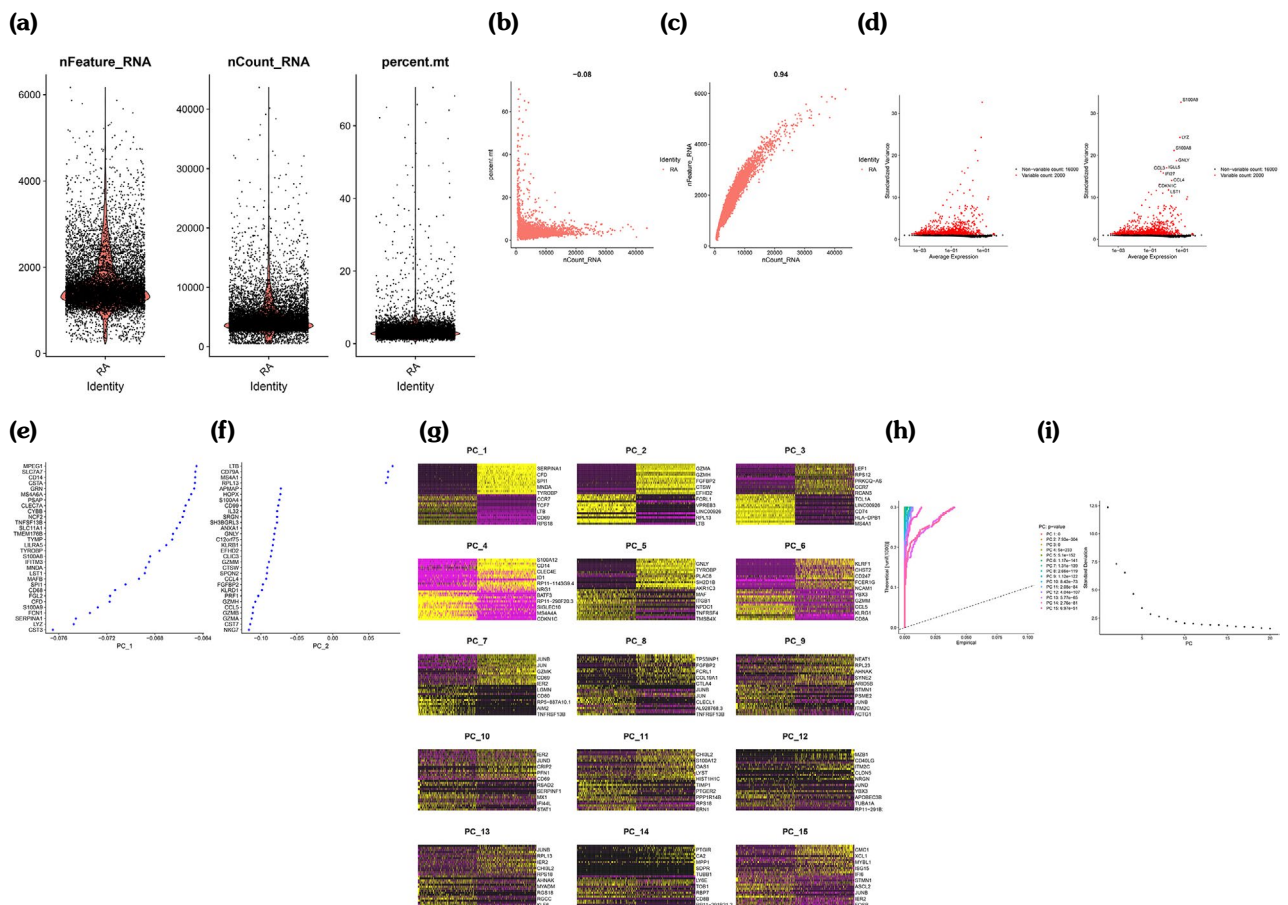

Figure 1. Characterization of single cell sequencing data from peripheral blood mononuclear cell (PBMC) samples isolated from peripheral blood of rheumatoid arthritis (RA) patients. **(a)** Quality control of single cell RNA sequencing data. The Y axes represented RNA features, RNA counts and percentage of mitochondrial genes. The cells with high quality were filtered out. **(b)** The relationship between the percentage of mitochondrial genes and the RNA counts. **(c)** The relationship between RNA features and the RNA counts. **(d)** Red dots represented the 2000 hypervariable genes. The top 10 gene names were labeled out. **(e)** and **(f)** The top 30 significantly correlated genes in the top 2 principal components. **(g)** The heatmap of top 10 genes in the top 15 principal components. **(h)** and **(i)** The JackStrawPlot **(h)** and ElbowPlot **(i)** of principal components, which were used to identify the significantly available dimensions of data sets with estimated *p* value and Elbow.

Monocle2 R package based on the identified marker genes from each cluster. As shown in the Figure 3C and 3D, T cell differentiation developed from cluster 0 and cluster 1 to cluster 5 and cluster 6.

GO and KEGG analysis

We used the Metascape online analysis tool to perform GO function enrichment analysis and KEGG pathway enrichment analysis on
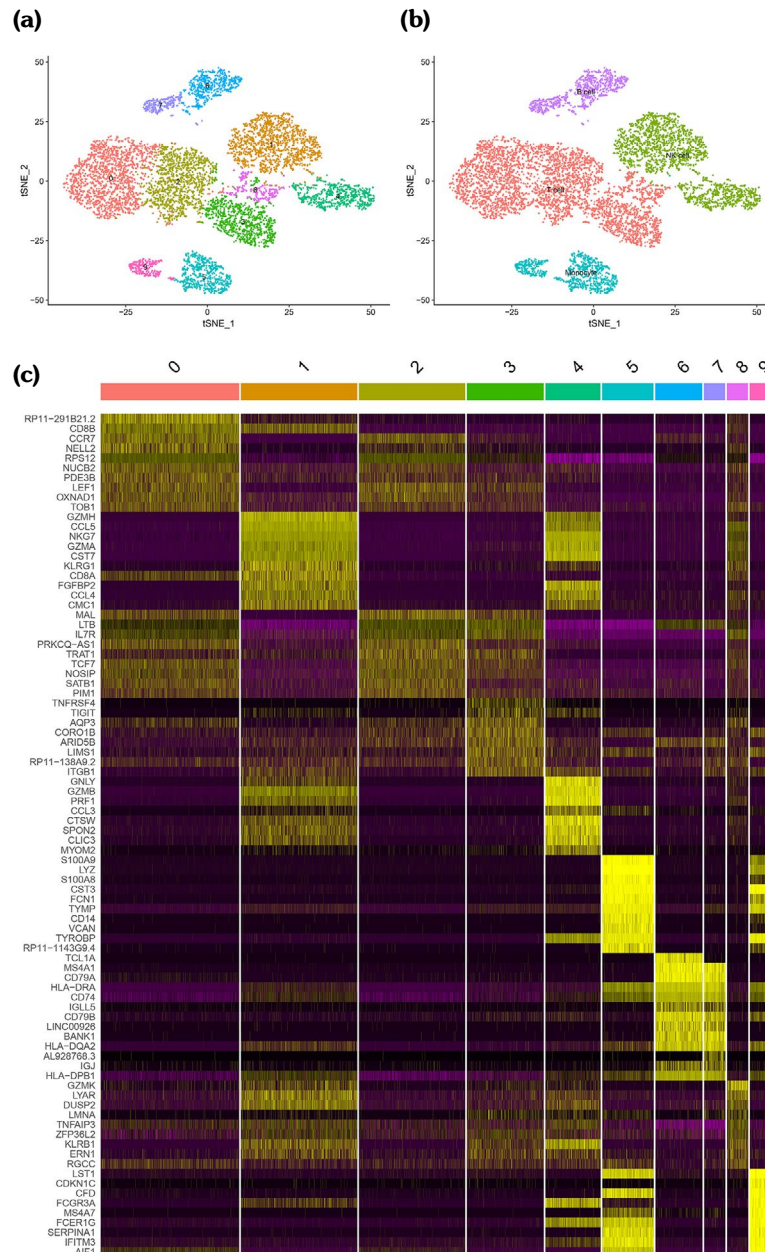

Figure 2. Clusters of single-cell RNA sequencing data. **(a)** Based on available significant components, we conducted TSNE algorithm to learn the underlying manifold of the data in order to place similar cells together in low-dimensional space. 10 cell clusters were obtained. **(b)** Cells were annotated into 4 main cell clusters: T cells, NK cells, B cells and Monocyte. **(c)** Heatmap of the top 10 differentially expressed genes in each cluster.

RNA: Ribonucleic acid: TSNE: T-Distributed Stochastic Neighbor Embedding; NK: Natural killer.

DEGs in each subcluster. The results revealed that the DEGs were markedly enriched in BPs such as T cell activation, Epstein-Barr virus infection, positive regulation of immune response, regulation of defense response, and Epstein-Barr virus infection pathway signaling (Figure 3e).

Construction of PPI network and identification of hub genes

The FindMarkers function in the Seurat package was used to find the DEGs between T cell subclusters, and the top 10 DEGs of each subcluster were selected to establish a PPI network, as shown in Figure 4a. The degree of association between nodes in the PPI network was calculated by the Cytohubba plugin, and the top 10 hub genes were identified according to the degree of association, including CD8A, CCL4, CCL5, GZMB, NKG7, PRF1, GZMH, CCR7, GZMK, and GZMA. The expression of the top 10 gene in each subcluster is shown in Figure 4b and 4c, suggesting that the top 10 genes had obvious differences between the subclusters.

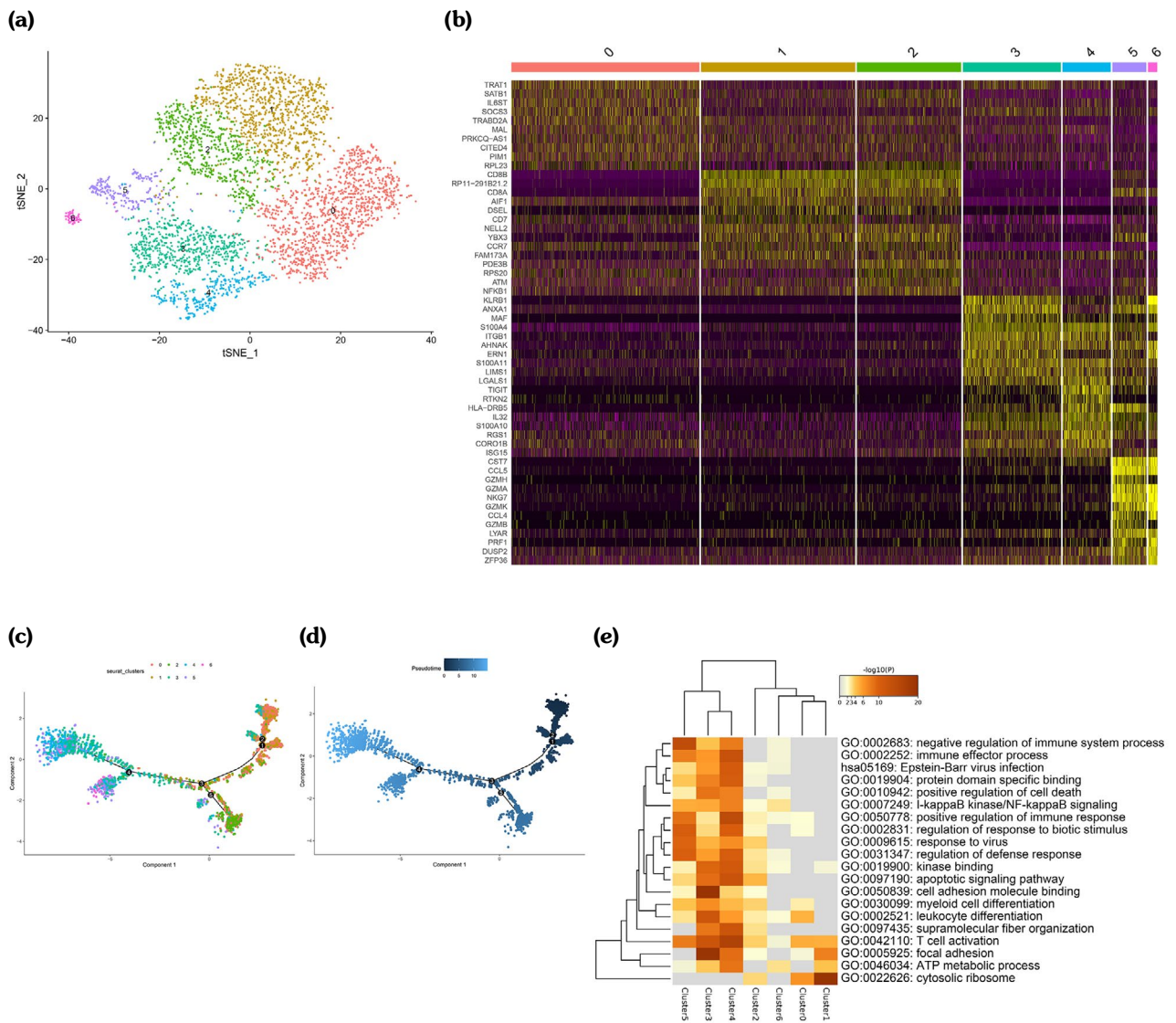

Figure 3. Subclusters of the T cells. **(a)** T cells were divided into 7 subclusters. **(b)** Heatmap of top 10 differentially expressed genes in each subclusters. **(c)** Pseudotemporal trajectory of T cells subclusters using the Monocle algorithm. **(d)** T cells evolved from dark blue to light blue. **(e)** Seven T cell subclusters were functionally annotated using the Metascape tool.

External dataset validation

The GSE12021, GSE55457, and GSE55584 were selected as validation datasets, including 35 RA patients and 19 healthy controls. The accuracy of the 10 hub genes in predicting RA was analyzed, and the receiver operating characteristic (ROC) curves of them in diagnosing RA were drawn. The results are shown in Figure 4d and Table 1. Only the area under the curve (AUC) of CCL4 was 0.57, and the AUCs of the other nine genes ranged from 0.789 to 0.922, indicating high accuracy. Therefore, nine genes including CD8A, CCL5, GZMB, NKG7, PRF1, GZMH, CCR7, GZMK and GZMA might

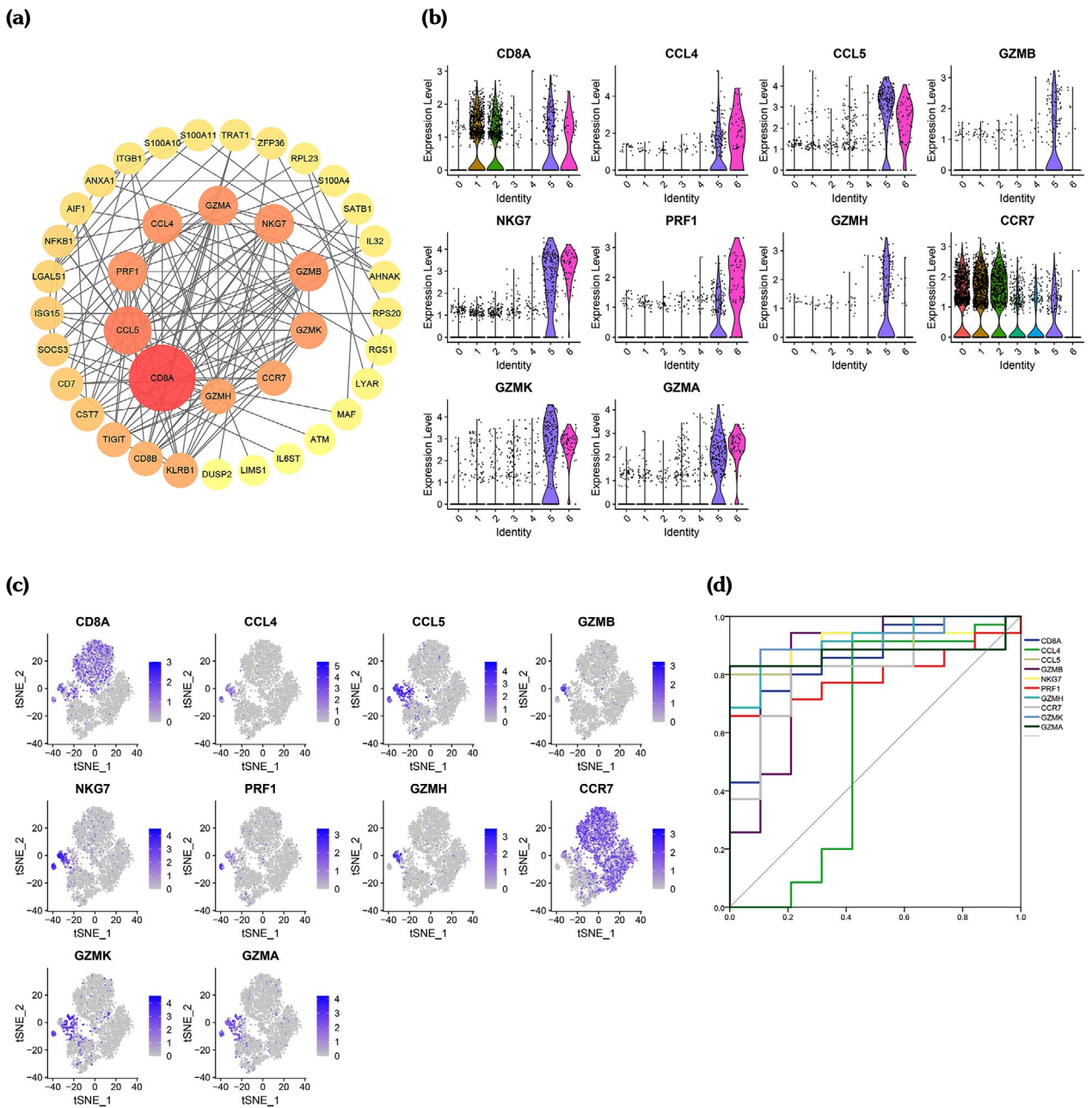**Figure 4.** Identification and validation of the hub genes. **(a)** Construction of PPI network using marker genes of each subclusters to identify 10 hub genes. **(b)** The expression level of the hub genes in all 7 cell subclusters. **(c)** The distributions of the hub genes in each subclusters. **(d)** Receiver operator characteristic (ROC) curves of the hub genes in the validation datasets.

Table 1. AUC values of top 10 genes in diagnosing RA.

Gene	AUC
CD8A	0.856
CCL4	0.57
CCL5	0.904
GZMB	0.847
NKG7	0.913
PRF1	0.792
GZMH	0.91
CCR7	0.789
GZMK	0.922
GZMA	0.874

AUC: Area under the curve; RA: Rheumatoid arthritis

be selected as the candidate genes for diagnosing RA.

DISCUSSION

Rheumatoid arthritis is a systemic autoimmune disease characterized by joint synovitis. Innate and adaptive immune responses play integral roles in the pathogenesis of RA. Abnormal numbers and functions of T cells, B cells, and NK cells play important roles in the pathogenesis and development of RA. Previous studies have mainly focused on the screening of biomarkers via analyzing DEGs between RA and osteoarthritis (OA).[17,18] However, bulk sample transcriptome analysis cannot consider the effects of heterogeneity between cells in different tissues on the results and, therefore, it has great limitations. Therefore, we use single-cell transcriptome analysis to group cells based on the single-cell data from the GEO database to screen the T cells for subcluster analysis.

T cells differentiate to Th1, Th2, or Th17 T cells, and provide help for B cells for the production of (auto) antibodies. Th1 secretes cytokines such as interleukin-2 (IL-2), tumor necrosis factor-alpha (TNF-α) and interferon-gamma (IFN-γ) to mediate cellular immunity. Th2 secretes IL-4, IL-6 and IL-10 and other cytokines to mediate fluid immunity; Th17

secretes IL-17, IL-21 and other inflammatory cytokines to mediate autoimmunity; Treg secretes IL-10, transforming growth factor-β (TGF-β) and other regulatory immune responses, which plays a negative regulatory role in the immune system.[19]

In the present study, we identified nine hub genes including CD8A, CCL5, GZMB, NKG7, PRF1, GZMH, CCR7, GZMK and GZMA, which had high accuracy in diagnosing RA. Among nine biomarkers, some have been reported as prognostic biomarkers of RA by other studies.

The CD8A encodes the CD8α chain of the dimeric CD8 protein, The CD8 antigen is a cell surface glycoprotein found on most cytotoxic T lymphocytes that mediates efficient cell-cell interactions within the immune system. The CCL5, belonging to the C-C chemokine family, mainly induces the recruitment and migration of immune cells to inflammatory sites.[20] The GZMA, a member of the serine protease family, is secreted by cytotoxic cells such as cytotoxic T cells and NK cells, and plays an important role in cell death, cytokine processing, and inflammation. It has been reported that CD8A, CCL5, and GZMA were significantly overexpressed in the early RA, were the hub genes of the entire process of RA development, and might be used as the diagnostic biomarkers for RA.[21]

Long et al.[22] introduced and successfully validated a 16-gene signature derived from gene expression profiling data of synovial tissues, and found that the gene expression levels of GZMB and GZMK were usually higher in RA than that in OA, and they were selected as marker genes of RA. Myles et al.[23] found that synovial fluid (SF) mononuclear cell gene expression profiling suggested dysregulation of innate immune genes in enthesitis-related arthritis patients, including GZMH and NKG7.

Perforin (PRF; encoded by the PRF1 gene) is a pore-forming protein stored in secretory granules of cytotoxic lymphocytes. The polymorphisms of PRF1 seem to be responsible for more aggressive RA disease phenotype.[24]

Expression of CCR7 was found to be the hallmark of RA synovial fluid M1 macrophages (Mϕs) and its levels were potentiated in response to M1 mediating factors.[25]

The above nine genes have been verified by external data sets, and their diagnostic accuracy rates are in the range of 0.789 to 0.922, which have a high diagnostic value. It can provide useful help for the diagnosis and treatment of RA. Our results are considered to be important for understanding the pathological mechanisms and in designing the strategies for preventing RA progression.

In addition, there are still some limitations in our study. First, this prognostic signature is developed using retrospective analysis based on GEO data, thus, the results need to be further confirmed in prospective trails. Second, further validation studies with a large cohort of patients are necessary to demonstrate its usefulness in clinical application.

In conclusion, we identified nine candidate genes (CD8A, CCL5, GZMB, NKG7, PRF1, GZMH, CCR7, GZMK and GZMA) for diagnosing RA based on single-cell sequencing analysis, and validated their diagnostic value for RA patients. Our findings may provide new sights for the diagnosis and treatment of RA.

Ethics Committee Approval: Ethics approval and consent to participate: Not applicable, because GEO belongs to public databases, the patients involved in the database have obtained ethical approval, users can download relevant data for free for research and publish relevant articles, and our study is based on open-source data, and the Affiliated Hospital of Weifang Medical University do not require research using publicly available data to be submitted for review to their ethics committee. The study was conducted in accordance with the principles of the Declaration of Helsinki.

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: Substantial contributions to the conception and design of the work: Y.L., S.M.; The acquisition, analysis, and interpretation of data for the work: Y.L., S.F.; Drafting the work: Y.L.; Revising it critically for important intellectual content: S.M.; Final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: Y.L., S.F., S.M.

Conflict of Interest: The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

REFERENCES

1. Smolen JS, Aletaha D, McInnes IB. Rheumatoid arthritis. Lancet 2016;388:2023-38.
2. Lin YJ, Anzaghe M, Schülke S. Update on the pathomechanism, diagnosis, and treatment options for rheumatoid arthritis. Cells 2020;9:880.
3. Rheumatoid arthritis. Nat Rev Dis Primers 2018;4:18002.
4. Aletaha D, Smolen JS. Diagnosis and management of rheumatoid arthritis: A review. JAMA 2018;320:1360-72.
5. Scherer HU, Häupl T, Burmester GR. The etiology of rheumatoid arthritis. J Autoimmun 2020;110:102400.
6. Croia C, Bursi R, Sutera D, Petrelli F, Alunno A, Puxeddu I. One year in review 2019: pathogenesis of rheumatoid arthritis. Clin Exp Rheumatol 2019;37:347-57.
7. Wu F, Gao J, Kang J, Wang X, Niu Q, Liu J, et al. B cells in rheumatoid arthritis: Pathogenic mechanisms and treatment prospects. Front Immunol 2021;12:750753.
8. Schwaneck EC, Renner R, Junker L, Tony HP, Kleinert S, Gernert M, et al. T cells, natural killer cells, and γδT cells in a large patient cohort with rheumatoid arthritis: Influence of age and anti-rheumatic therapy. Scand J Rheumatol 2020;49:8-12.
9. Jiang Q, Yang G, Liu Q, Wang S, Cui D. Function and role of regulatory T cells in rheumatoid arthritis. Front Immunol 2021;12:626193.
10. Fathollahi A, Samimi LN, Akhlaghi M, Jamshidi A, Mahmoudi M, Farhadi E. The role of NK cells in rheumatoid arthritis. Inflamm Res 2021;70:1063-73.
11. Picerno V, Ferro F, Adinolfi A, Valentini E, Tani C, Alunno A. One year in review: The pathogenesis of rheumatoid arthritis. Clin Exp Rheumatol 2015;33:551-8.
12. Luckheeram RV, Zhou R, Verma AD, Xia B. CD4$^+$ T cells: Differentiation and functions. Clin Dev Immunol 2012;2012:925135.
13. Wang W, Shao S, Jiao Z, Guo M, Xu H, Wang S. The Th17/Treg imbalance and cytokine environment in peripheral blood of patients with rheumatoid arthritis. Rheumatol Int 2012;32:887-93.
14. Zhang R, Zhou X, Jin Y, Chang C, Wang R, Liu J, et al. Identification of differential key biomarkers in the synovial tissue between rheumatoid arthritis and osteoarthritis using bioinformatics analysis. Clin Rheumatol 2021;40:5103-10.
15. Cheng L, Wang Y, Wu R, Ding T, Xue H, Gao C, et al. New insights from single-cell sequencing data: Synovial fibroblasts and synovial macrophages in rheumatoid arthritis. Front Immunol 2021;12:709178.

16. Cai S, Ming B, Ye C, Lin S, Hu P, Tang J, et al. Similar transition processes in synovial fibroblasts from rheumatoid arthritis and osteoarthritis: A single-cell study. J Immunol Res 2019;2019:4080735.

17. Zhu N, Hou J, Wu Y, Li G, Liu J, Ma G, et al. Identification of key genes in rheumatoid arthritis and osteoarthritis based on bioinformatics analysis. Medicine (Baltimore) 2018;97:e10997.

18. Cai P, Jiang T, Li B, Qin X, Lu Z, Le Y, et al. Comparison of rheumatoid arthritis (RA) and osteoarthritis (OA) based on microarray profiles of human joint fibroblast-like synoviocytes. Cell Biochem Funct 2019;37:31-41.

19. Cope AP, Schulze-Koops H, Aringer M. The central role of T cells in rheumatoid arthritis. Clin Exp Rheumatol 2007;25(5 Suppl 46):S4-11.

20. Aldinucci D, Colombatti A. The inflammatory chemokine CCL5 and cancer progression. Mediators Inflamm 2014;2014:292376.

21. Zhou S, Lu H, Xiong M. Identifying Immune cell infiltration and effective diagnostic biomarkers in rheumatoid arthritis by bioinformatics analysis. Front Immunol 2021;12:726747.

22. Long NP, Park S, Anh NH, Min JE, Yoon SJ, Kim HM, et al. Efficacy of integrating a novel 16-gene biomarker panel and intelligence classifiers for differential diagnosis of rheumatoid arthritis and osteoarthritis. J Clin Med 2019;8:50.

23. Myles A, Tuteja A, Aggarwal A. Synovial fluid mononuclear cell gene expression profiling suggests dysregulation of innate immune genes in enthesitis-related arthritis patients. Rheumatology (Oxford) 2012;51:1785-9.

24. Perricone C, Ceccarelli F, Valesini G. An overview on the genetic of rheumatoid arthritis: a never-ending story. Autoimmun Rev 2011;10:599-608.

25. Van Raemdonck K, Umar S, Shahrara S. The pathogenic importance of CCL21 and CCR7 in rheumatoid arthritis. Cytokine Growth Factor Rev 2020;55:86-93.