Three hematologic/immune system-specific expressed genes are considered as the potential biomarkers for the diagnosis of early rheumatoid arthritis through bioinformatics analysis

Qi Cheng
Second Affiliated Hosptial of Zhejiang University School of Medicine https://orcid.org/0000-0003-0720-3188

Huaxiang Wu
Department of Rheumatology, the Second Affiliated Hospital of Zhejiang University School of Medicine, 88 Jiefang Road, Hangzhou, 310009, China

Yan Du (duyan2014@zju.edu.cn)
Department of Rheumatology, the Second Affiliated Hospital of Zhejiang University School of Medicine, 88 Jiefang Road, Hangzhou, 310009, China https://orcid.org/0000-0001-9931-4197

Research

Keywords: early rheumatoid arthritis, tissue-specific expressed genes, biomarker, microarray, bioinformatics analysis, RNA regulatory pathways

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-73142/v1

Version of Record: A version of this preprint was published on January 6th, 2021. See the published version at https://doi.org/10.1186/s12967-020-02689-y.

Abstract

Background Rheumatoid arthritis (RA) is the most common chronic autoimmune connective tissue disease. However, early RA is difficult to diagnose due to the lack of effective biomarkers. This study aimed to identify new biomarkers and mechanisms for RA disease progression at the transcriptome level through a combination of microarray and bioinformatics analyses.

Methods Microarray datasets for synovial tissue in RA or osteoarthritis (OA) were downloaded from the Gene Expression Omnibus (GEO) database, and differentially expressed genes (DEGs) were identified by R software. Tissue/organ-specific genes were recognized by BioGPS. Enrichment analyses were performed and protein-protein interaction (PPI) networks were constructed to understand the functions and enriched pathways of DEGs and to identify hub genes. Cytoscape was used to construct the co-expressed network and competitive endogenous RNA (ceRNA) networks. Biomarkers with high diagnostic value for the early diagnosis of RA were validated by GEO datasets. The ggpubr package was used to perform statistical analyses with Student's t-test.

Results A total of 275 DEGs, including 197 downregulated genes and 78 upregulated genes, were identified between the samples from RA and OA. Among these DEGs, 71 tissue/organ-specific expressed genes were recognized. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis indicated that DEGs are mostly enriched in immune response, immune-related biological process, immune system, and cytokine signal pathways. Fifteen hub genes and 4 gene cluster modules were identified by Cytoscape. Eight haematologic/immune system-specific expressed hub genes were verified by GEO datasets, and three genes (granzyme A (GZMA), protein regulator of cytokinesis 1 (PRC1), and threonine/tyrosine kinase protein kinase (TTK)) that have a high diagnostic value for early RA were identified. NEAT1-miR-212-3p/miR-132-3p/miR-129-5p-TTK, XIST-miR-25-3p/miR-129-5p-GZMA, and TTK_hsa_circ_0077158- miR-212-3p/miR-132-3p/miR-129-5p-TTK might be potential RNA regulatory pathways to regulate the disease progression of early RA.

Conclusions This work identified three haematologic/immune system-specific expressed genes, namely, GZMA, PRC1, and TTK, as potential biomarkers for the early diagnosis and treatment of RA and provided insight into the mechanisms of disease development in RA at the transcriptome level. In addition, we proposed that NEAT1-miR-212-3p/miR-132-3p/miR-129-5p-TTK, XIST-miR-25-3p/miR-129-5p-GZMA, and TTK_hsa_circ_0077158-miR-212-3p/miR-132-3p/miR-129-5p-TTK are potential RNA regulatory pathways that control disease progression in early RA.

Background

Rheumatoid arthritis (RA) is a common chronic autoimmune connective tissue disease that mainly involves the joints. The incidence of RA is 5 to 10 per 1000 people[1]. With the progression of the disease and the continuation of synovial inflammation, the involved joint tissue is gradually eroded. Eventually, RA leads to irreversible damage to the joint, which is a very large burden on individuals and society.

However, early diagnosis and treatment of RA can effectively prevent disease progression, joint damage, and other complications in 90% of patients[2]. Therefore, the earlier a patient with RA is diagnosed, the less burden will be placed on the patient and society. At present, serum biomarkers used in the diagnosis of established RA are rheumatoid factor and anti-cyclic citrullinated peptide antibody[3]. However, early RA especially serum RF and anti-CCP antibody-negative is difficult to diagnose due to the lack of effective biomarkers. Although studies have reported that some biomarkers, such as 14-3-3η autoantibodies and calprotectin, may be effective in the diagnosis of early RA, they have not been used in its clinical diagnosis[4–7]. Therefore, it is vital to identify new and effective biomarkers for the early diagnosis and treatment of RA.

Currently, transcriptomic and microarray analyses have been widely used in a variety of diseases, including RA, to identify new biomarkers to improve diagnosis and treatment[8–10]. In addition, competitive endogenous RNA (ceRNA) networks can elucidate a new mechanism for promoting the development of the disease in a transcriptional regulatory network[11]. Through the combination of microarray and bioinformatics analyses, it is possible to explore potential key genes and pathway networks that are closely related to the development of diseases.

In the present study, we first downloaded microarray datasets for synovial tissue in RA or osteoarthritis (OA) from the Gene Expression Omnibus (GEO) database. After pre-processing and normalizing the data, we identified differentially expressed genes (DEGs) based on the screening criteria and obtained the tissue/organ-specific expressed genes by the online tool BioGPS. Next, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed by the Gene Set Enrichment Analysis (GSEA) software, R software clusterProfiler package, and online tool KEGG Orthology-Based Annotation System (KOBAS) 3.0. A protein-protein interaction (PPI) network was constructed using the online tool STRING, and Cytoscape was used to identify cluster modules and hub genes related to RA. Then, target microRNAs (miRNAs) of selected hub genes were predicted by five online miRNA databases, and a co-expressed network was constructed with Cytoscape. Subsequently, we validated the selected hub genes using GEO datasets, and ceRNA networks were constructed based on prediction results of long noncoding RNAs (lncRNAs) and circular RNAs (circRNAs). This work provides insight into the mechanisms of disease development in RA at the transcriptome level and explores potential biomarkers for the early diagnosis and treatment of RA.

Methods

Microarray data acquisition

The GEO database was used to obtain microarray data for synovial tissue in RA or OA. Screening criteria included the following: 1) Homo sapiens Expression Profiling by array; 2) synovial tissue from RA or OA; 3) datasets contain more than five samples, and 4) datasets contain complete information about the samples. Finally, two GPL570 datasets GSE77298 and GSE82107, which included 16 RA samples and 10 OA samples, were selected to analyse the DEGs; three GPL96 datasets GSE55584, GSE55457, and

GSE55235, which included 33 RA samples and 26 OA samples, and the GPL11154 GSE89408 dataset, which included 57 early RA samples, 95 established RA samples and 22 OA samples, were selected to verify the hub genes (Table 1).

Table 1
Information for selected microarray datasets

GEO accession	Platform	Samples		Attribute
		OA	RA	
GSE77298	GPL570	0	16	Identify DEGs
GSE82107	GPL570	10	0	
GSE55584	GPL96	6	10	Verify hub genes
GSE55457	GPL96	10	13	
GSE55235	GPL96	10	10	
GSE89408	GPL11154	22	152 (57 early and 95 established)	

Annotation: GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array; GPL96: [HG-U133A] Affymetrix Human Genome U133A Array; GPL11154: Illumina HiSeq 2000 (Homo sapiens); DEGs: differentially expressed genes.

Data Normalization And Identification Of DEGs

The original files that were downloaded from the GEO database were preprocessed by the Robust Multiarray Average (RMA) method based on the R software (version 4.0.1) affy package. The limma package was used to identify DEGs. The screening criteria were log2 (fold change) > 1 or < -1 and adjusted p value (Q value) < 0.05.

Heatmap And Volcano Plot Analyses

To better visualize these DEGs, R software was used to make heatmaps and volcano plots. Heatmaps were made with the pheatmap package.

Identification Of Tissue/organ-specific Expressed Genes

To understand the tissue/organ-specific expression of these DEGs, the online tool BioGPS (http://biogps.org/) was used to analyse the tissue distribution[12]. The screening criteria were as follows: 1) transcripts that mapped to a single organ system with an expression value of > 10 multiples of the median, and 2) second-most-abundant tissue's expression was no more than a third as high[13]. The genes obtained by these criteria were considered to be tissue-specific genes.

Enrichment Analysis

GSEA was used to assess the distribution trend of the genes of a predefined set in the gene table to determine their contribution to the phenotype[14]. We downloaded GSEA_4.1.0 and c5: GO gene sets (c5.all.v7.1.symbols.gmt) for functional enrichment analyses. The R software clusterProfiler package was used to analyse the GO enrichment of DEGs, and a chord plot was created for the visualization of these enrichment results. KEGG Orthology-Based Annotation System (KOBAS) 3.0 is an online database widely used for gene/protein function annotation and pathway enrichment (http://kobas.cbi.pku.edu.cn/kobas3) [15]. KOBAS 3.0 was used for the KEGG pathway and Reactome enrichment analyses of DEGs.

Construction Of The PPI Network

The PPI network was constructed based on all DEGs by the online tool STRING (https://string-db.org/) with a filter condition (combined score > 0.4). Next, we downloaded the interaction information and optimized the PPI network with Cytoscape software (v3.8.0) for better visualization. Minimal Common Oncology Data Elements (MCODE) was used to identify significant gene clusters and obtain cluster scores (filter criteria: degree cut-off = 2; node score cut-off = 0.2; k-core = 2; max depth = 100). CytoHubba was used to identify significant genes in this network as hub genes. We used five algorithms, namely Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC), Degree, and Clustering Coefficient, to calculate the top 30 hub genes. Finally, all the results were intersected to obtain the final hub genes.

Prediction Of Target MiRNAs

We used five miRNA databases, namely, RNA22, DIANA-micro T, miRWalk, miRDB, and miRcode, to predict target miRNAs of hub genes and selected miRNAs that were found in at least four databases as the target miRNAs. The messenger RNA (mRNA)-miRNA co-expressed network based on the relationship between mRNAs and miRNAs was constructed by using Cytoscape.

Construction Of CeRNA Networks

StarBase (version 3.0) (http://starbase.sysu.edu.cn/index.php) was used to predict lncRNAs and circRNAs that interacted with the selected miRNAs[16]. The intersections of the predicted results were used as the target lncRNAs and circRNAs. CeRNA networks based on the interactions among mRNAs, miRNAs, and noncoding RNAs (ncRNAs) were constructed by using Cytoscape.

Statistics Analysis

The R software ggpubr package was used to perform statistical analyses, and the ggplot2 package was used to draw boxplots. Student’s t-test was used to compare the differences between the two groups. IBM SPSS Statistics 25 (SPSS, Inc., Chicago, IL, USA) was used to analyse the data and draw the receiver operating characteristic (ROC) curve.

Results

1. Identification of DEGs

Compared with genes in the OA samples, we identified a total of 275 DEGs in the RA samples, which comprised 197 downregulated genes and 78 upregulated genes. Next, heatmap and volcano plot analyses were used to visualize these DEGs, which are shown in Fig. 1A and 1B.

2. Identification of the tissue/organ-specific expressed genes

A total of 71 tissue/organ-specific expressed genes were identified by BioGPS (Table 2). We observed that most of these genes were specifically expressed in the haematologic/immune system (35/71, 49.29%). The second organ-specific expressed system was the nervous system, which included 13 genes (13/71, 18.31%). This was followed by the digestive system (7/71, 9.86%), respiratory system (4/71, 5.63%), circulatory system (4/71, 5.63%), and placenta (3/71, 4.22%). Finally, the endocrine system, genital system, and tongue, prostate, and adipose tissues had the lowest specific expressed genes (1/71, 1.41%).

Table 2
Distribution of tissue/o-specific expressed genes identified by BioGPS

System/Organ		Genes	Counts
Haematologic/Immune	Haematologic/Immune cells	PLA2G7, SLC50A1, T, MSC, MATK, PRKCD, CCR7, CYB561A3, P2RY8, CD3G, EMR2, NOV, BCL2A1, CD52, CD27, IL7R, TTK, MAP3K7CL, PNOC, FCGR1B, GZMB, GZMA, DLGAP5, TRBC1, MYOM2, CORO1A, PRC1, CEP55, CD3D, IER2, ITK, TNFRSF17	32
	Immune organs	CXCL13, LCK, CD163	3
Nervous		PALM, KCND2, RASL10A, DACH1, STXBP1, DNM1, IL17D, PLP1, WRB, RCAN2, ZNF423, LRRN4CL, LPHN3	13
Digestive		GIPC2, AKR1B10, IGJ, ADAMDEC1, C6, TOX3, C15orf48	7
Respiratory		CHAD, MFAP4, CLDN5, LAMP3	4
Circulatory		ACTC1, CASQ2, LRRC2, CKMT2	4
Placenta		PVRL3, RHOBTB1, AGTR1	3
Endocrine		DUOX2	1
Genital		MLF1	1
Others	Tongue	MAL	1
	Prostate	PPAP2A	1
	Adipose	HOXC10	1

3. Enrichment Analysis

The GSEA software, R software clusterProfiler package, and online tool KOBAS 3.0 were used for functional and pathway enrichment analyses. First, we uploaded the expression profile of all genes in the RA and OA samples to GSEA, and the c5: GO gene set was used to perform the GO enrichment analysis of the expression profile at an overall level. The screening criterion for significant gene sets was $p < 0.05$. We observed that most of the enriched gene sets were related to the innate immune cell-mediated immune response, immune-related biological processes, and pathways (Fig. 2).

Next, the R software clusterProfiler package and KOBAS 3.0 were used to perform GO, KEGG pathway, and Reactome enrichment analyses of DEGs, respectively. GO enrichment analysis of DEGs also revealed that the immune response in RA samples was stronger than that in OA samples, and this included the regulation of humoral immune response, complement activation, leukocyte activation, and migration. The top 10 biological processes were selected based on a Q value < 0.05 and were drawn in a chord plot (Fig. 3A). KEGG pathway enrichment analysis showed that DEGs were enriched in cytokine-cytokine receptor interaction, primary immunodeficiency, JAK-STAT signalling pathway, Fc gamma R-mediated phagocytosis, and neuroactive ligand-receptor interaction. Reactome enrichment analysis showed that DEGs were mostly enriched in the immune system and signal transduction. According to a Q value < 0.05, we selected the top five KEGG pathways and the top five Reactome terms and showed them in a bubble plot (Fig. 3B).

4. PPI network analysis, MCODE cluster modules and hub gene identification

The interaction network between proteins coded by DEGs, which comprised 187 nodes and 307 edges, was constructed by STRING and visualized by Cytoscape (Fig. 4A). The MCODE plugin was used to identify gene cluster modules. In this network, we identified four modules, which are shown in Fig. 4B-4E, according to the filter criteria. Cluster 1 had the highest cluster score (score: 9, 9 nodes and 36 edges), followed by cluster 2 (score: 5.167, 13 nodes, 31 edges), cluster 3 (score: 3.333, 4 nodes, 5 edges), and cluster 4 (score: 2.8, 6 nodes, 7 edges). Next, we used the cytoHubba plugin to identify hub genes. Fifteen hub genes were identified by intersecting the results from the five algorithms. These hub genes with detailed information are shown in Table 3. Because GO and KEGG enrichment analyses showed that DEGs were mainly enriched in immune-related biological processes and pathways, we intersected 15 hub genes and genes specifically expressed in the haematologic/immune system. Ultimately, we obtained eight haematologic/immune system-specific expressed hub genes, including CD52, CD27, threonine/tyrosine kinase protein kinase (TTK), granzyme A (GZMA), DLGAP5, protein regulator of cytokinesis 1 (PRC1), CEP55, and CXCL13 (Table 3, in bold).

Table 3
15 hub genes identified by five algorithms of cytoHubba

Gene symbol	Description	log2FC	Q value	Regulation
CXCL13	C-X-C motif chemokine ligand 13	2.846	0.012	up
CD52	CD52 molecule	1.928	0.004	up
GZMA	granzyme A	1.753	0.022	up
CD27	CD27 molecule	1.419	0.004	up
CEP55	centrosomal protein 55	1.264	0.01	up
SKA3	spindle and kinetochore associated complex subunit 3	1.215	3.03E-07	up
IL21R	interleukin 21 receptor	1.208	7E-04	up
DLGAP5	DLG associated protein 5	1.172	0.027	up
PRC1	protein regulator of cytokinesis 1	1.112	0.042	up
EOMES	eomesodermi	1.108	0.031	up
TTK	TTK protein kinas	1.065	0.006	up
CDCA8	cell division cycle associated 8	1.049	1.74E-08	up
UHRF1	ubiquitin-like with PHD and ring finger domains 1	1.048	0.016	up
KIAA0101	PCNA clamp associated factor	1.017	0.013	up
FNBP1L	formin binding protein 1 like	-1.325	0.004	down

Annotation: FC: fold change, Q value: adjust P-value. Eight hematologic/immune system-specific expressed hub genes shown in bold

5. Prediction of target miRNAs and construction of the co-expressed network

We obtained 95 target miRNAs of 8 specifically expressed hub genes by predicting five miRNA databases and determined 105 mRNA-miRNA pairs. According to the prediction results, a coexpressed network of mRNAs and miRNAs, which comprised 103 nodes and 105 edges, was constructed by Cytoscape (Fig. 5).

6. Verification of the 8 specifically expressed hub genes by 4 datasets from the GEO database

Three GPL96 datasets, namely, GSE55584, GSE55457 and GSE55235, and the GPL11154 GSE89408 dataset were selected to verify the 8 specifically expressed hub genes. The R software ggplot2 package and ggpubr package were used to draw boxplots and perform Student's t-test statistical analyses.

Consistent with our predictions, the mRNA expression levels of the 8 specifically expressed hub genes in the RA samples were significantly increased compared with those in the OA samples ($p < 0.01$) (Fig. 6A and 6B). In addition, we observed that the mRNA expression levels of GZMA, PRC1, and TTK in the 57 early RA samples were significantly increased compared with those in the 95 established RA samples ($p < 0.05$) (Fig. 6B).

7. ROC curve of the 8 specifically expressed hub genes in early RA samples and established RA samples

We used IBM SPSS Statistics 25 to analyse the 8 specifically expressed hub gene expression profiles of OA samples, early RA samples, and established RA samples and draw receiver operating characteristic (ROC) curves. The difference values of the area under the ROC curve (AUC) between early RA samples and established RA samples were greater than 0.5, suggesting that these genes have a higher diagnostic value as biomarkers. Compared to the diagnostic value in established RA samples, CD27 (AUC: 0.872 vs 0.817), CEP55 (AUC: 0.805 vs 0.731), GZMA (AUC: 0.906 vs 0.852), PRC1 (AUC: 0.809 vs 0.703), and TTK (AUC: 0.793 vs 0.726) have higher diagnostic value in early RA samples (Fig. 7). Moreover, CXCL13 has a very high diagnostic value in both early RA samples (AUC: 0.893) and established RA samples (AUC: 0.9). DLGAP5 (AUC: 0.810 and 0.786) and CD52 (AUC: 0.863 and 0.837) also have a high diagnostic value in both groups. According to the verified results, we considered GZMA, PRC1, and TTK as biomarkers for the diagnosis of early RA.

8. Prediction of target ncRNAs and construction of ceRNA networks

miRNAs are well known to induce gene silencing and downregulate gene expression by binding mRNAs. However, its upstream molecules, circRNAs, and lncRNAs, can affect the function of miRNA by combining miRNA response elements, thus upregulating gene expression. This interaction between RNAs is called a ceRNA network[11]. Next, we used the online database Starbase 3.0 to predict the lncRNAs and circRNAs that interact with the selected miRNAs. The screening criteria were: mammalian, human h19 genome, strict stringency (> = 5) of CLIP-Data, and with or without degradome data. We chose the ncRNAs that exist in most of the prediction results of miRNAs as our predicted lncRNAs and circRNAs. In addition, since a transcript has multiple circRNA shear sites in the prediction results of the Starbase database, we selected the circRNA with the most samples and highest score in the circBase database as the target circRNA. Finally, we obtained 3 target lncRNAs and 4 target circRNAs of the target miRNAs of PRC1; 1 target lncRNA and 19 target circRNAs of the target miRNAs of GZMA; and 1 target lncRNA and 14 target circRNAs of the target miRNAs of TTK. Three ceRNA networks based on the prediction results were constructed and illustrated by Cytoscape and are shown in Fig. 8A-8C. Subsequently, according to the ceRNA hypothesis, we conducted a literature search and selected four reported downregulated miRNAs and an upregulated lncRNA in RA and upregulated lncRNA in another autoimmune disease, Sjogren's syndrome, for further analysis. We propose that NEAT1-miR-212-3p/miR-132-3p/miR-129-5p-TTK (Fig. 8D) and XIST-miR-25-3p/miR-129-5p-GZMA (Fig. 8E) might be potential RNA regulatory pathways to regulate the disease progression of early RA. Regarding the prediction results of circRNAs, we found a circRNA (TTK_hsa_circ_0077158) predicted by target miRNAs of TTK, and its target is TTK. Hence, we

propose the following circRNA-miRNA-mRNA pathway: TTK_hsa_circ_0077158-miR-212-3p/miR-132-3p/miR-129-5p-TTK (Fig. 8F); it might be a key regulatory pathway in the pathogenesis of early RA.

Discussion

The main characteristic of RA is chronic synovial inflammation, which leads to erosion and damage of joints. Early diagnosis and treatment of RA will effectively prevent joint damage and improve quality of life. However, early RA is difficult to diagnose due to the lack of effective biomarkers. It is crucial to identify new and effective biomarkers for the early diagnosis and treatment of RA.

In our study, we identified 275 DEGs, including 71 tissue/organ-specific expressed genes, by comparing genes expressed in RA and OA samples. GO enrichment analysis of all genes and DEGs indicated that the immune responses, such as the immune cell-mediated immune response and the regulation of humoral immune response, were stronger in RA samples than in OA samples. KEGG pathways that were enriched included cytokine-cytokine receptor interaction, primary immunodeficiency, JAK-STAT signalling pathway, Fc gamma R-mediated phagocytosis, and neuroactive ligand-receptor interaction. Reactome enrichment analysis also showed that DEGs were mostly enriched in the immune system and signal transduction. We observed that the results of these enrichment analyses were consistent with actual differences between RA and OA.

After the hub genes that were screened by the PPI network were validated using the GEO datasets, we identified three haematologic/immune system-specific expressed genes, namely, GZMA, PRC1, and TTK, as biomarkers for the diagnosis of early RA. In addition, we constructed an mRNA-miRNA co-expression network and ceRNA networks to elucidate the pathogenesis of RA at the transcriptome level.

GZMA, a member of the serine protease family, is secreted by cytotoxic cells such as cytotoxic T cells and natural killer (NK) cells and plays an important role in cell death, cytokine processing, and inflammation[17, 18]. Several studies have reported that compared with the expression level of GZMA in OA patients, the expression level of GZMA increases in plasma, synovial tissues, and synovial membranes in patients with RA[19, 20]. This indicates that GZMA plays a significant role in the pathogenesis of RA. Consistent with this research, our study found that GZMA was upregulated in the synovial membrane of RA, especially in early RA. In addition, the ROC curve of GZMA indicated that it has a very high diagnostic value in early RA (AUC = 0.906). We considered GZMA a very effective biomarker for the diagnosis of early RA.

PRC1 (also called ASE1), a human mitotic spindle-associated CDK substrate protein, is a key regulator of cell division[21]. According to BioGPS, PRC1 is specifically expressed in early erythrocytes, endotheliocytes, and B lymphocytes. At present, PRC1 has not been reported in RA-related studies. However, in our study, PRC1 was upregulated in the synovial membrane of RA, especially in early RA. Compared with OA, synovial inflammation and hyperplasia are obvious in RA. In addition, it has been reported that the metabolic level of the synovial membrane is elevated, similar to that of tumour tissue[22]. These results all reflect the increased proliferation of cells in the synovial membrane of RA to

some extent. Therefore, PRC1 may play an important role in the proliferation of synovial cells and the disease progression of RA.

TTK (also called MPS1 and CT96), which encodes a dual specificity protein kinase that phosphorylates a variety of amino acids such as tyrosine and serine, is related to cell proliferation[23]. Similar to PRC1, TTK is also highly specifically expressed in early red blood cells and endothelial cells. A study by H Ah-Kim *et al.* reported that tumour necrosis factor-alpha (TNF-α) can increase TTK expression in human articular chondrocytes[24], suggesting that TTK is regulated by TNF-α in some biological processes. We know that TNF-α plays a very significant role in the pathogenesis of RA[25].

Thus, we hypothesized that TTK plays an important role in synovial cell proliferation and TNF-α-mediated pathogenesis. In addition, we identified that TTK was highly expressed in the synovial membrane of RA and has a high diagnostic value in early RA (AUC = 0.793). We considered TTK as a novel and effective biomarker for the diagnosis of early RA.

Furthermore, target miRNAs and the target lncRNAs and circRNAs of these miRNAs were predicted for GZMA, PRC1, and TTK, and a ceRNA network was constructed with Cytoscape.

This network reveals the mechanism by which selected genes are regulated at the transcriptome level. According to the ceRNA hypothesis, we performed a literature search to select downregulated miRNAs in RA for further analysis. Among the target miRNAs of GZMA, PRC1, and TTK, the expression of the following miRNAs was downregulated in RA: miR-129-5p (in RA synovial tissue and synovial fibroblasts), miR-132-3p (in RA synovial fibroblasts), miR-212-3p (in RA synovial tissue and synovial fibroblasts), and miR-25-3p (in peripheral blood mononuclear cells)[26–29]. In addition, it has been reported that the lncRNA NEAT1 is upregulated in peripheral blood mononuclear cells of patients with RA[30]. Therefore, we propose that NEAT1-miR-212-3p/miR-132-3p/miR-129-5p-TTK might be a potential RNA regulatory pathway to regulate the disease progression of early RA. Additionally, although lncRNA XIST has not been reported in RA, it has been reported to be upregulated in another autoimmune disease, Sjogren's syndrome[31]. We hypothesize that XIST-miR-25-3p/miR-129-5p-GZMA has an important regulatory role in RA. Regarding the prediction results of circRNAs, we found a circRNA (TTK_hsa_circ_0077158) predicted by target miRNAs of TTK, and its target was TTK. Hence, we proposed a circRNA-miRNA-mRNA pathway: TTK_hsa_circ_0077158-miR-212-3p/miR-132-3p/miR-129-5p-TTK; it might be a key regulatory pathway in the pathogenesis of early RA. Of course, these RNA regulatory pathways need to be further experimentally verified, which is also a limitation of our study.

Conclusions

Our work identified three haematologic/immune system-specific expressed genes, GZMA, PRC1, and TTK, as potential biomarkers for the early diagnosis and treatment of RA and provided insight into the mechanisms of disease development in RA at the transcriptome level. In addition, we propose that NEAT1-miR-212-3p/miR-132-3p/miR-129-5p-TTK, XIST-miR-25-3p/miR-129-5p-GZMA, and

TTK_hsa_circ_0077158- miR-212-3p/miR-132-3p/miR-129-5p-TTK are potential RNA regulatory pathways that control disease progression in early RA.

List Of Abbreviations

Rheumatoid arthritis: RA; osteoarthritis: OA; Gene Expression Omnibus: GEO; differentially expressed genes: DEGs; Gene Ontology: GO; Kyoto Encyclopedia of Genes and Genomes: KEGG; competitive endogenous RNA: ceRNA; Gene Set Enrichment Analysis: GSEA; Robust Multiarray Average: RMA; KEGG Orthology-Based Annotation System: KOBAS; protein-protein interaction: PPI; receiver operating characteristic: ROC; area under the ROC curve: AUC; noncoding RNAs: ncRNAs.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The [GSE datasets] data that support the findings of this study are available in the GEO database (https://www.ncbi.nlm.nih.gov/geo/) with the following data accession identifier(s): GSE77298, GSE82107, GSE55584, GSE55457, GSE55235, and GSE89408.

Competing interests: The authors declare no conflicts of interest regarding this work. The corresponding authors have the right to and do speak on behalf of all authors.

Funding: The National Natural Science Foundation of China (No. 81501388) and the National Natural Science Foundation of Zhejiang Province (Nos. LY20H100007 and LQ17H160007).

Authors' contributions: Conception and design of the work, Qi Cheng and Yan Du; acquisition and analysis of data, Qi Cheng; interpretation of data, Qi Cheng, Yan Du, and Huaxiang Wu; writing and preparing the original draft, Qi Cheng; writing, reviewing and editing the paper, Qi Cheng and Yan Du; supervision, Yan Du and Huaxiang Wu; project administration, Yan Du and Huaxiang Wu; and funding acquisition, Yan Du. All authors have read and agreed to the published version of the manuscript and to have agreed to both be personally accountable for the author's contributions and ensure to answer any questions related to the accuracy or integrity of any part of the work.

Acknowledgements: Not applicable.

References

1. Smolen JS, Aletaha D, McInnes IB: **Rheumatoid arthritis.** *Lancet* 2016, **388:**2023-2038.
2. Aletaha D, Smolen JS: **Diagnosis and Management of Rheumatoid Arthritis: A Review.** *Jama* 2018, **320:**1360-1372.

3. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, 3rd, Birnbaum NS, Burmester GR, Bykerk VP, Cohen MD, et al: **2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative.** *Arthritis Rheum* 2010, **62:**2569-2581.
4. Maksymowych WP, Boire G, van Schaardenburg D, Wichuk S, Turk S, Boers M, Siminovitch KA, Bykerk V, Keystone E, Tak PP, et al: **14-3-3η Autoantibodies: Diagnostic Use in Early Rheumatoid Arthritis.** *J Rheumatol* 2015, **42:**1587-1594.
5. Jonsson MK, Sundlisæter NP, Nordal HH, Hammer HB, Aga AB, Olsen IC, Brokstad KA, van der Heijde D, Kvien TK, Fevang BS, et al: **Calprotectin as a marker of inflammation in patients with early rheumatoid arthritis.** *Ann Rheum Dis* 2017, **76:**2031-2037.
6. De Winter LM, Hansen WL, van Steenbergen HW, Geusens P, Lenaerts J, Somers K, Stinissen P, van der Helm-van Mil AH, Somers V: **Autoantibodies to two novel peptides in seronegative and early rheumatoid arthritis.** *Rheumatology (Oxford)* 2016, **55:**1431-1436.
7. Dunaeva M, Blom J, Thurlings R, Pruijn GJM: **Circulating serum miR-223-3p and miR-16-5p as possible biomarkers of early rheumatoid arthritis.** *Clin Exp Immunol* 2018, **193:**376-385.
8. Carr HL, Turner JD, Major T, Scheel-Toellner D, Filer A: **New Developments in Transcriptomic Analysis of Synovial Tissue.** *Front Med (Lausanne)* 2020, **7:**21.
9. Li WC, Bai L, Xu Y, Chen H, Ma R, Hou WB, Xu RJ: **Identification of differentially expressed genes in synovial tissue of rheumatoid arthritis and osteoarthritis in patients.** *J Cell Biochem* 2019, **120:**4533-4544.
10. Macías-Segura N, Castañeda-Delgado JE, Bastian Y, Santiago-Algarra D, Castillo-Ortiz JD, Alemán-Navarro AL, Jaime-Sánchez E, Gomez-Moreno M, Saucedo-Toral CA, Lara-Ramírez EE, et al: **Transcriptional signature associated with early rheumatoid arthritis and healthy individuals at high risk to develop the disease.** *PLoS One* 2018, **13:**e0194205.
11. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP: **A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?** *Cell* 2011, **146:**353-358.
12. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, 3rd, Su AI: **BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.** *Genome Biol* 2009, **10:**R130.
13. Wang H, Zhu H, Zhu W, Xu Y, Wang N, Han B, Song H, Qiao J: **Bioinformatic Analysis Identifies Potential Key Genes in the Pathogenesis of Turner Syndrome.** *Front Endocrinol (Lausanne)* 2020, **11:**104.
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.
15. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L: **KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases.** *Nucleic Acids Res* 2011, **39:**W316-322.

16. Li JH, Liu S, Zhou H, Qu LH, Yang JH: **starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA, and protein-RNA interaction networks from large-scale CLIP-Seq data.** *Nucleic Acids Res* 2014, **42:**D92-97.
17. Anthony DA, Andrews DM, Watt SV, Trapani JA, Smyth MJ: **Functional dissection of the granzyme family: cell death and inflammation.** *Immunol Rev* 2010, **235:**73-92.
18. van Daalen KR, Reijneveld JF, Bovenschen N: **Modulation of Inflammation by Extracellular Granzyme A.** *Front Immunol* 2020, **11:**931.
19. Tak PP, Spaeny-Dekking L, Kraan MC, Breedveld FC, Froelich CJ, Hack CE: **The levels of soluble granzyme A and B are elevated in plasma and synovial fluid of patients with rheumatoid arthritis (RA).** *Clin Exp Immunol* 1999, **116:**366-370.
20. Kummer JA, Tak PP, Brinkman BM, van Tilborg AA, Kamp AM, Verweij CL, Daha MR, Meinders AE, Hack CE, Breedveld FC: **Expression of granzymes A and B in synovial tissue from patients with rheumatoid arthritis and osteoarthritis.** *Clin Immunol Immunopathol* 1994, **73:**88-95.
21. Jiang W, Jimenez G, Wells NJ, Hope TJ, Wahl GM, Hunter T, Fukunaga R: **PRC1: a human mitotic spindle-associated CDK substrate protein required for cytokinesis.** *Mol Cell* 1998, **2:**877-885.
22. Falconer J, Murphy AN, Young SP, Clark AR, Tiziani S, Guma M, Buckley CD: **Review: Synovial Cell Metabolism and Chronic Inflammation in Rheumatoid Arthritis.** *Arthritis Rheumatol* 2018, **70:**984-999.
23. Liu X, Winey M: **The MPS1 family of protein kinases.** *Annu Rev Biochem* 2012, **81:**561-585.
24. Ah-Kim H, Zhang X, Islam S, Sofi JI, Glickberg Y, Malemud CJ, Moskowitz RW, Haqqi TM: **Tumour necrosis factor-alpha enhances the expression of hydroxyl lyase, cytoplasmic antiproteinase-2, and a dual-specificity kinase TTK in human chondrocyte-like cells.** *Cytokine* 2000, **12:**142-150.
25. Brennan FM, Maini RN, Feldmann M: **TNF alpha--a pivotal role in rheumatoid arthritis?** *Br J Rheumatol* 1992, **31:**293-298.
26. Zhang Y, Yan N, Wang X, Chang Y, Wang Y: **MiR-129-5p regulates cell proliferation and apoptosis via IGF-1R/Src/ERK/Egr-1 pathway in RA-fibroblast-like synoviocytes.** *Biosci Rep* 2019, **39**.
27. Liu Y, Zhang XL, Li XF, Tang YC, Zhao X: **miR-212-3p reduced proliferation, and promoted apoptosis of fibroblast-like synoviocytes via down-regulating SOX5 in rheumatoid arthritis.** *Eur Rev Med Pharmacol Sci* 2018, **22:**461-471.
28. Tseng CC, Wu LY, Tsai WC, Ou TT, Wu CC, Sung WY, Kuo PL, Yen JH: **Differential Expression Profiles of the Transcriptome and miRNA Interactome in Synovial Fibroblasts of Rheumatoid Arthritis Revealed by Next Generation Sequencing.** *Diagnostics (Basel)* 2019, **9**.
29. Kurowska W, Kuca-Warnawin E, Radzikowska A, Jakubaszek M, Maślińska M, Kwiatkowska B, Maśliński W: **Monocyte-related biomarkers of rheumatoid arthritis development in undifferentiated arthritis patients - a pilot study.** *Reumatologia* 2018, **56:**10-16.
30. Shui X, Chen S, Lin J, Kong J, Zhou C, Wu J: **Knockdown of lncRNA NEAT1 inhibits Th17/CD4(+) T cell differentiation through reducing the STAT3 protein level.** *J Cell Physiol* 2019, **234:**22477-22484.

31. Mougeot JL, Noll BD, Bahrani Mougeot FK: **Sjögren's syndrome X-chromosome dose effect: An epigenetic perspective.** *Oral Dis* 2019, **25:**372-384.

Figures

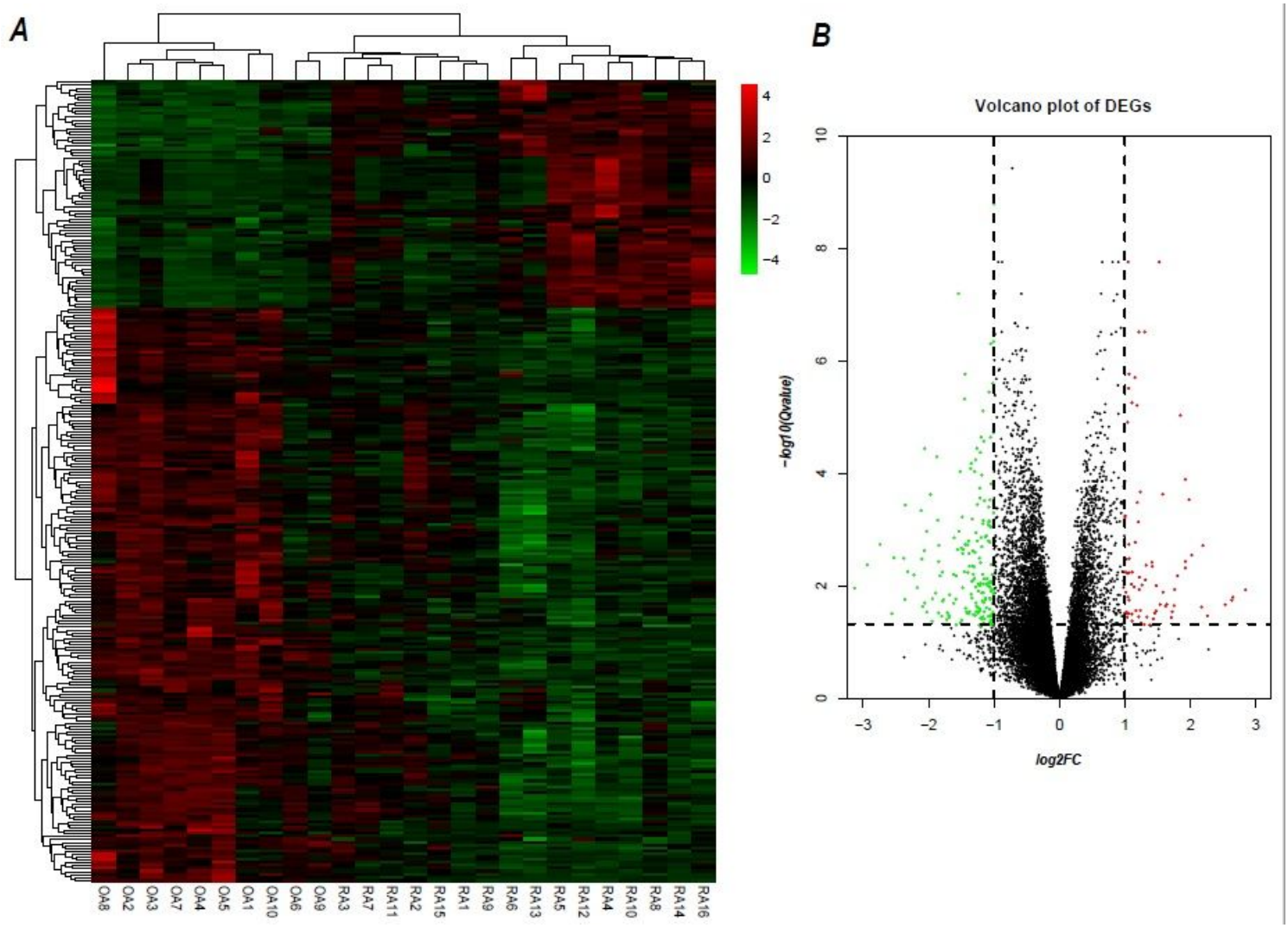

Figure 1

Identification of DEGs. A. Heatmap of DEGs between the RA samples and the OA samples. Red rectangles represent high expression, and green rectangles represent low expression. B. Volcano plot of DEGs between the RA samples and the OA samples. The red plots represent upregulated genes, the black plots represent nonsignificant genes, and the green plots represent downregulated genes.

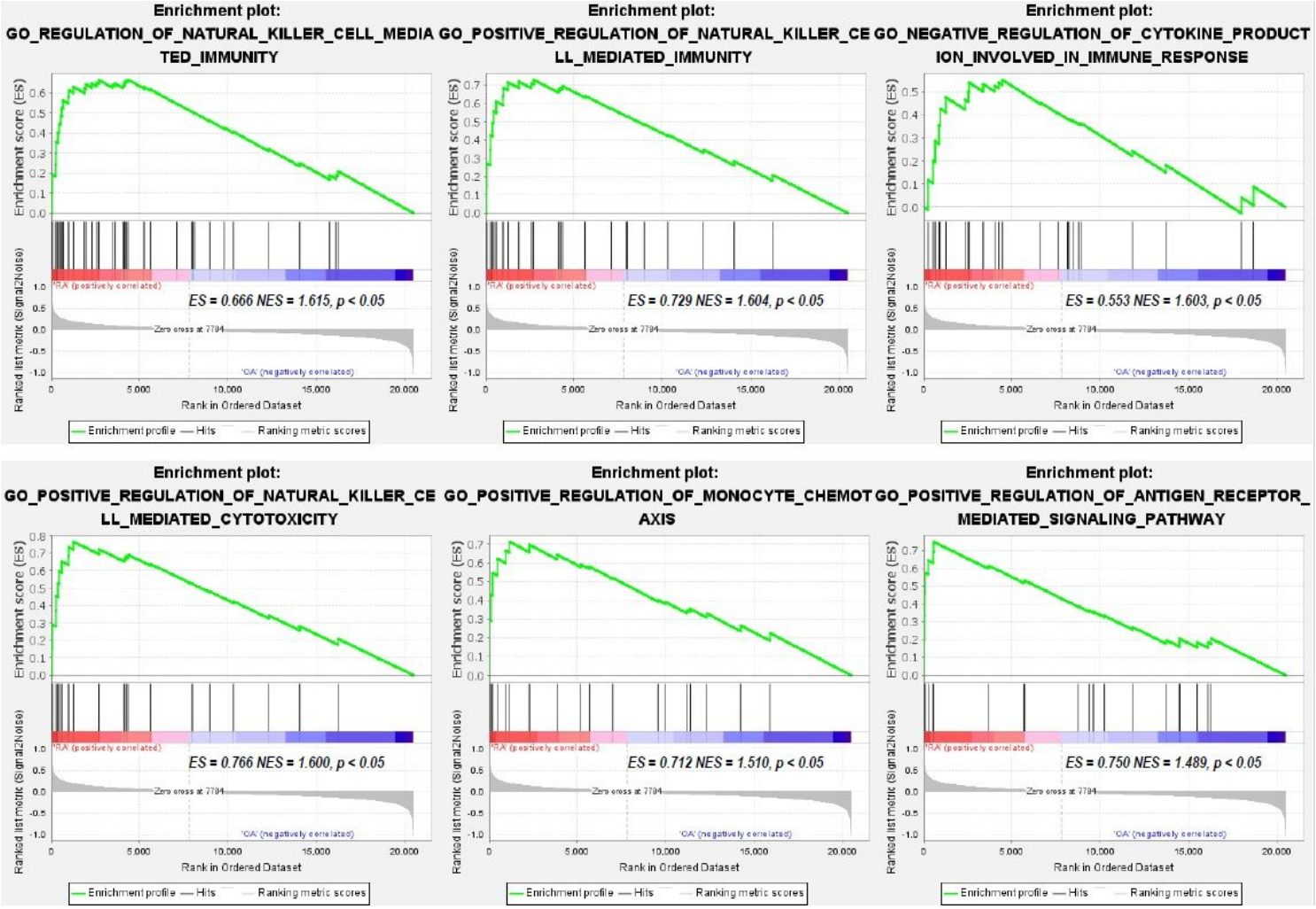

Figure 2

GO enrichment analysis of expression profile in GSEA software at the overall level. The c5: GO gene set was used to perform the GO enrichment analysis of the expression profile at the overall level. The screening criteria for significant gene sets were $p < 0.05$.

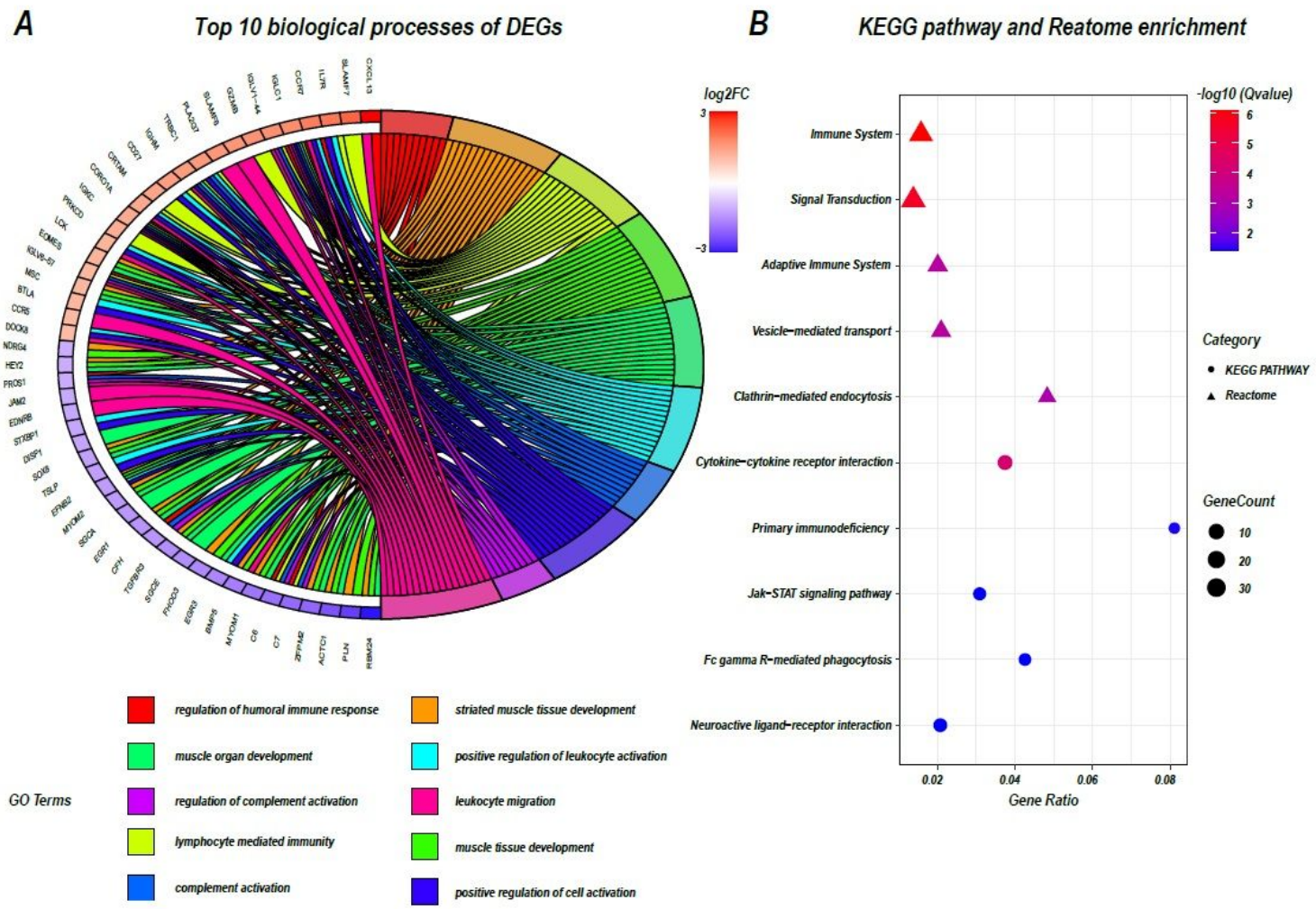

Figure 3

GO, KEGG pathway, and Reactome enrichment analyses of DEGs. A. The top 10 biological processes of DEGs were selected based on a Q value < 0.05 and were drawn in a chord plot. B. The top five KEGG pathways and the top five Reactome terms were selected according to Q < 0.05 and are shown in a bubble plot.

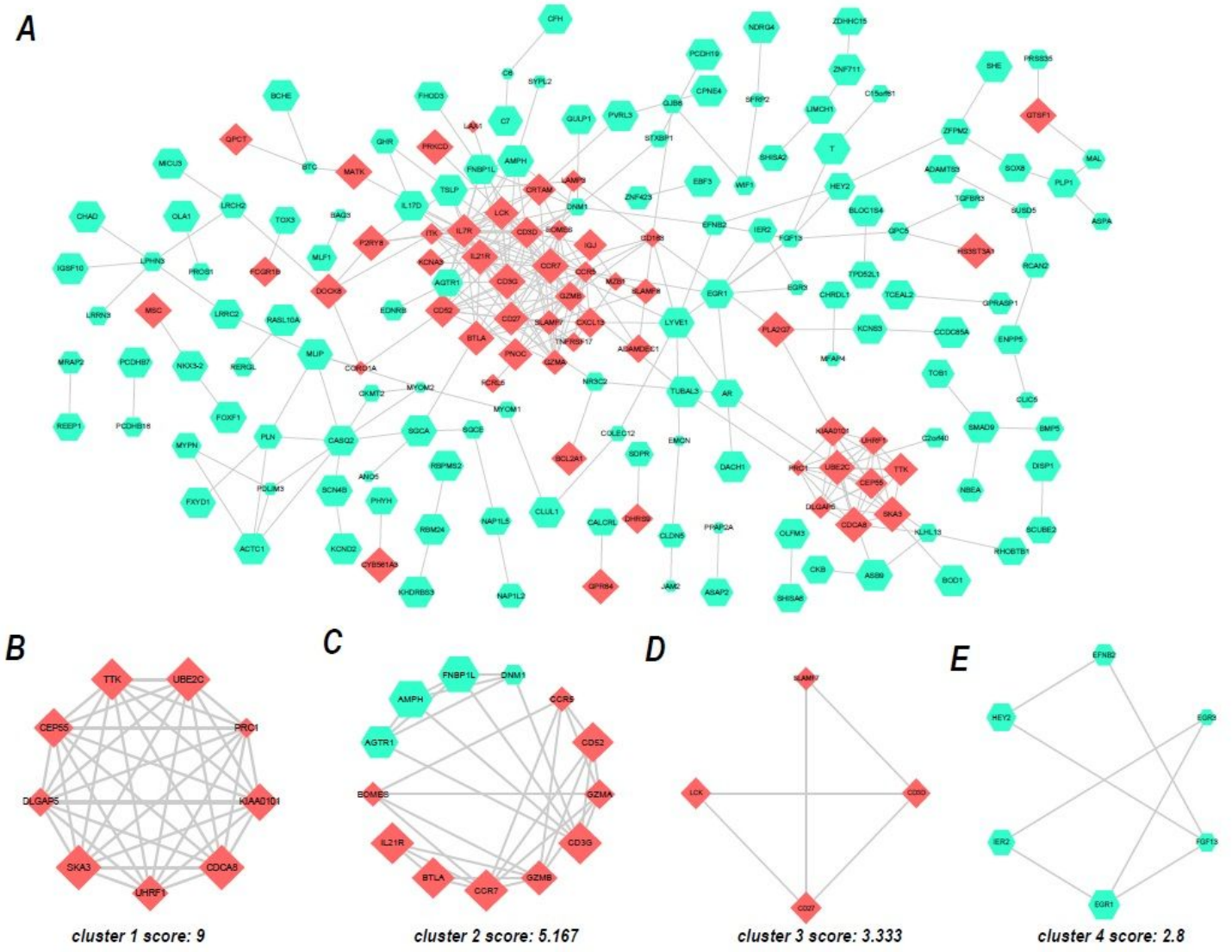

Figure 4

PPI network of DEGs and four cluster modules extracted by MCODE. A. The interaction network between proteins coded by DEGs. Red diamonds represent the upregulated genes, and green hexagons represent the downregulated genes. The smaller the value of Q is, the larger the shape size. B-E. Four cluster modules extracted by MCODE and their scores.

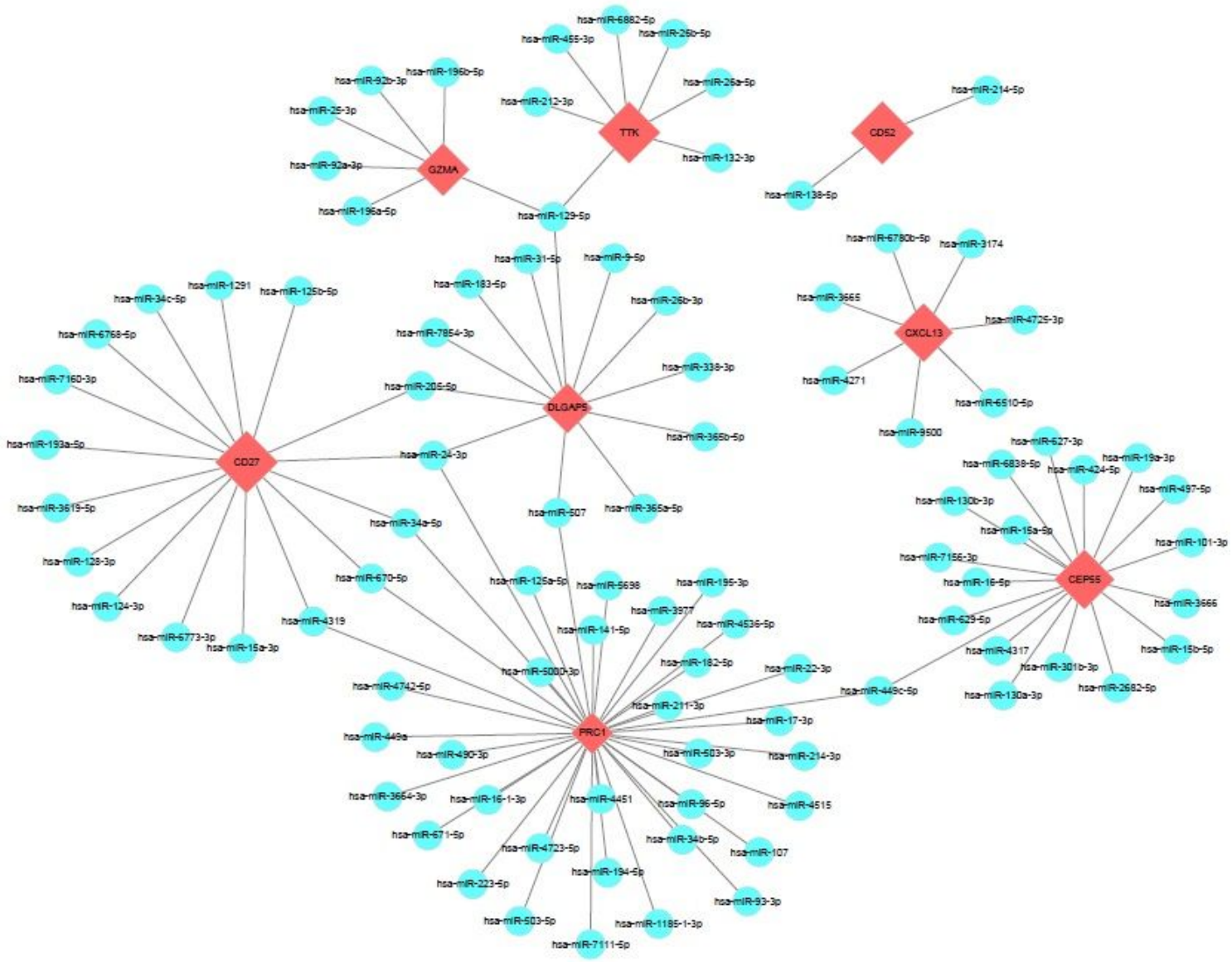

Figure 5

A co-expressed network of mRNAs and target miRNAs. Red diamonds represent the hub genes, and blue circles represent miRNAs.

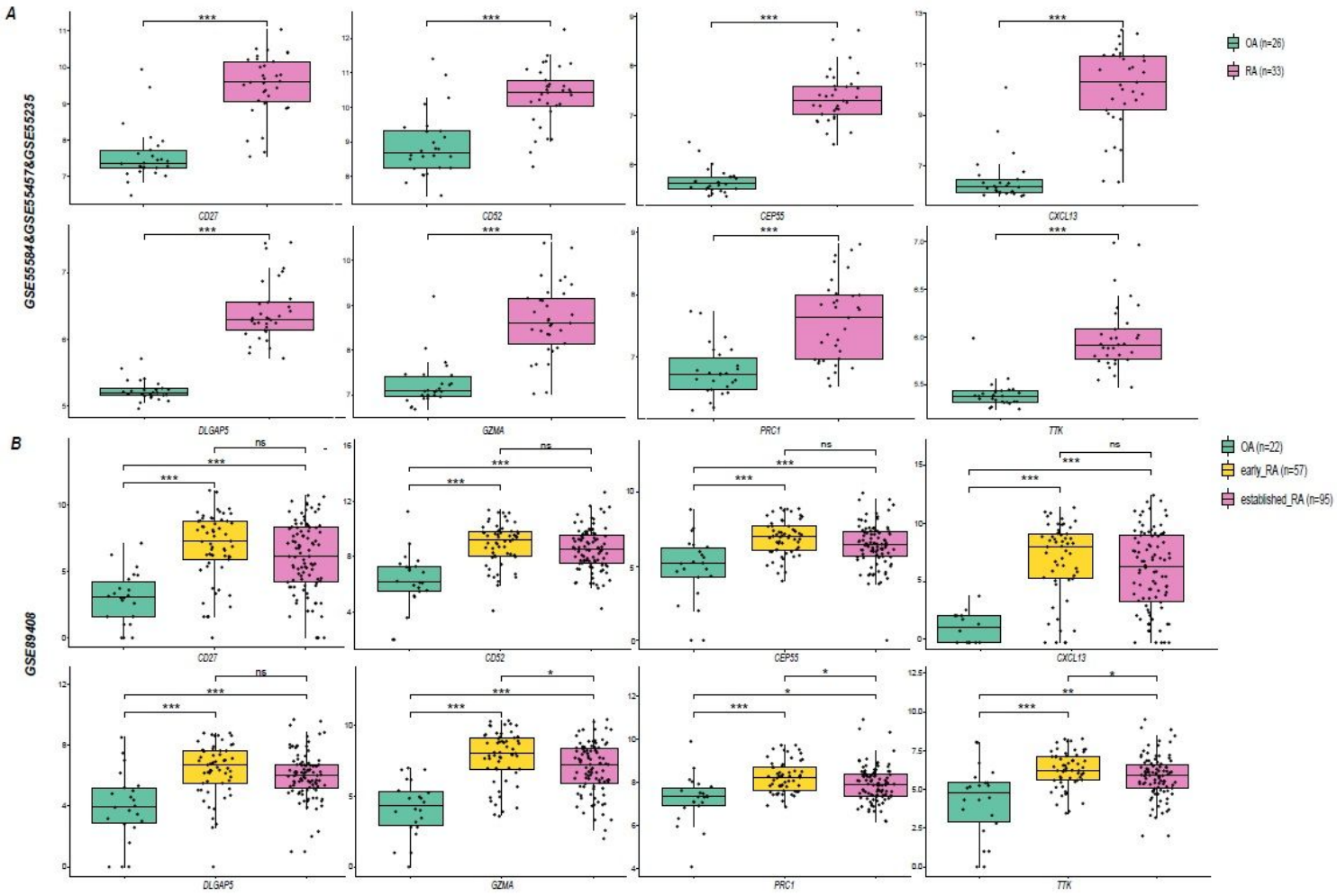

Figure 6

Verification of the 8 specifically expressed hub genes by 4 datasets of the GEO database. A. Verification by three GPL96 datasets: GSE55584, GSE55457 and GSE55235. B. Verification by the GPL11154 GSE89408 dataset. ***: $p < 0.001$, *: $p < 0.05$, ns: no significant difference.

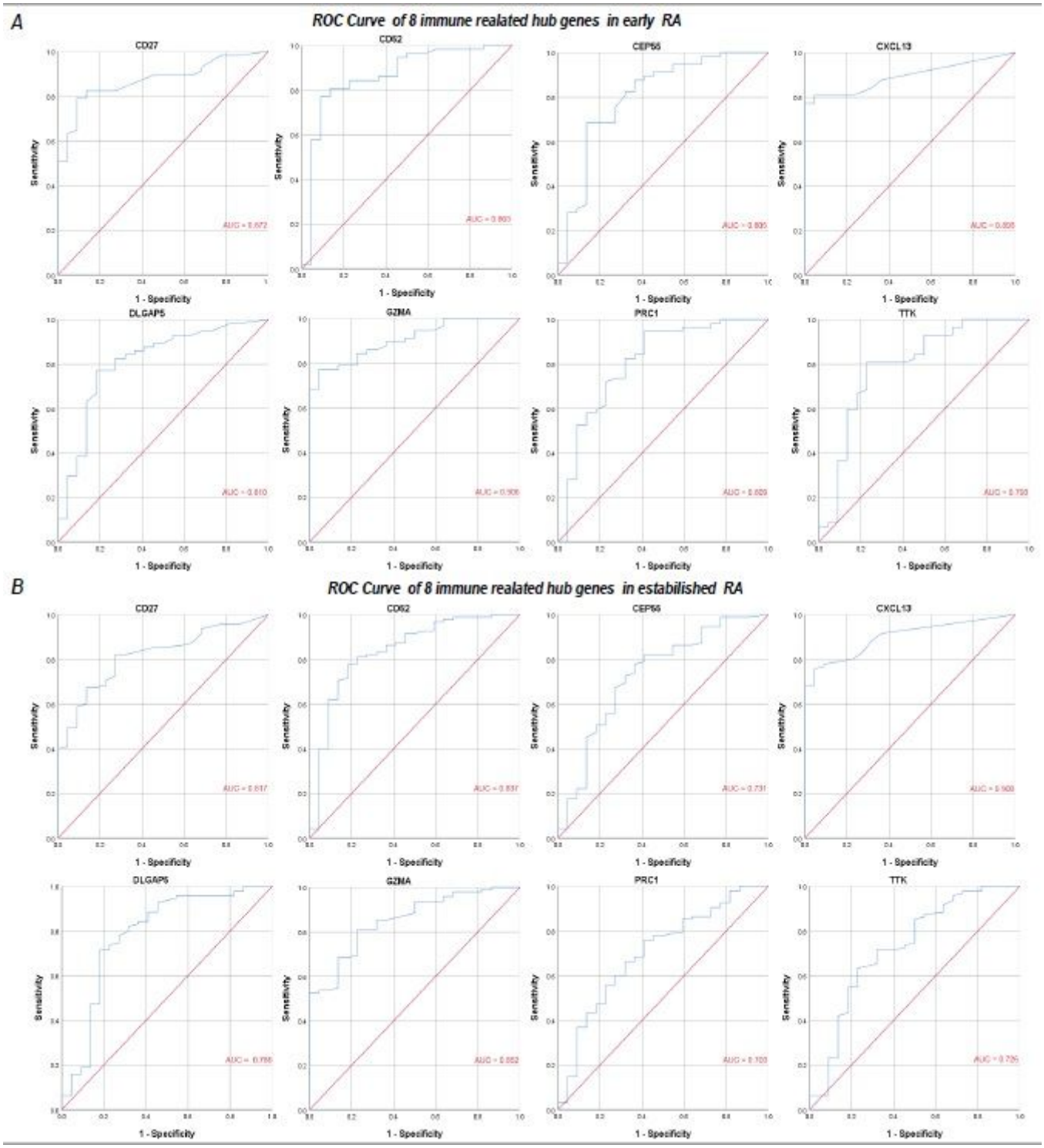

Figure 7

ROC curve of the 8 specifically expressed hub genes. A. ROC curve of the 8 specifically expressed hub genes in early RA samples. B. ROC curve of the 8 specifically expressed hub genes in established RA samples. AUC: area under the ROC curve.

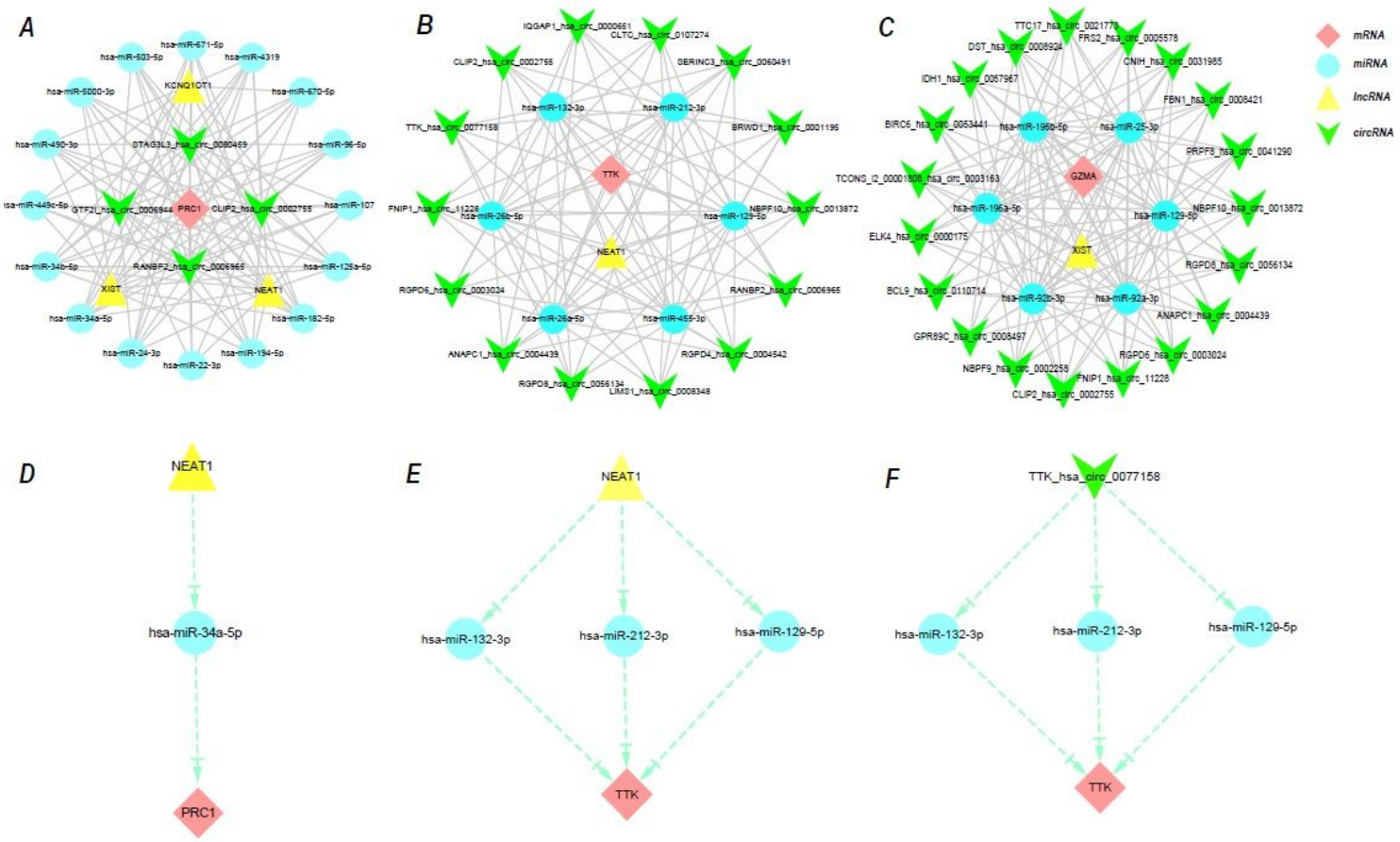

Figure 8

Three ceRNA networks of PRC1, TTK, and GZMA and the potential RNA regulatory pathways. A. ceRNA network of PRC1. B. ceRNA network of TTK. C. ceRNA network of GZMA. D. NEAT1-miR-212-3p/miR-132-3p/miR-129-5p-TTK. E XIST-miR-25-3p/miR-129-5p-GZMA. F. TTK_hsa_circ_0077158-miR-212-3p/miR-132-3p/miR-129-5p-TTK. Red diamonds represent the hub genes, blue circles represent miRNAs, yellow triangle represents lncRNAs, and the V represents the circRNA.