CAPZA2 and TRIB1 genomes contribute to the pathogenesis of ankylosing spondylitis

Yuanlin Yao
The First Affiliated Hospital of Guangxi Medical University
Xinli Zhan
The First Affiliated Hospital of Guangxi Medical University
Jie Jiang
The First Affiliated Hospital of Guangxi Medical University
Tuo Liang
The First Affiliated Hospital of Guangxi Medical University
Tianyou Chen
The First Affiliated Hospital of Guangxi Medical University
Hao guo
The First Affiliated Hospital of Guangxi Medical University
Hao Li
The First Affiliated Hospital of Guangxi Medical University
Zhen Ye
The First Affiliated Hospital of Guangxi Medical University
Wuhua Chen
The First Affiliated Hospital of Guangxi Medical University
Liyi Chen
The First Affiliated Hospital of Guangxi Medical University
Jiarui Chen
The First Affiliated Hospital of Guangxi Medical University
Shengsheng Huang
The First Affiliated Hospital of Guangxi Medical University
Xuhua Sun
The First Affiliated Hospital of Guangxi Medical University
Shaofeng Wu
The First Affiliated Hospital of Guangxi Medical University
Jichong Zhu
The First Affiliated Hospital of Guangxi Medical University
Chong Liu (liuchong@stu.gxmu.edu.cn)
The First Affiliated Hospital of Guangxi Medical University

Research Article

Keywords: Ankylosing spondylitis, single-cell sequencing, Pseudo-time and trajectory analysis, diagnostic models, veen

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1660404/v1

Abstract

Purpose: The pathogenesis of ankylosing spondylitis is not yet clear. Identification of new marker genes and cellular subtypes of immune cells in the blood of patients and exploration of new genetic loci in the pathogenesis of ankylosing spondylitis.

Methods: We downloaded single-cell RNA-seq profiles (GSE157595) and microarray datasets (GSE25101 and GSE73754) from the Gene Expression Omnibus (GEO) database and performed principal component analysis, GO enrichment analysis, KEGG pathway analysis, t-distribution random-adjacent embedding analysis, and analysis using SVM and WGCNA to identify marker genes. The intersection of SVM and single cell sequencing marker genes: CAPZA2 and TRIB1, were taken for testing of diagnostic model building ROC curves and immunohistochemical validation analysis.

Results: CAPZA2 and TRIB1 were identified as new diagnostic genes for AS. These two genes showed significant differential expression between AS and controls, and immunohistochemical results showed that the expression of CAPZA2 and TRIB1 was significantly higher in AS than in controls.

Conclusion: CAPZA2 and TRIB1 may be key factors in the pathogenesis of ankylosing spondylitis.

Introduction

Ankylosing spondylitis is a chronic inflammatory disease, but the mechanism is not clear[1]. The main sites of disease are the sacroiliac joints, spine and peripheral joints, with associated reactive arthritis, psoriasis and inflammatory bowel disease, etc. AS was once thought to be more common in men, with a male to female ratio of 10.6:1, but recent reports suggest that the prevalence of ankylosing spondylitis in men and women is comparable, almost equal, except that women have a more delayed onset of the disease, which is less severe and progresses more slowly than men[2]. The disease progresses more slowly in women than in men. The age of onset is usually between 15 and 30 years, with a significant decrease in the incidence in children over 30 years of age and under 8 years of age. Studies have found a strong association between AS patients and human leukocyte antigen B27 (HLA-B27, or B27 for short); the rate of B27 positivity is 5%-7% in the general population in China, but up to 90% in AS patients, and there is a tendency for AS to run in families[3].

The cause of AS is not known, but there is a genetic, infectious and autoimmune correlation, and the joint changes in AS are characterised by granulomatous synovitis with synovial thickening, fibrosis and ossification, and infiltration of plasma cells, macrophages and lymphocytes[4]. As a chronic disease, AS causes severe irreversible physical damage as well as psychological and financial devastation to the patient. There is therefore an urgent need to identify its pathogenesis and more accurate biomarkers.

However, there have been few reports on the analysis of ankylosing spondylitis, particularly its pathogenesis and biomarkers, through the use of single cell DNA. Spine surgeons can use single-cell RNA-Seq to identify differences between each cell and also to discover more precisely how different

genes are differentially expressed in different cell populations. On the one hand, the identification of marker genes from individual cell-to-cell transcriptional data reveals more accurate cellular biomarker genes and differences in gene expression between cell populations[5]. On the other hand, the identification of cell subtypes from each cell and a three-dimensional time tree of cell differentiation to understand the progression of ankylosing spondylitis may have implications for the clinical management of patients[6]. In this study, we used single-cell RNA-seq data downloaded from the Gene Expression Omnibus (GEO) database to illustrate marker genes and cell subtypes in ankylosing spondylitis peripheral blood cells, and validated them with microarray data, and correlation experiments.

Methods

Data download, quality control, data filtering and normalisation

We downloaded three AS datasets from the Gene Expression Omnibus (GEO) database: the single-cell RNA-seq profile GSE157595 and the microarray datasets GSE25101 and GSE73754 for the analysis of diagnostic markers for AS. The data used in this study came from patients with and without ankylosing spondylitis and the specific inclusion criteria can be reviewed in the original article. The dataset contained data from a total of 114 patients, with 72 patients with ankylosing spondylitis as the experimental group and 42 patients without ankylosing spondylitis as the control group. Single cell sequencing analysis included only cells of sufficient quality for amplification and next generation RNA sequencing[7]. We used the software package Seurat for data analysis to assess the number of genes in each cell and the number of genes sequenced, with quality control and data filtering performed accordingly. The average of duplicate gene expression was taken. Gene expression found in fewer than three cells was excluded, and cells with fewer than 50 identified genes sequenced were also excluded. The percentage of mitochondrial genes was calculated for this study, and violin plots showing the number and sequence of genes, and the percentage of mitochondrial genes. Data were normalised after the above adjustments[8]. The top 1500 genes with large normalised variance were selected for the following analysis. The symbols of the top 10 genes were labelled[8]. In addition, two AS datasets were downloaded from the Gene Expression Omnibus (GEO) repository and used for analysis of possible diagnostic markers. In addition the GSE25101 and GSE73754 datasets contained a total of 102 samples. We first matched the probe names of both datasets to their corresponding gene symbols using the programming language PERL, and then de-batched and normalised both datasets so that they could be analysed at the same level[9]. We combined the AS samples from the dataset and controls into one file for uniform analysis. In this study, we used the turning language PERL for text processing and the turning language R for statistical analysis and image plotting[10].

Principal component analysis

Principal component analysis was performed on the single-cell dataset GSE157595. The principal component analysis was performed using 1500 screened genes with large normalised variance and the top 20 principal components were selected for the following analysis. The top 20 genes in each principal

component will be plotted. an overview of the PCA and heat map will be plotted in the figure. P-values for each principal component are used for the following analyses.

TSNE analysis and identification of marker genes

Principal component analysis was followed by t-distribution random neighbourhood embedding (TSNE) analysis. Cells were sorted into different clusters. Heat maps were then used to show the marker genes in the different clusters. A violin plot was used to show the expression levels of marker genes in each cluster. Bubble plots were used to show the expression levels in each cluster.

Clustering of ankylosing spondylitis cells

Ankylosing spondylitis cells are clustered using the algorithm developed in the R package "Monocle". The number of clusters was chosen automatically by the R package "Monocle" according to the method described earlier. Fourteen clusters were identified: clusters 0 to 13.

Notes on cell types

We used the "SingleR" package to perform cell annotation. The identified clusters were annotated. Each cell is also annotated.

Pseudo-time and trajectory analysis

The R packages "Monocle", "clusterProfiler", "org.Hs.eg.db", " enrichplot" and "ggplot2" for cell trajectory and pseudo-time analysis.

GO and KEGG analysis

To gain more insight into the molecular function (MF), biological process (BP), cellular composition (CC) and pathway enrichment of these differentially expressed genes, we used the 'clusterProfiler' package, the 'org.Hs.eg.db' package, the 'richplot' package, the 'ggplot2' package and the 'GOplot' package. "package, the "richplot" package, the "ggplot2" package and the "GOplot" package. The "enrichment plot" package, the "ggplot2" package, and the "GOplot" package perform GO enrichment on GO enrichment analysis and KEGG pathway enrichment analysis were performed on differentially expressed genes. Enrichment analysis. GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed using the 'digest' and 'GOplot' R packages. The R package of "digest" and "GOplot" was used to perform the analysis. The names of the tagged genes were transferred to the gene IDs. Bar charts, bubble charts, circle charts and cluster charts are used to show the results in different directions.

Disease Ontology Enrichment Analysis (DO) and Gene Set Enrichment Analysis (GSEA)

DO enrichment analysis is important for understanding complex pathogenesis, diagnosis and early prevention of major diseases. We use the 'ggplot2' package, the 'org.Hs.eg.db' package, the 'enrichment plot' package, the 'clusterProfiler' package, the 'GSEABase' package and the 'DOSE' package.

"clusterProfiler" package, "GSEABase" package and "DOSE" package to perform DO enrichment analysis and visualize the results of all differentially expressed genes. GSEA enrichment analysis is a genome-wide expression profiling microarray data analysis method that integrates a priori knowledge of gene localisation, biological significance and function. We use the 'clusterProfiler' package, the 'org. h.s.eg .db' package, the 'limma' package, and the 'richplot' package. 'richplot' package to enrich the genes and visualise the results.

Support vector machine (SVM) analysis, veen plots, ROC diagnostic curve analysis

Support vector machine analysis is a class of generalised linear classifiers that perform binary classification of data in a supervised learning manner, where the convenience of the decision is to learn the maximum marginal plane that the sample can be solved for, i.e. to separate the hyperplanes and solve for the correctly partitioned dataset with the maximum geometric spacing. The venn diagram was constructed using the R data package "VennDiagram" to obtain the intersection of the marker genes from the GSE157595 dataset and the difference genes from the GSE25101 and GSE73754 datasets. The ROC curve was constructed using the R package "pROC".

Differential analysis of immune cell correlations

The 'limma' package, the 'ggExtra' package, the 'ggpubr' package and the 'reshape2' package are used to associate immune cells with genes. "package were used to associate immune cells with genes. The proportion of immune cells and the differences between the AS and control groups were also plotted using Lollipop.

Immunohistochemical analysis

In this paper, five cases of interspinous ligaments diagnosed as AS with posterior convexity deformity and surgically removed from the First Clinical Affiliated Hospital of Guangxi Medical University were used as the experimental group, and three cases of interspinous ligaments diagnosed as spinal fracture and removed intraoperatively were used as the normal control group to detect the difference in expression of CAPZA2 and TRIB1 between as and control groups: //abclonal.com.cn/catalog/A2054, catalogue number:A2054) at a dilution ratio of 1:200. TRIB1-specific antibody was purchased from ABclonal (https://abclonal.com.cn/catalog/A10134, item number:A10134) at a dilution ratio of 1: 1000. Interspinous ligament tissue was isolated and preserved by immersion in formalin solution for 10 minutes. Subsequently, after laboratory operations such as wax sealing, sectioning, antigen repair, antibody hybridization, colour development and tissue closure, we obtained all 16 immunohistochemical sections with completed staining. We observed them under an inverted microscope, with separate image intercepts for the AS and control groups.

Results

Data download, quality control, data filtering, standardisation of single cell sequencing

We downloaded three AS datasets from the Gene Expression Omnibus (GEO) database: the single-cell RNA-seq profile GSE157595 and the microarray datasets GSE25101 and GSE73754 and. The single-cell RNA-seq profile GSE157595 contains six AS patients extracted and mixed into one experimental sample and six healthy patients extracted and mixed into one control sample. The X-axis represents the names of the samples and the dots in the graph represent the cells for each assessment[11]. The Y-axis reflects the number of genes in each cell. The distribution of genes in each cell is shown in Figure 1A and the depth of sequencing of the samples is shown in Figure 1B, the mitochondrial gene content was not screened in this study (Figure 1C). The X-axis represents the sequencing depth of each sample, and the Y-axis represents the number of mitochondrial genes. A correlation coefficient of 0 means that there is no direct relationship between the sequencing depth of the sample and the number of mitochondrial genes (Figure 2A). x-axis represents the sequencing depth of each sample, and the Y-axis represents the number of sequenced genes. A correlation coefficient of 0.13 means that there is a positive relationship between the sequencing depth of the sample and the number of sequenced genes (Figure 2B). x-axis represents the average expression of each gene. The X-axis represents the mean expression of each gene. The Y-axis represents the normalized variance. Genes with large normalised variances were screened and selected for the following analysis as they represent heterogeneity between cells. A total of 581 genes with large normalised variances were obtained (Figure 3A). The top 10 genes with normalised variance are shown in Figure 3B. The top 10 genes with normalised variance can be seen as: IGHA, IGHG3, IGLC2, LCN2, SRGN, CAMP, IGHG4, JCHAIN, IGHM, IGHG1.

Principal Component Analysis (PCA)

The main sources of variation were calculated in this study using variable genes as shown in PCA[12]. the aim of PCA was to identify the characteristics of these variable genes. Figure 4A shows the overall distribution of cells in the samples in PC1 and PC2. Figure 4B shows the p-value values for each PC, with smaller p-values indicating greater importance in the main components of the samples. Figure 4C shows the top 20 signature genes expressed by each PC. Figure 4D shows the expression of the top 20 signature genes for each cell in each PC expression. pc1 was considered to be one of the most dominant principal components in PCA, with the top 20 genes being:DENND4B, KLK3, CCDC14, AR, RCE1, SAE1, ZNF577, TXNDC16, AMACR, ASH1L FBXO41, HNRNPC, SBNO1, KRR1, MALAT1, KLK2, COMMD2, ZNF254, C22orf29, FAM219B.

TSNE analysis of cell types and identification of marker genes

The results of the TSNE analysis classified the cells into 14 clusters (Figure 5C). After annotation (Figure 5D), we found that these cells could be divided into cluster 3 for T cells, clusters 0 and 12 for tissue stem cells, clusters 2 and 11 for monocytes, cluster 10 for progenitor B cells, cluster 13 for macrophages, clusters 1, 5 and 8 for B cells, clusters 4, 7 and 9 for erythrocytes and cluster 6 for myeloid cells. The scatter plot (Fig. 5A) shows the expression of the top 10 genes with the highest normalised variance in each cluster. The violin plot (Figure 5B) shows the expression of the top two genes in cluster 0, LEPR and

COL1A2, in each cluster. The heat map shows the expression of the signature genes in each cluster 0-13 (Fig. 5E).

Pseudo-time and trajectory analysis

Fourteen clusters were annotated according to marker genes: clusters 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. Figure 6 shows the pseudo-temporal and trajectory analysis, clusters 0, 1 due to stem cells; clusters 11, 12, 13 by macrophages, monocytes; clusters 2, 5, 6 by T cells; clusters 3, 4, 10 by B cells; clusters 7, 8, 9 by erythrocytes.

GO and KEGG analysis

We performed GO enrichment analysis on these differentially expressed genes to analyse their BP, CC and MF and obtained the results of all enrichment analyses while visualising the top 10 most important genes (Fig. 7A)[13]. Figure 7A shows that GO entries were mainly enriched for neutrophil, leukocyte chemotaxis, granulocyte, interferon, extracellular matrix, intracellular matrix, cell adhesion molecule binding, myocardin, actin and endoplasmic reticulum composition. The clustering diagram in Figure 7B demonstrates the enriched GO term[14]. The KEGG pathway (Figure 7C) is mainly enriched for phagosomes, regulatory actin, atherosclerosis, cell adhesion molecules, Rap1 signalling pathway, rheumatoid arthritis, Th17 cell differentiation, haematopoietic cell lineage and allograft rejection[15]. The clustering diagram in Figure 7D demonstrates the enriched KEGG term[14].

Disease Ontology Enrichment Analysis (DO) and Gene Set Enrichment Analysis (GSEA)

Disease ontology enrichment analysis (Figure 8E) showed enrichment mainly in primary immunodeficiency diseases, hereditary coagulation, atherosclerosis, bacterial infections, fungal infections, renal diseases, and urinary tract diseases[16]. Analysis of GSEA enrichment in the AS group showed that the main pathways of GO enrichment were cotranslational protein targeting to membrane, establishment of protein localization to endoplasmic reticulum, mitochondrial translation, ncrna metabolic process and ncrna processing (Figure 8A), whereas the KEGG pathway was predominantly enriched. in complement and coagulation cascades, citrate cycle tca cycle, graft versus host disease, oxidative phosphorylation and proteasome (Fig. 8C)[17].

CAPZA2 and TRIB1 are genes used as diagnostic models for AS

We obtained the intersection by constructing venn diagrams from the single-cell dataset marker genes and the microarray dataset SVM, and the two genes CAPZA2 and TRIB1 met all our screening requirements (Figure 8G)[18]. We then constructed ROC diagnostic model curves to test our analysis and found that the area under the curve for the AS diagnostic model constructed by CAPZA2 was 0.741 with 95% CI of 0.637-0.835, (Figure 9C)[19]. The area under the curve for the diagnostic model of AS constructed by TRIB1 was 0.727, 95% CI 0.617-0.828 (Figure 9D). the area under the curve for the diagnostic models of CAPZA2 and TRIB1 was much greater than 0.5, indicating a more accurate diagnostic curve. box plot analysis of the differences between the CAPZA2 and TRIB1 diagnostic models

in AS and controls (Figure 9A, 9B), the The differences were statistically significant ($p<0.05$), suggesting the significance of the CAPZA2 and TRIB1 diagnostic models in the diagnosis of AS. Pseudotime and trajectory analysis demonstrated a possible correlation between the development of ankylosing spondylitis and inflammatory cell differentiation[20]. Pathogenic enrichment analysis revealed enrichment mainly in primary immunodeficiency disease, hereditary coagulation, atherosclerosis, and bacterial infection[21].

Differential analysis of immune cell correlations

Fitted curves were constructed for the correlation between the two genes and immune cells for the diagnostic model (Figures 10, 11)[22]. Figure 12A shows that 7 immune cells were associated with CAPZA2 ($p < 0.05$) and Figure 12B shows that 6 immune cells were associated with TRIB1 ($p < 0.05$).

Immunohistochemical analysis

We performed immunohistochemical staining of CAPZA2 and TRIB1 in the interspinous ligaments of five patients with AS and three patients with spinal fractures, respectively. The results showed that the specific expression of each of CAPZA2 and TRIB1was significantly higher in AS than in the control group (Figure 13A1-D2). After detecting the positive rate of all immunohistochemical images with Image J software, the positive rate data of CAPZA2 and TRIB1 were imported into R software for statistical analysis[23]. The positive rate of CAPZA2 and TRIB1 were statistically found to be much higher in AS than in the control group, and the difference was statistically significant ($P<0.05$), indicating that CAPZA2 and TRIB1 in AS and the control group there was a significant difference.

Discussion

In our study, we performed GO enrichment analysis of 114 samples obtained from the GEO database for differentially expressed genes and showed that GO entries were mainly enriched in neutrophils, leukocyte chemotaxis, granulocytes, interferons, extracellular matrix, intracellular matrix, cell adhesion molecule binding, myocardin, actin and endoplasmic reticulum components[24]. The KEGG pathway is mainly enriched in phagosomes, regulatory actin, atherosclerosis, cell adhesion molecules, Rap1 signalling pathway, rheumatoid arthritis, Th17 cell differentiation, haematopoietic cell lineage and allograft rejection. Cells were classified into 14 clusters by TSNE analysis. The first 2 upregulated genes in cluster 0 were LEPR and COL1A2. venn diagrams were constructed from the single cell dataset marker genes and the microarray dataset SVM to obtain 2 intersecting genes, CAPZA2 and TRIB1[25].

CAPZA2 (Capping Actin Protein Of Muscle Z-Line Subunit Alpha 2) is a Protein Coding gene. Diseases associated with CAPZA2 include Alacrima, Achalasia, And Mental Retardation Syndrome. Among its related pathways are Cytoskeleton remodeling Neurofilaments and Innate Immune System . Yan Huang et al. reported two pediatric probands who carry damaging heterozygous de novo mutations in CAPZA2 (HGNC: 1490) and exhibit neurological symptoms with shared phenotypes including ...expression of the CAPZA2 (HGNC: 1490) and exhibit neurological symptoms with shared phenotypes including global

motor development delay, speech delay, intellectual disability, hypotonia and a history of seizures .Expression of the CAPZA2 variants affects bristle morphogenesis, a process that requires extensive actin polymerization and bundling during development[26].

TRIB1 (Tribbles Pseudokinase 1) is a Protein Coding gene. Hamish D McMillanet al. reported that TRIB1 has been most well characterised structurally and plays roles in diverse cancer types[27]. The most well-understood role of TRIB1 is in acute myeloid leukaemia, where it can regulate C/EBP transcription factors and kinase pathways. Structure-function studies have uncovered conformational switching of TRIB1 from an inactive to an active state when it This conformational switching is centred on the active site of TRIB1, which appears to be accessible to small-molecule inhibitors in Beyond myeloid neoplasms, TRIB1 plays diverse roles in signalling pathways with well-established roles in tumour Beyond myeloid neoplasms, TRIB1 plays diverse roles in signalling pathways with well-established roles in tumour progression[28].

Our study showed that the area under the curve for the diagnostic models of CAPZA2 and TRIB1 was much greater than 0.5, indicating a more accurate diagnostic curve[29]. box plot analysis of the differences between the CAPZA2 and TRIB1 diagnostic models in AS and controls showed a statistically significant difference ($p < 0.05$)[30]. Correlations between the two genes used to construct the diagnostic model and the six immune cells were found to be $p < 0.05$[31]. We performed immunohistochemistry for CAPZA2 and TRIB1 in the interspinous ligaments of five AS cases and three spinal fractures, respectively, and found that the expression of CAPZA2 and TRIB1 was significantly higher in the AS group than in the normal control group, with a difference of statistically significant ($P < 0.05$).

We first screened the differentially expressed genes in AS by GEO single-cell sequencing dataset, then performed GO enrichment analysis, KEGG pathway enrichment analysis[32]. We then combined the two microarray datasets to perform GO enrichment analysis, KEGG pathway enrichment analysis, DO enrichment analysis, GSEA enrichment analysis, followed by SVM analysis, WGCNA, and Venn diagram cross-tabulation of differentially tagged genes from single cell sequencing analysis and SVM analysis of the two microarray datasets. Diagnostic modelling and immune cell differential analysis of the genes CAPZA2 and TRIB1 were also performed. Finally, we also analysed CAPZA2 and TRIB1 staining by immunohistochemistry separately and the results showed that CAPZA2 and TRIB1 had more gene expression in AS than in normal controls. Similar to other studies, there are limitations to our study. First, the sample size was insufficient. A total of 114 samples were used in our bioinformatics analysis, including 72 AS samples and 42 control samples. Secondly, we did not have more laboratory analyses to test our results, which is not nearly enough.

Conclusion

Upregulation of CAPZA2 and TRIB1 may be a key factor in the progression of ankylosing spondylitis, providing a new direction for research and treatment of the pathogenesis of ankylosing spondylitis.

Declarations

Data Availability Statement

The datasets supporting the conclusions of this article are available in the GEO atabase:https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157595 https//www.ncbi.nlm.nih.gov/geo/query/acc.cgifacc=GSE25101and https//www.ncbi.nlm.mih.gov/geo/query/acc,cgiracc-GSF73754.

Ethics Statement

The studies involving human participants were reviewed and approved by Ethics Department of the First Affiliated Hospital of Guangxi Medical University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author Contributions

Zhan Xinli, Liu Chong, Yao Yuanlin: Conceptualization, Methodology.**Jie Jiang, Tuo Liang,Tianyou Chen, Hao guo:** Data curation, Investigation. **Hao Li, Zhen Ye, Wuhua Chen, Liyi Chen,Jiarui Chen:** Formal analysis, Software.**Shengsheng Huang, Xuhua Sun, Shaofeng Wu, Jichong Zhu:** Visualization. **Yao Yuanlin, Liu Chong:** Writing- Reviewing and Editing.All co-authors participated in the laboratory operation. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Youth Science Foundation of Guangxi Medical University, Grant/Award Numbers: GXMUYFY201712, Guangxi Young and Middle-aged Teacher's Basic Ability Promoting Project, Grant/Award Number: 2019KY0119.

Acknowledgments

The authors are grateful to Dr. Xin Li Zhan (Spine and Osteopathy Ward, The First Affiliated Hospital 21 of Guangxi Medical University) for his kindly assistance in all present study stages.

References

1. Yang, L., et al., *A Possible Role of Intestinal Microbiota in the Pathogenesis of Ankylosing Spondylitis.* Int J Mol Sci, 2016. **17**(12).
2. Wang, R., S.E. Gabriel, and M.M. Ward, *Progression of Nonradiographic Axial Spondyloarthritis to Ankylosing Spondylitis: A Population-Based Cohort Study.* Arthritis Rheumatol, 2016. **68**(6): p. 1415-21.
3. Pelaez-Ballestas, I., M. Romero-Mendoza, and R. Burgos-Vargas, *If three of my brothers have ankylosing spondylitis, why does the doctor say it is not necessarily hereditary? The meaning of risk*

in multiplex case families with ankylosing spondylitis. Chronic Illn, 2016. **12**(1): p. 58-70.

4. Shi, H., et al., *GM-CSF Primes Proinflammatory Monocyte Responses in Ankylosing Spondylitis.* Front Immunol, 2020. **11**: p. 1520.
5. Svensson, V., et al., *Power analysis of single-cell RNA-sequencing experiments.* Nat Methods, 2017. **14**(4): p. 381-387.
6. Garcia-Montoya, L., H. Gul, and P. Emery, *Recent advances in ankylosing spondylitis: understanding the disease and management.* F1000Res, 2018. **7**.
7. Zhang, H., L. He, and L. Cai, *Transcriptome Sequencing: RNA-Seq.* Methods Mol Biol, 2018. **1754**: p. 15-27.
8. Lin, X.D., et al., *Identification of marker genes and cell subtypes in castration-resistant prostate cancer cells.* J Cancer, 2021. **12**(4): p. 1249-1257.
9. Suwazono, S. and H. Arao, *A newly developed free software tool set for averaging electroencephalogram implemented in the Perl programming language.* Heliyon, 2020. **6**(11): p. e05580.
10. Li, X., et al., *CT features and quantitative analysis of subsolid nodule lung adenocarcinoma for pathological classification prediction.* BMC Cancer, 2020. **20**(1): p. 60.
11. Park, M., et al., *Nuclear image analysis study of neuroendocrine tumors.* Korean J Pathol, 2012. **46**(1): p. 38-41.
12. Abdelhafez, O.H., et al., *Metabolomics analysis and biological investigation of three Malvaceae plants.* Phytochem Anal, 2020. **31**(2): p. 204-214.
13. Sun, S., et al., *Identification and Validation of Autophagy-Related Genes in Chronic Obstructive Pulmonary Disease.* Int J Chron Obstruct Pulmon Dis, 2021. **16**: p. 67-78.
14. Zhang, C., et al., *The identification of key genes and pathways in hepatocellular carcinoma by bioinformatics analysis of high-throughput data.* Med Oncol, 2017. **34**(6): p. 101.
15. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D457-62.
16. Liang, W., et al., *Identification of Susceptibility Modules and Genes for Cardiovascular Disease in Diabetic Patients Using WGCNA Analysis.* J Diabetes Res, 2020. **2020**: p. 4178639.
17. Qi, B., et al., *Integrated Weighted Gene Co-expression Network Analysis Identified That TLR2 and CD40 Are Related to Coronary Artery Disease.* Front Genet, 2020. **11**: p. 613744.
18. Jia, A., L. Xu, and Y. Wang, *Venn diagrams in bioinformatics.* Brief Bioinform, 2021. **22**(5).
19. Metz, C.E., *Basic principles of ROC analysis.* Semin Nucl Med, 1978. **8**(4): p. 283-98.
20. Street, K., et al., *Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.* BMC Genomics, 2018. **19**(1): p. 477.
21. Amorim-Vaz, S., et al., *RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens In Vivo Applied to the Fungus Candida albicans.* mBio, 2015. **6**(5): p. e00942-15.

22. Jiang, J., et al., *Upregulated of ANXA3, SORL1, and Neutrophils May Be Key Factors in the Progressionof Ankylosing Spondylitis.* Front Immunol, 2022. **13**: p. 861459.
23. Rha, E.Y., J.M. Kim, and G. Yoo, *Volume Measurement of Various Tissues Using the Image J Software.* J Craniofac Surg, 2015. **26**(6): p. e505-6.
24. Oakes, S.A. and F.R. Papa, *The role of endoplasmic reticulum stress in human pathology.* Annu Rev Pathol, 2015. **10**: p. 173-94.
25. Huang, S., et al., *Applications of Support Vector Machine (SVM) Learning in Cancer Genomics.* Cancer Genomics Proteomics, 2018. **15**(1): p. 41-51.
26. Huang, Y., et al., *Variants in CAPZA2, a member of an F-actin capping complex, cause intellectual disability and developmental delay.* Hum Mol Genet, 2020. **29**(9): p. 1537-1546.
27. Yoshino, S., et al., *Trib1 promotes acute myeloid leukemia progression by modulating the transcriptional programs of Hoxa9.* Blood, 2021. **137**(1): p. 75-88.
28. Iwamoto, S., et al., *The role of TRIB1 in lipid metabolism; from genetics to pathways.* Biochem Soc Trans, 2015. **43**(5): p. 1063-8.
29. Cheng, J., et al., *Validation of Ten Noninvasive Diagnostic Models for Prediction of Liver Fibrosis in Patients with Chronic Hepatitis B.* PLoS One, 2015. **10**(12): p. e0144425.
30. Noel, C.W., et al., *The fragility of statistically significant findings from randomized trials in head and neck surgery.* Laryngoscope, 2018. **128**(9): p. 2094-2100.
31. Poldrack, R.A., G. Huckins, and G. Varoquaux, *Establishment of Best Practices for Evidence for Prediction: A Review.* JAMA Psychiatry, 2020. **77**(5): p. 534-540.
32. Schatten, H., et al., *Microtubules are required for centrosome expansion and positioning while microfilaments are required for centrosome separation in sea urchin eggs during fertilization and mitosis.* Cell Motil Cytoskeleton, 1988. **11**(4): p. 248-59.

Figures

Figure 1

Results of quality control, data filtering. **(A)** Distribution of identified genes in each cell. **(B)** Distribution of identified genes by sequencing depth. **(C)** Percentage of mitochondrial genes.

Figure 2

Results of data filtering and normalization. **(A)** Sequencing depth versus mitochondrial gene content. **(B)** Sequencing depth versus number of genes sequenced.

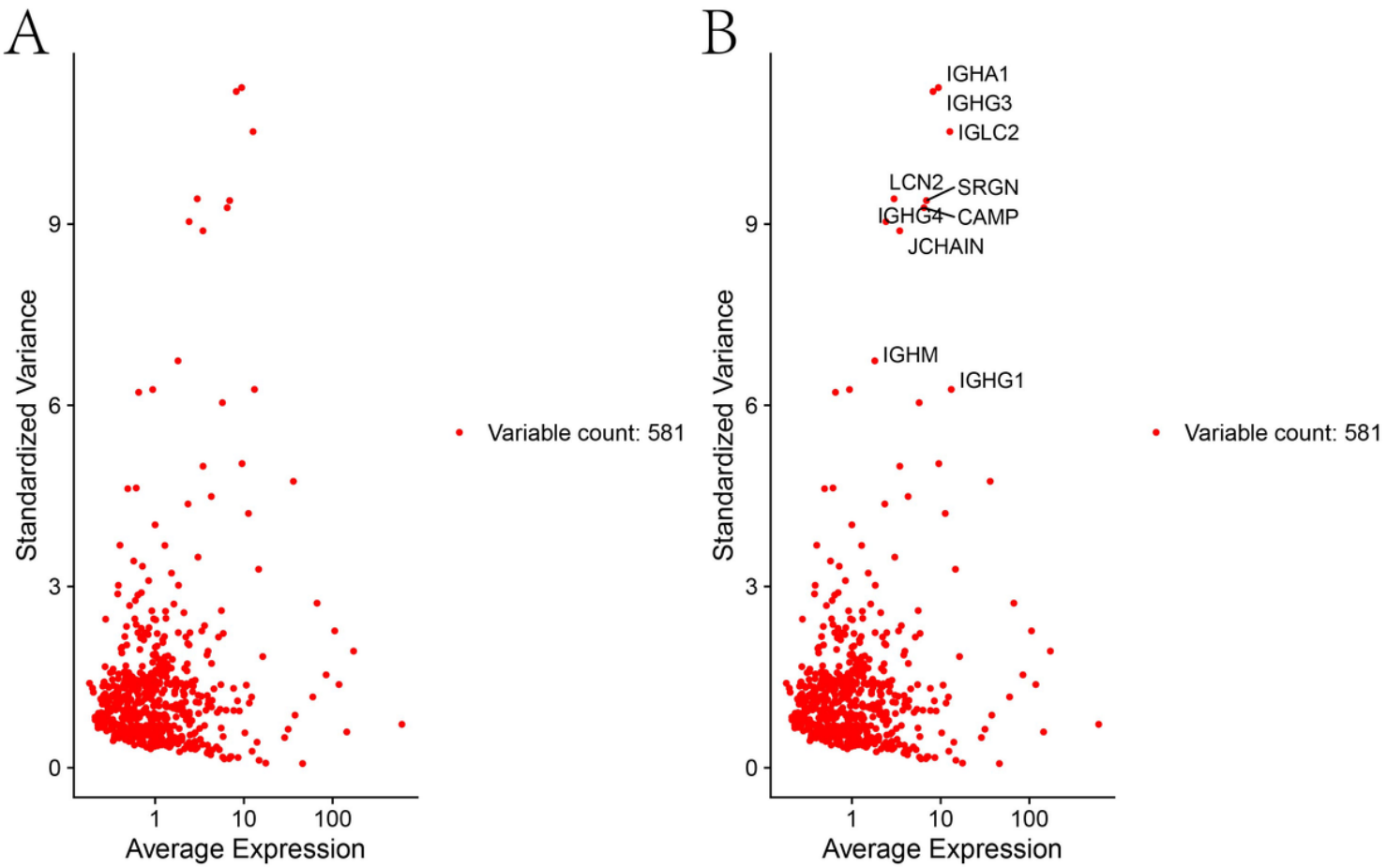

Figure 3

Results of data filtering and normalization. **(A)** A total of 581 genes with large normalized variances. **(B)** The top 10 genes of the 581 genes with normalized variance are shown. The X-axis represents the mean expression of each gene. The Y-axis represents the normalized variance.

Figure 4

Results of principal component (PC) analysis. **(A)** Overview of the cell distribution for each donor in PC1 and PC2. **(B)** P-value values for each PC. **(C)** Heat map showing the expression of key genes in each pc1-pc4. **(D)** The top 4 principal components (PCs) and the top 20 genes expressed in each PC.

Figure 5

TSNE and marker gene results for cell types. **(A)** Scatter plot showing the distribution of each cell among the top 10 normalized variance gene marker genes. **(B)** Gene expression of the top 2 key genes highly

expressed in cluster 0 in each cluster. **(C, D)** Cell annotations. **(E)** Expression of signature genes in each cluster 0-13.

Figure 6

Pseudo-time and trajectory analysis.

Figure 7

Results of GO and KEGG analysis. **(A)** Results of GO analysis for identified marker genes. **(B)** Presentation of GO terms in the clustering diagram. **(C)** Results of KEGG analysis in identified marker genes. **(D)** The KEGG pathway is shown in the clustering diagram.

Figure 8

GSEA, DO analysis, SVM analysis, venn diagram. **(A-D)** shows GO and KEGG results from GSEA enrichment analysis showing the results of SVM analysis. **(E)** Shows results from DO enrichment analysis. **(F)** SVM analysis. **(G)** shows the crossover results of single cell sequencing analysis and SVM analysis.

Figure 9

ROC curve analysis and difference box plot analysis. **(A, B)** Difference box plots for the 2 genes of the model. **(C -D)** ROC curves for the 2 genes of the model are shown.

Figure 10

Correlation analysis of CAPZA2 with immune cells. **(A-F)** Correlation analysis of CAPZA2 and 6 immune cells is shown. If $R > 0$, this indicates higher gene expression and higher immune cell content; if $R < 0$, this indicates higher gene expression and lower immune cell content.

Figure 11

Correlation analysis of TRIB1 with immune cells. **(A-F)** shows the correlation analysis between CAPZA2 and 6 immune cells. If R > 0, this indicates higher gene expression and higher immune cell content; if R < 0, this indicates higher gene expression and lower immune cell content

Figure 12

Immune-related lollipop plot. **(A)** Immune cells showing significant differences in TRIB1 in AS and controls. **(B)** Immune cells showing significant differences in CAPZA2 in AS and controls.

Figure 13

Immunohistochemistry. **(A1-D2)** shows specific expression of CAPZA2 and TRIB1 in AS and controls.