Original Article

Prediction of target genes in community-acquired pneumonia based on the bioinformatics method

Yangsong Zuo[1#], Wenyi Shen[1#], Guiming Chen[2], Huailian Liu[3], Na Liu[4], Ting Xu[4], Juan Pu[5]

[1]Department of Aspiration Medicine, Lianshui County People's Hospital, Huai'an, China; [2]Department of Cardiothoracic Surgery, Lianshui County People's Hospital, Huai'an, China; [3]Hospital Infection Control Division, Huai'an Maternal and Child Health Centre, Huai'an, China; [4]Department of Respiratory Medicine, Nanjing Chest Hospital, Affiliated Nanjing Brain Hospital, Nanjing Medical University, Nanjing, China; [5]Department of Radiotherapy, Lianshui County People's Hospital, Huai'an, China

Contributions: (I) Conception and design: Y Zuo, J Pu; (II) Administrative support: W Shen; (III) Provision of study materials or patients: G Chen; (IV) Collection and assembly of data: N Liu; (V) Data analysis and interpretation: T Xu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Juan Pu, MD. Radiotherapy Department, Lianshui County People's Hospital, 6 Hongri Avenue, Huai'an 223400, China. Email: lspuj@sina.com; Ting Xu, MD. Department of Respiratory Medicine, Nanjing Chest Hospital, Affiliated Nanjing Brain Hospital, Nanjing Medical University, 215 Guangzhou Road, Nanjing 210029, China. Email: shootingkytg@163.com.

Background: To screen the related genes of community-acquired pneumonia (CAP) by bioinformatics technology, and to analyze the clinical value of key genes.

Methods: Gene chip data sets containing CAP patients and normal controls were screened from the Gene Expression Omnibus (GEO) database. The downregulated differentially expressed genes (DEGs) were screened using a gene expression analysis tool (GEO2R). Simultaneously, gene set enrichment analysis (GSEA) was used to explore the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and core genes related to CAP. The candidate genes were then intersected with the genes reported in Online Mendelian Inheritance in Man (OMIM), and the clinical value of these candidate genes was examined based on a literature search. Finally, the clinical data of the CAP patients were retrospectively analyzed. Detect the type of pathogenic bacteria in bronchial-alveolar lavage fluid (BALF) using metagenomics next-generation sequencing (mNGS) high throughput sequencing technology, and detect the expression of key genes through liquid based cell immunohistochemistry to analyze the correlation between pathogenic bacteria and key genes.

Results: Through the intersection of Venn diagrams, 175 co-expressed downregulated DEGs related to CAP were identified. A total of 4 candidate genes, including *ICOS*, *IL7R*, *ITK*, and *ZAP70*, were obtained by constructing the protein mutual aid network and conducting a module analysis of the common differentially expressed genes. The core genes in the GSEA enrichment pathways were intersected with the CAP-related genes reported in the relevant literature retrieved from the OMIM database. In the Venn diagram, two genes that coexist with OMIM include *IL7R* and *PIK3R1*. After considering our findings and the relevant literature, we determined that the key gene related to the occurrence and development of CAP was *IL7R*. The mNGS detected 13 kinds of bacteria, 4 kinds of fungi, and 2 kinds of viruses. Based on immunohistochemical results, it was found that there were relatively more bacteria detected in the *IL7R* high expression group.

Conclusions: The identification of the key gene *IL7R* and the related signaling pathways extend understanding of the pathogenesis of CAP and provide a theoretical basis for clinical targeted therapy research.

Keywords: Community-acquired pneumonia; public microarray database; gene set enrichment analysis (GSEA); *IL7R*; bioinformatics method; metagenomics next-generation sequencing (mNGS)

Submitted Feb 08, 2023. Accepted for publication May 11, 2023. Published online May 22, 2023.
doi: 10.21037/jtd-23-592
View this article at: https://dx.doi.org/10.21037/jtd-23-592

Introduction

Community-acquired pneumonia (CAP) refers to pneumonia that occurs when the human body is infected with bacteria, fungi, viruses, or mixed infections outside the hospital or in the community environment. CAP is a common infectious disease that threatens human health. Epidemiological studies have shown that the mortality rate of hospitalized CAP patients is 4–14% worldwide (1-3). When pneumonia develops into severe pneumonia, in addition to causing abnormal responses of the respiratory system, it also damages other systems of the body, which may eventually induce acute respiratory distress syndrome, sepsis, and multiple organ damage (4). Currently, CAP is the leading cause of death from infectious diseases in developing countries (5).

There has been great progress in terms of the etiology and imaging diagnosis of pneumonia, and various anti-infective treatments have also reduced the intensive care unit admission rate of CAP patients; however, the mortality rate of CAP patients remains high (6). Additionally, clinicians are prone to bias in diagnosing CAP based on their experience, and many pneumonia scores are relatively complex and often involve the combined application of multiple indicators. These objective conditions increase the difficulty of CAP diagnosis and treatment (7,8). Thus, a new and rapid diagnostic method or biomarker urgently needs to be found to improve the diagnosis and treatment of CAP.

In recent years, the rapid development of gene chip technology has become an effective tool for exploring and monitoring genetic changes in diseases (9). Studies have shown that gene mutations, uncontrolled gene expression, and epigenetic changes play key roles in the development of CAP (10,11). Targeted lipidomics has identified phospholipids and lysophospholipids as biomarkers for assessing CAP (12). The circulating level of matrix metalloproteinase 9, and its ratio to tissue inhibitor of metalloproteinases 1 (TIMP-1) can be used as a predictor of severity in patients with CAP (13). The expression of Krueppel-like factor 4 in phagocytes regulates the expression of the early inflammatory response and disease severity of CAP patients (14). The genetic polymorphisms of T-cell immunoglobulin and mucin domain-1 (TIM-1) gene promoter rs9313422 G>C, and rs41297579 G>A are associated with the risk of CAP in children (15). Thus, searching for new biomarkers by conducting systematic analyses of gene chip technology and using the bioinformatics method is very important for the early diagnosis, treatment, and prognosis of CAP patients.

This study was based on 2 public databases: the Gene Expression Omnibus (GEO) and the Online Mendelian Inheritance in Man (OMIM) databases. Bioinformatics analysis methods were used to explore the differentially expressed genes (DEGs) of CAP, the disease-related signaling pathways, and the protein-protein interaction (PPI) network to predict the relevant genes that may play an important role in the progression of CAP at the molecular level and reveal the molecular mechanism of CAP. Finally, we used the clinical data of the CAP patients admitted to our hospital to further verify the clinical value of the key genes identified. Our research results will provide new ideas for the clinical diagnosis and treatment of CAP. We present this article in accordance with the STREGA reporting checklist (available at https://jtd.amegroups.com/article/view/10.21037/jtd-23-592/rc).

Highlight box

Key findings

- We found that *IL7R* is the key gene related to the occurrence and development of community-acquired pneumonia (CAP).

What is known and what is new?

- Gene mutations, uncontrolled gene expression, and epigenetic changes play key roles in the development of CAP.
- Using the bioinformatics method, we explored the gene chips in the Gene Expression Omnibus (GEO) database, a public microarray database, and Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic diseases, to identify the genes related to CAP, and analyzed the clinical value of the key genes.

What is the implication, and what should change now?

- Our identification of the key gene *IL7R* and the related signaling pathways extends understanding of the pathogenesis of CAP and provides a theoretical basis for clinical targeted therapy research.

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Data sources

Gene chip data sets (GSE42830 and GSE94916) containing the data of CAP patients and normal controls were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). To be included in the study, the data sets had to meet the following criteria: (I) comprise genome-wide expression messenger RNA microarray data; (II) comprise data on CAP patients and normal controls; (III) comprise standardized or original data sets; (IV) include >3 samples.

Data processing and differential gene screening

R language (https://www.r-project.org/) was used to conduct the principal component analysis (PCA) of the data from the above 2 data sets according to the different chips and observe the distributions between the groups. The GEO2R online tool (https://www.ncbi.nlm.nih.gov/geo/ info/geo2r.html) was used to analyze the DEGs in each of the data sets, and the filter conditions to identify the DEGs were set as follows: |log fold change| ≥1 and P<0.01. However, the co-expressed DEGs from the 2 chips contained genes that were inconsistent in terms of upregulation and downregulation, and among them, co-expression differentially expressed genes from two microarrays will contain genes with inconsistent up-regulation and down-regulation. Directly conducting a biochemical analysis of all co-expression differentially expressed genes and pathways related to CAP will be mixed with the impact of false positive co-expression genes. To exclude this confounding factor, and to screen for predictive targets for clinical diagnosis and prognosis. Compared to healthy individuals, down-regulated genes may have more clinical application feasibility and research value, so we only select the down-regulated genes among the common DEGs for analysis Heat maps and volcano maps of the DEGs obtained from the 2 data sets were drawn using R language (https://www.r-project.org/). Using Venn Diagram, the differentially expressed down-regulated genes screened from the two data sets were intersected to obtain differentially expressed down-regulated genes consistent with CAP-related expression.

Gene Ontology (GO) enrichment analysis of the common DEGs and Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway enrichment analysis

The above-mentioned common DEGs were extracted, and the DAVID online database (https://david.ncifcrf.gov/) was used to conduct functional enrichment analyses of the GO and KEGG signaling pathways. An adjusted P value <0.05 was used as the threshold to identify the main enriched functions and pathways of the DEGs.

PPI network analysis of the common DEGs

The STRING (https://string-db.org/) online database was used to analyze the PPI network of the common DEGs, and the results were imported into Cytoscape software(https://cytoscape.org/) for the visualization and correlation analyses. The Molecular Complex Detection (MCODE) plug-in was used to screen out the key protein expression molecules.

CAP-related core gene screening

OMIM (https://omim.org/) was searched using the keyword "myelodysplastic syndrome" to find the genes related to CAP as reported in the literature. A Venn diagram was used to intersect the key protein molecules obtained by the MCODE analysis, and located the reported literature explaining the role of our intersection genes in the progress of CAP.

Gene set enrichment analysis (GSEA) screening of the DEGs

GSEA was conducted to screen the core DEGs of the 2 microarray data sets, and the KEGG gene set in the C2 data set was used as the preset gene set for the enrichment analysis. NOM.P (nominal P value) <0.05 and false discovery rate (FDR) q value <0.25 were set as the screening standard to identify the enrichment pathways related to the CAP disease traits in the 2 chips. The core genes that play a key role in these enrichment pathways were selected, and candidate genes related to CAP reported in the literature obtained from the OMIM database were intersected again using a Wayne diagram, and located in the reported literature explaining the association between the common

genes that intersect GSEA and OMIM and CAP.

Metagenomics next-generation sequencing (mNGS) detection of bronchial-alveolar lavage fluid (BALF)

mNGS uses BALF for genetic testing to clarify the cause of lung infection. mNGS detection is a commonly used type of next-generation sequencing (NGS) technology in clinical practice. It is characterized by massive parallel sequencing, is highly sensitive, and can simultaneously detect hundreds of pathogens. The clinical application of mNGS can improve the detection rate of infectious pathogenic bacteria in BALF and reveal the distribution of pathogenic bacteria and infection types in patients with pulmonary infections with different underlying diseases.

After the BALF was obtained, it was quickly frozen (at –80 ℃), and mNGS detection was performed within 24 h. First, glass microbeads were added to the human BALF samples to extract the nucleic acids. The constructed library was sequenced on the Nextseq 550 (Illumina, San Diego, CA, USA) sequencer. The detection range covered >24,000 pathogens, including bacteria, fungi, viruses, parasites and other pathogenic microorganisms. After removing the human sequences, the remaining data were compared against bacterial, fungal, viral, and parasite data to distinguish between colonization, contamination, and infection.

The criteria for positive mNGS results were as follows: (I) the relative third degree of pathogenic bacteria identified by mNGS was >30%; and (II) the microorganisms identified only by mNGS were considered new potential pathogens. Regardless of whether mNGS or traditional cultures were used, when microorganisms were detected and the above criteria were met, at least 2 clinicians had to distinguish colonization, contamination, and infection in combination with actual situation of the patient.

Immunohistochemistry

Immunohistochemical detection was performed on the BALF of the CAP patients included in the study. The recovered washing solution was centrifuged at 1,500 r/min for 10 minutes at 4 ℃, and the recovered supernatant was placed at –20 ℃ for cytokine detection. Then, 1 mL of PBS containing 1% bovine serum albumin (BSA) was used to resuspend cell precipitation, and 10 μL of resuspended solution was taken for cell counting. Centrifuge the remaining liquid at 4 ℃ again. After resuspension with 80–100 μL, three slides can be coated. Allow the slides to dry naturally, and then fix them in a 10% neutral formaldehyde solution for more than 10 minutes. Afterward, heat-mediated antigen repair was performed using Tris/EDTA buffer at pH 9.0, followed by blocking endogenous peroxidase and blocking with blocking solution. The blocking solution was gently shaken off. When the immunohistochemical evaluation was performed, the immunohistochemical results were scored based on the proportion of positive cells as follows: 0 (negative), 1 (<25%), 2 (25–50%), 3 (51–75%), and 4 (>75%). While the intensity of the cell staining was scored based on the proportion of positively stained cells as follows: 0 (negative or no staining), 1 (weak positive), 2 (moderate positive), and 3 (strong positive). The value obtained by multiplying the 2 scores was the final score corresponding to each sample, and the arithmetic means of these scores were then calculated to determine the high and low expression cut-off scores for the key genes.

Statistical analysis

GraphPad Prism 7.0 software (https://www.graphpad-prism.cn/) was used to draw the graphs and conduct the statistical analysis. All the data are presented as the mean ± standard deviation (SD). In the analysis of the DGEs, *t*-tests were used to determine the P values and adjusted P values, where the P values were adjusted by the FDR. A P value <0.05 indicated that the difference was statistically significant.

Results

Screening of CAP-related DEGs

The gene chip data sets (GSE42830 and GSE94916) containing the data of CAP patients and normal controls were screened from the GEO database, and then the volcano plot, PCA plot, heat map, and sample normalized box plot of the DEGs were generated according to the screening conditions (*Figure 1A*,*1B*). In these results, we focused on the downregulated DEGs in the CAP group. GSE42830 had 409 downregulated DEGs, and GSE94916 had 2984 downregulated DEGs. The downregulated DEGs of the 2 data sets were intersected by a Venn diagram, and 175 downregulated DEGs related to CAP were identified (*Figure 1C*).

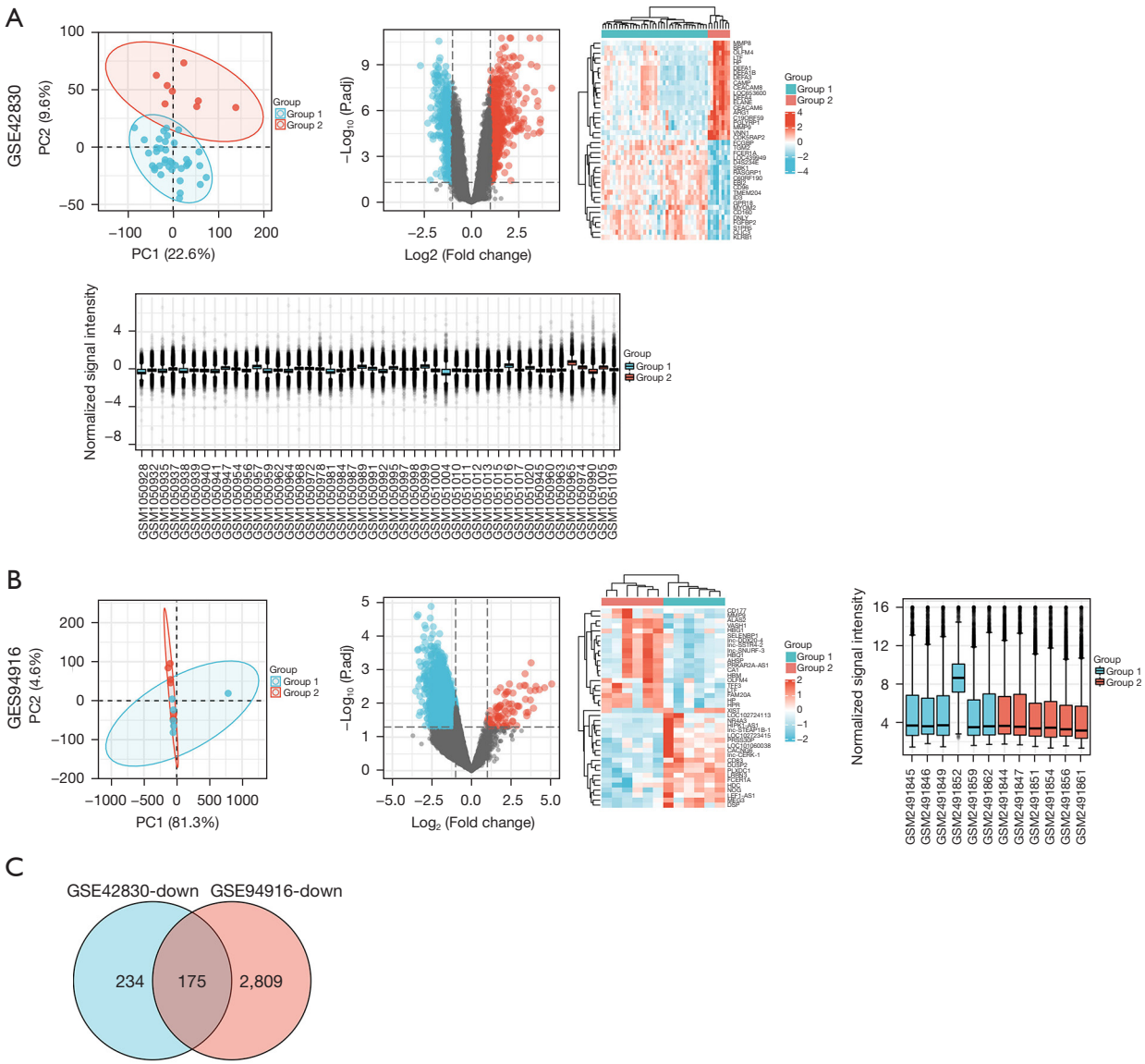

Figure 1 Screening of differential genes in community-acquired pneumonia. (A) Volcano map, principal component analysis map, heat map, and sample normalized box plot of the DEGs in the GSE42830 data set. Blue dots mean down-regulated genes, red dots mean up-regulated genes, black dots mean |log FC| <1. (B) Volcano map, principal component analysis map, heat map, and sample normalized box plot of the DEGs in the GSE94916 data set. Blue dots mean down-regulated genes, red dots mean up-regulated genes, black dots mean |log FC| <1. (C) Venn diagram of the downregulated DGEs in the 2 data sets, GSE42830 and GSE94916.

GO and KEGG signaling pathway enrichment analyses of the CAP-related common DEGs

Using the DAVID online database (https://david.ncifcrf.gov/), we conducted an enrichment analysis of the screened common DEGs with the background of Homo sapiens, and obtained the GO enrichment information. The biological processes enriched by the downregulated DEGs mainly included T cell activation, lymphocyte differentiation, intercellular adhesion regulation, T cell differentiation, the antigen receptor-mediated signaling pathways, the positive regulation of T cell activation, the T cell receptor signaling pathway, α-β T cell differentiation, and T cell selection (*Figure 2A* and *Table 1* for further details). In terms of the cellular composition, the DEGs were mainly

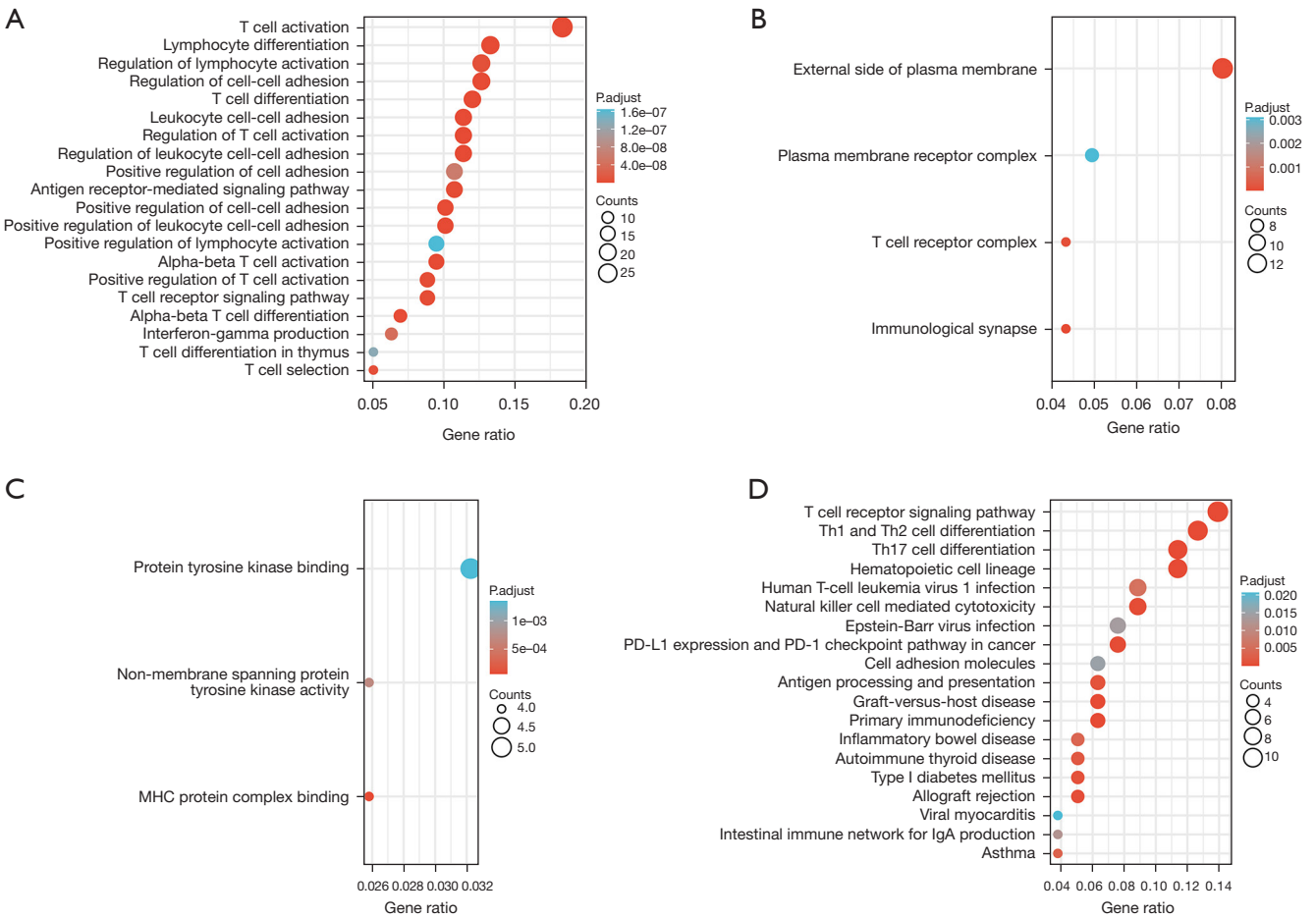

Figure 2 Gene Ontology and KEGG enrichment analysis results of the downregulated differential genes. (A) Enrichment results of the biological processes. (B) Enrichment results of the cellular components. (C) Enrichment results of the molecular functions. (D) Enrichment results of the KEGG signaling pathways. KEGG, Kyoto Encyclopedia of Genes and Genomes.

enriched outside the plasma membrane, plasma membrane receptor complex, and T cell receptor complex (*Figure 2B* and *Table 2* for further details). In terms of molecular function, DEGs mainly focus on protein tyrosine kinase, non-transmembrane protein tyrosine kinase activity, and the major histocompatibility complex protein complex (*Figure 2C* and *Table 3* for further details). The enrichment analysis of the KEGG signaling pathways showed that the downregulated DEGs were mainly related to the T cell receptor signaling pathway Th1 and Th2 cell differentiation, Th17 cell differentiation hematopoietic cell lineage, human T-cell leukemia virus 1 infection, natural-killer cell-mediated cytotoxicity, Epstein-Barr virus infection, programmed death-ligand 1 (PD-L1) expression in cancer, the programmed cell death protein 1 (PD-1) checkpoint pathway, allograft rejection, viral myocarditis, and the intestinal immune network for immunoglobulin A production (*Figure 2D* and *Table 4*).

Construction of the protein mutual aid network, module analysis of the common DEGs, and screening of the core genes

STRING (https://string-db .org/) was used to construct the PPI network for the common DEGs. The obtained results were imported into Cytoscape (https://cytoscape.org/) software for the visual analysis and screening of the PPI networks. The Network Analyzer tool in Cytoscape (https://cytoscape.org/) was used to calculate the non-directional score of each node in the PPI network and obtain the

Table 1 Results of the GO BP analysis

Ontology	ID	Description	Gene ratio	Bg ratio	P value	Adjusted P value	Q value
BP	GO:0042110	T cell activation	29/158	464/18,670	2.70e–17	2.70e–17	6.08e–14
BP	GO:0030217	T cell differentiation	19/158	240/18,670	1.88e–13	1.88e–13	2.12e–10
BP	GO:0046631	Alpha-beta T cell activation	15/158	138/18,670	7.42e–13	7.42e–13	5.57e–10
BP	GO:0030098	Lymphocyte differentiation	21/158	353/18,670	2.45e–12	2.45e–12	1.38e–09
BP	GO:1903039	Positive regulation of leukocyte cell-cell adhesion	16/158	218/18,670	5.24e–11	5.24e–11	2.36e–08
BP	GO:1903037	Regulation of leukocyte cell-cell adhesion	18/158	304/18,670	1.08e–10	1.08e–10	4.06e–08
BP	GO:0050863	Regulation of T cell activation	18/158	314/18,670	1.83e–10	1.83e–10	5.90e–08
BP	GO:0022407	Regulation of cell-cell adhesion	20/158	403/18,670	2.17e–10	2.17e–10	6.12e–08
BP	GO:0022409	Positive regulation of cell-cell adhesion	16/158	255/18,670	5.35e–10	5.35e–10	1.29e–07
BP	GO:0007159	Leukocyte cell-cell adhesion	18/158	337/18,670	5.74e–10	5.74e–10	1.29e–07
BP	GO:0046632	Alpha-beta T cell differentiation	11/158	101/18,670	9.27e–10	9.27e–10	1.90e–07
BP	GO:0050851	Antigen receptor-mediated signaling pathway	17/158	316/18,670	1.59e–09	1.59e–09	2.99e–07
BP	GO:0050852	T cell receptor signaling pathway	14/158	202/18,670	1.87e–09	1.87e–09	3.01e–07
BP	GO:0050870	Positive regulation of T cell activation	14/158	202/18,670	1.87e–09	1.87e–09	3.01e–07
BP	GO:0045058	T cell selection	8/158	47/18,670	5.23e–09	5.23e–09	7.50e–07
BP	GO:0051249	Regulation of lymphocyte activation	20/158	485/18,670	5.33e–09	5.33e–09	7.50e–07
BP	GO:0032609	Interferon-gamma production	10/158	113/18,670	4.16e–08	4.16e–08	5.51e–06
BP	GO:0045785	Positive regulation of cell adhesion	17/158	403/18,670	5.84e–08	5.84e–08	7.31e–06
BP	GO:0033077	T cell differentiation in thymus	8/158	70/18,670	1.33e–07	1.33e–07	1.58e–05
BP	GO:0051251	Positive regulation of lymphocyte activation	15/158	334/18,670	1.64e–07	1.64e–07	1.85e–05

GO, Gene Ontology; BP, biological process.

Table 2 Results of the GO CC analysis

Ontology	ID	Description	Gene ratio	Bg ratio	P value	Adjusted P value	Q value
CC	GO:0001772	Immunological synapse	7/162	36/19,717	1.52e–08	1.52e–08	3.88e–06
CC	GO:0009897	External side of plasma membrane	13/162	393/19,717	2.29e–05	2.29e–05	0.003
CC	GO:0042101	T cell receptor complex	7/162	127/19,717	8.69e–05	8.69e–05	0.007
CC	GO:0098802	Plasma membrane receptor complex	8/162	295/19,717	0.003	0.003	0.198

GO, Gene Ontology; CC, cellular component.

Table 3 Results of the GO MF analysis

Ontology	ID	Description	Gene ratio	Bg ratio	P value	Adjusted P value	Q value
MF	GO:0023023	MHC protein complex binding	4/155	25/17,697	6.20e–05	6.20e–05	0.021
MF	GO:0004715	Non-membrane spanning protein tyrosine kinase activity	4/155	46/17,697	6.94e–04	6.94e–04	0.118
MF	GO:1990782	Protein tyrosine kinase binding	5/155	93/17,697	0.001	0.001	0.153

GO, Gene Ontology; MF, molecular function; MHC, major histocompatibility complex.

Table 4 Results of the KEGG analysis

Ontology	ID	Description	Gene ratio	Bg ratio	P value	Adjusted P value	Q value
KEGG	hsa04660	T cell receptor signaling pathway	11/79	104/8,076	4.12e–09	4.12e–09	6.68e–07
KEGG	hsa04658	Th1 and Th2 cell differentiation	10/79	92/8,076	1.69e–08	1.69e–08	1.37e–06
KEGG	hsa04640	Hematopoietic cell lineage	9/79	99/8,076	4.38e–07	4.38e–07	2.37e–05
KEGG	hsa04659	Th17 cell differentiation	9/79	107/8,076	8.53e–07	8.53e–07	3.46e–05
KEGG	hsa05340	Primary immunodeficiency	5/79	38/8,076	3.07e–05	3.07e–05	9.96e–04
KEGG	hsa05332	Graft-versus-host disease	5/79	42/8,076	5.05e–05	5.05e–05	0.001
KEGG	hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	6/79	89/8,076	2.20e–04	2.20e–04	0.005
KEGG	hsa04650	Natural-killer cell-mediated cytotoxicity	7/79	131/8,076	2.76e–04	2.76e–04	0.006
KEGG	hsa05330	Allograft rejection	4/79	38/8,076	4.86e–04	4.86e–04	0.009
KEGG	hsa04940	Type I diabetes mellitus	4/79	43/8,076	7.84e–04	7.84e–04	0.013
KEGG	hsa04612	Antigen processing and presentation	5/79	78/8,076	9.53e–04	9.53e–04	0.014
KEGG	hsa05320	Autoimmune thyroid disease	4/79	53/8,076	0.002	0.002	0.023
KEGG	hsa05310	Asthma	3/79	31/8,076	0.003	0.003	0.041
KEGG	hsa05321	Inflammatory bowel disease	4/79	65/8,076	0.004	0.004	0.042
KEGG	hsa05166	Human T-cell leukemia virus 1 infection	7/79	219/8,076	0.005	0.005	0.059
KEGG	hsa04672	Intestinal immune network for IgA production	3/79	49/8,076	0.012	0.012	0.122
KEGG	hsa05169	Epstein-Barr virus infection	6/79	202/8,076	0.014	0.014	0.133
KEGG	hsa04514	Cell adhesion molecules	5/79	149/8,076	0.015	0.015	0.136
KEGG	hsa05416	Viral myocarditis	3/79	60/8,076	0.021	0.021	0.176

KEGG, Kyoto Encyclopedia of Genes and Genomes.

degree value of each node. The degree value is represented by the size of the node, the color of the node from red to green represents the neighborhood connectivity of each node from high to low, the thickness of the edge represents the combined score value of the edge, all the protein nodes are arranged an attribute circle layout, and nodes with degree values ≥4 can be found in the inner layer (*Figure 3A*).

The MCODE (https://cytoscape.org/) plug-in was used to conduct a cluster correlation analysis of the important protein molecules with the following default parameters: node score cut-off value: 0.2, K-core: 2, and maximum depth: 100. The following 5 clusters had high scores: Cluster 1 (*Figure 3B*), Cluster 2 (*Figure 3C*), Cluster 3 (*Figure 3D*), Cluster 4 (*Figure 3E*), and Cluster 5 (*Figure 3F*). We then searched the OMIM (https://www.omim.org/) online catalog of human genes and genetic diseases and found that 148 CAP-related genes have been reported in the literature, and these genes intersected with the 41 key protein molecules obtained by the MCODE analysis. A Venn diagram was used to identify the 4 candidate genes: *ICOS*, *IL7R*, *ITK*, and *ZAP70* (*Figure 3G*).

Screening of DEGs by GSEA

For the GSEA, the KEGG gene set of the C2 data set was used as the preset gene set and conducts the 2 chip data sets, respectively. The core DEGs were screened, and a NOM.P<0.05 was used as the screening criteria to identify the enriched pathways related to CAP. In total, 459 KEGG-related core genes were obtained from the GSE42830 chip, and 131 KEGG-related core genes were obtained from the GSE94916 chip. The core genes in these enriched pathways were intersected with the CAP-related genes reported in the literature retrieved from the OMIM database. In

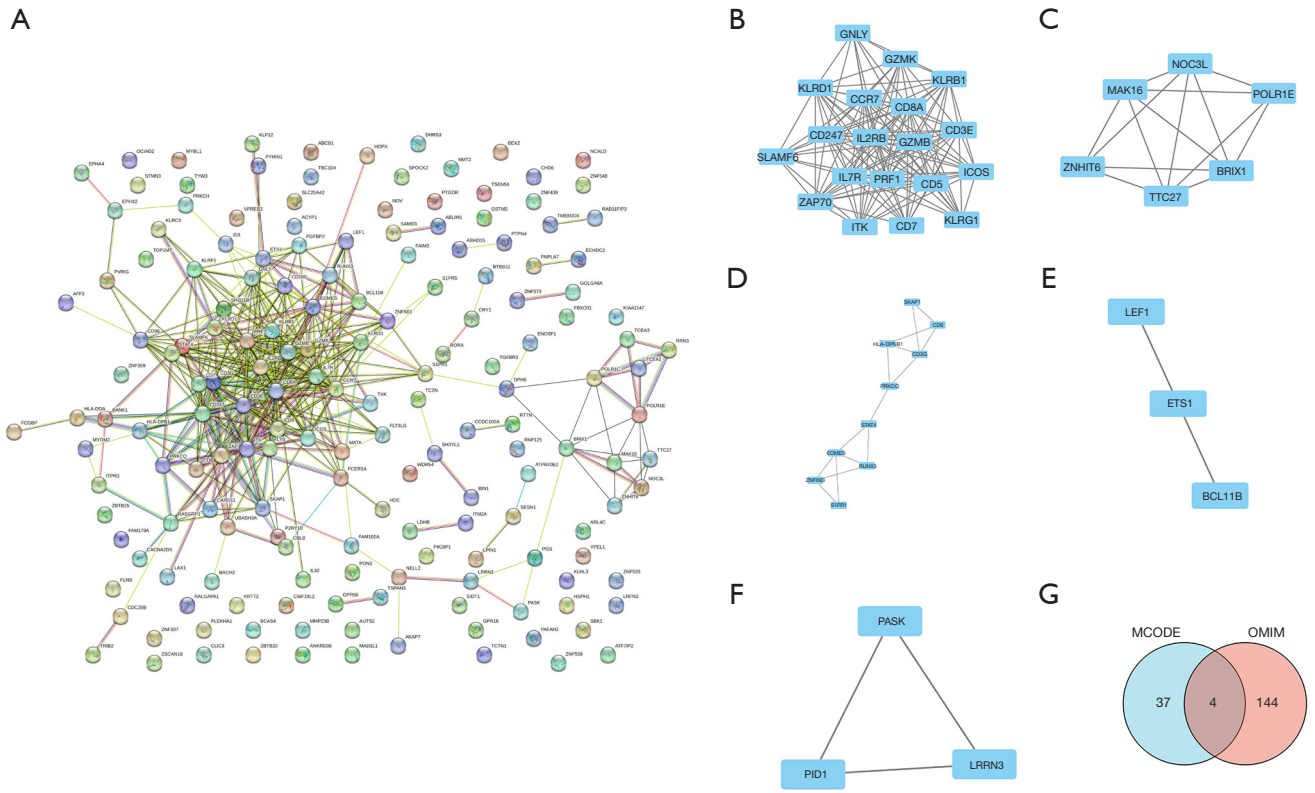

Figure 3 Protein-protein interaction network and Venn diagram of the candidate genes. (A) Protein-protein interaction network diagram; (B) cluster 1; (C) cluster 2; (D) cluster 3; (E) cluster 4; (F) cluster 5; (G) candidate gene Venn diagram.

the Venn diagram, two gene sets co-existed with OMIM, namely *IL7R* and *PIK3R1* (*Figure 4A*). We also found that the pathways involved in the KEGG pathways of these 2 genes in the GSE42830 data set were immunodeficiency and the T-cell receptor signaling pathway (*Figure 4B*), while the pathways involved in the KEGG pathways of the GSE94916 data set were cytokines and the inflammation-related pathway (*Figure 4C*). Based on the above research results, we were of the view that *IL7R* was the key gene most likely to be related to the occurrence and development of CAP. This finding may inform our cytological research in the future.

mNGS test results of BALF

Our study included 113 patients with CAP. The BLAF of these patients was analyzed by immunohistochemistry. As described above, samples with scores of <7 were considered to have low expressions of *IL7R*. The BALF of these patients was also tested by mNGS, and 113 strains of microorganisms (71 strains of bacteria, 39 strains of fungi, and 3 strains of viruses) were found. Among the bacteria, Acinetobacter was the most common, followed by Streptococcus, Klebsiella, and Pseudomonas. Among the fungi, candida was the most common. Three strains of virus were detected in the specimens, including coronavirus and cytomegalovirus. According to the level of *IL7R* expression, patients were divided into 2 groups, and the mNGS detection results of the 2 groups of patients are shown in *Tables 5,6*.

Discussion

CAP has high morbidity and fatality rates worldwide. The main symptoms of CAP are fever, chills, cough, sputum, chest pain, and shortness of breath. Generally, young and middle-aged patients can be diagnosed with typical clinical symptoms, signs, and laboratory tests and chest imaging. However, due to the irrational use of antibiotics and the change of pathogens, the phenomenon of pathogen resistance is increasing, which has led to the failure of initial empiric antimicrobial therapy, and the progression of

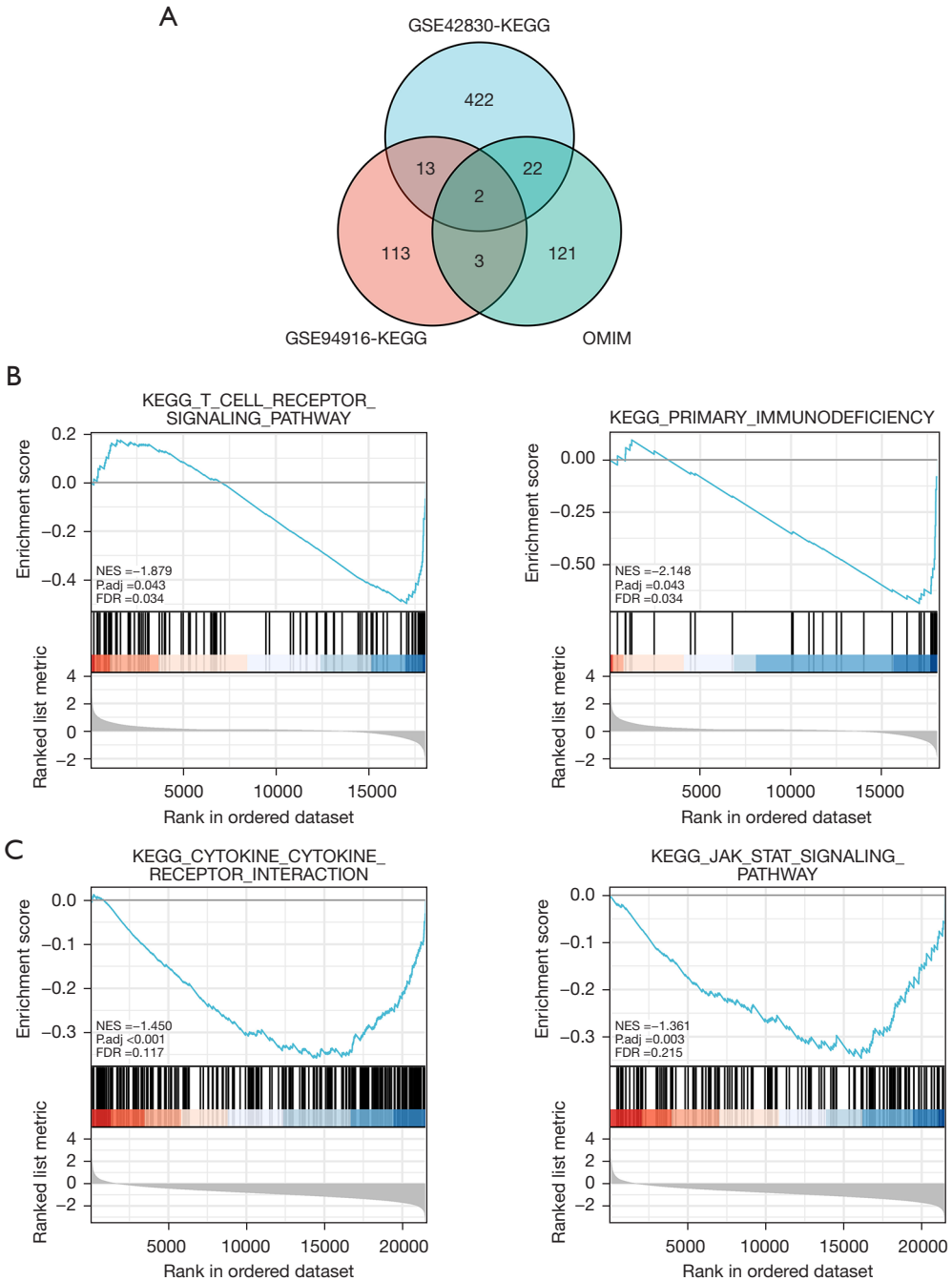

Figure 4 KEGG preset gene set enrichment results of the gene set enrichment analysis. (A) Venn diagram of the 3 data sets; (B) KEGG pathways involved in the core key genes in the GSE42830 data set; (C) KEGG pathways involved in the core key genes in the GSE94916 data set. KEGG, Kyoto Encyclopedia of Genes and Genomes; OMIM, Online Mendelian Inheritance in Man; NES, normalized enrichment score value after correction; FDR, false discovery rate.

pulmonary inflammation and exacerbation of the disease, which in turn has caused inflammation to become complex and difficult to control. The exact etiology of CAP is still unclear, and its risk factors may include cytogenetic abnormalities, gene mutations, cellular microenvironment, and immune regulation (16,17). This study retrieved CAP-

Table 5 Distribution of the pathogens detected by metagenomics next-generation sequencing in patients with high and low expressions of *IL7R*

Pathogen	High expressions of *IL7R*	Low expressions of *IL7R*
Bacteria	44	27
Acinetobacter	10	6
Streptococcus	7	5
Klebsiella	6	5
Pseudomonas	3	3
Enterococcus	4	1
Enterobacter	3	2
Staphylococcus	5	0
Achromobacter	1	1
Stenotrophomonas	1	1
Burkholderia	2	0
Micromonas	1	1
Moraxella catarrhalis	0	2
Corynebacterium striata	1	0
Fungus	26	13
Candida	20	10
Aspergillus	4	2
Pneumocystis	1	1
Cryptococcus	1	0
Virus	2	1
Coronavirus	1	1
Cytomegalovirus	1	0

Table 6 Distribution of the infection types of pathogens detected by metagenomics next-generation sequencing in patients with high and low expressions of *IL7R*

Pathogen detection	High expressions of *IL7R*	Low expressions of *IL7R*
Single pathogen infection		
Only bacteria	15	7
Only fungi	9	4
Only virus	1	0
Mixed infection		
Bacteria + Bacteria	12	11
Bacteria + Fungi	16	8
Bacteria + Virus	0	0
Bacteria + Fungi + Viruses	2	0

related data sets from the GEO database and used the bioinformatics method to analyze the data to identify the biomarkers related to CAP, and thus provide a theoretical basis for determining the molecular mechanism of CAP occurrence.

In recent years, gene chips have been widely used in disease research (18). In this study, a total of 175 common downregulated DEGs were identified by searching the CAP-related data sets in the GEO database and analyzing them using bioinformatics methods. To explore the biological functions of the DEGs, GO and KEGG enrichment analyses were performed using DAVID online software. The GO and KEGG enriched pathways included the immune microenvironment, inflammatory microenvironment, and cytokine receptors. These findings provided a reference for our subsequent cytological experiments. In this study, 5 high-scoring clusters were ultimately obtained by constructing a PPI mutual aid network and conducting a module analysis of the common DEGs. Four candidate genes were obtained after intersection with genes searched in the OMIM online catalog of human genes and genetic diseases: *ICOS*, *IL7R*, *ITK*, and *ZAP70*. Some of these genes have previously been studied in pneumonia. For example, a study showed that targeting *ICOS* molecules on T cells alleviates immune-mediated influenza pneumonia (19). Additionally, research has shown that *PD-1* and *ICOS* counter-regulates the development of tissue-resident regulatory T cells and the production of interleukin (IL)-10 in influenza (20), and sIL7R levels are significantly associated with SNPs in the *IL7R* gene in adult CAP, and increased plasma sIL7R may help identify the adult CAP patients at risk of death (21). Further, research has shown that Bruton's tyrosine kinase/ITK dual inhibitors modulate immunopathology and lymphopenia in coronavirus disease 2019 (COVID-19) treatment (22). However, to date, only *IL7R* has been the subject of relevant research in CAP.

The core genes in the GSEA enrichment pathways were intersected with the CAP-related genes reported in the literature retrieved from the OMIM database search, and there were 2 genes that overlapped in OMIM and the 2 gene sets in the Venn diagram; that is, *IL7R* and *PIK3R1*. The signaling pathways involved included immunodeficiency, the T cell receptor signaling pathway, cytokines, and the inflammation-related pathways, and these pathways have been proven to be related to the occurrence and development of CAP in previous studies (23,24). Based on the above research results, we were of the view that *IL7R* was the key gene most likely to be related to the occurrence and development of CAP.

Next, we retrospectively analyzed the clinical data of CAP patients admitted to our hospital from January 1, 2022 to November 30, 2022, and used high-throughput mNGS to detect the type of pathogenic bacteria in the BALF, and detected the expression of the key genes by liquid-based cell immunohistochemistry, and the correlation between Pathogen Types and Liquid Based Cell Immunohistochemistry was analyzed. mNGS is based on NGS technology and can directly perform the high-throughput sequencing of all the nucleic acids in clinical specimens without specific amplification or microbial cultures, so it the results of high throughput sequencing can be obtained in a short period. Therefore, the nucleic acid sequence in the sample can be obtained in a relatively short time, and then information such as the type and abundance of pathogens can be analyzed through bioinformatics interpretation. mNGS technology can be used to perform the random and unbiased full sequencing of nucleic acids, which solves the problem of early prediction and targeted detection and covers a wide range of pathogens. Viruses, bacteria, fungi, and parasites can be detected at the same time (25). Additionally, mNGS has a high sensitivity detection capability. Theoretically, as long as a characteristic sequence of a pathogenic microorganism is detected, the existence of the pathogen can be identified (26). The mNGS detected 13 kinds of bacteria, 4 kinds of fungi, and 2 kinds of viruses, and combined with the immunohistochemistry results, we found that there were relatively more genera detected in the IL7R high-expressing species. This is also consistent with the previous research results mentioned above, which also show the credibility of the key genes we identified.

Conclusions

In summary, immune and inflammatory factors may be regulated during the occurrence of CAP. Our identification of the key gene *IL7R* and the related signaling pathways extends understanding of the molecular mechanism of CAP, and provides a basis for clinical targeted therapy.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the STREGA reporting checklist. Available at https://jtd.amegroups.com/article/view/10.21037/jtd-23-592/rc

Peer Review File: Available at https://jtd.amegroups.com/article/view/10.21037/jtd-23-592/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jtd.amegroups.com/article/view/10.21037/jtd-23-592/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are 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. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Lim WS, Baudouin SV, George RC, et al. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. Thorax 2009;64 Suppl 3:iii1-55.
2. Fauci AS, Morens DM. The perpetual challenge of infectious diseases. N Engl J Med 2012;366:454-61.
3. Tsilogianni Z, Grapatsas K, Vasileios L, et al. Community-acquired pneumonia: current data. Ann Res Hasp 2017;1:25.
4. Jeon I, Jung GP, Seo HG, et al. Proportion of Aspiration Pneumonia Cases Among Patients With Community-Acquired Pneumonia: A Single-Center Study in Korea. Ann Rehabil Med 2019;43:121-8.
5. Kanwar M, Brar N, Khatib R, et al. Misdiagnosis of community-acquired pneumonia and inappropriate utilization of antibiotics: side effects of the 4-h antibiotic administration rule. Chest 2007;131:1865-9.
6. Bruns AH, Oosterheert JJ, Cucciolillo MC, et al. Cause-specific long-term mortality rates in patients recovered from community-acquired pneumonia as compared with the general Dutch population. Clin Microbiol Infect 2011;17:763-8.
7. Sun Y, Li H, Pei Z, et al. Incidence of community-acquired pneumonia in urban China: A national population-based study. Vaccine 2020;38:8362-70.
8. Ramirez JA, Wiemken TL, Peyrani P, et al. Adults Hospitalized With Pneumonia in the United States: Incidence, Epidemiology, and Mortality. Clin Infect Dis 2017;65:1806-12.
9. Hephzibah Cathryn R, Udhaya Kumar S, Younes S, et al. A review of bioinformatics tools and web servers in different microarray platforms used in cancer research. Adv Protein Chem Struct Biol 2022;131:85-164.
10. Yoo SJ, Kim HB, Choi SH, et al. Differences in the frequency of 23S rRNA gene mutations in Mycoplasma pneumoniae between children and adults with community-acquired pneumonia: clinical impact of mutations conferring macrolide resistance. Antimicrob Agents Chemother 2012;56:6393-6.
11. Waterer GW. Community-acquired pneumonia: genomics, epigenomics, transcriptomics, proteomics, and metabolomics. Semin Respir Crit Care Med 2012;33:257-65.
12. Ma X, Chen L, He Y, et al. Targeted lipidomics reveals phospholipids and lysophospholipids as biomarkers for evaluating community-acquired pneumonia. Ann Transl Med 2022;10:395.
13. Chiang TY, Yu YL, Lin CW, et al. The circulating level of MMP-9 and its ratio to TIMP-1 as a predictor of severity in patients with community-acquired pneumonia. Clin Chim Acta 2013;424:261-6.
14. Herta T, Bhattacharyya A, Rosolowski M, et al. Krueppel-Like Factor 4 Expression in Phagocytes Regulates Early Inflammatory Response and Disease Severity in Pneumococcal Pneumonia. Front Immunol 2021;12:726135.
15. Liu Y, Xu HB. Genetic polymorphisms of rs9313422 G>C and rs41297579 G>A at the promoter of TIM-1 gene contribute to the risk of community-acquired pneumonia in children. J Clin Lab Anal 2020;34:e23095.
16. Alves DW, Kennedy MT. Community-acquired pneumonia in casualty: etiology, clinical features, diagnosis, and management (or a look at the "new" in pneumonia since 2002). Curr Opin Pulm Med 2004;10:166-70.

17. Bjarnason A, Westin J, Lindh M, et al. Incidence, Etiology, and Outcomes of Community-Acquired Pneumonia: A Population-Based Study. Open Forum Infect Dis 2018;5:ofy010.
18. Afable MG 2nd, Wlodarski M, Makishima H, et al. SNP array-based karyotyping: differences and similarities between aplastic anemia and hypocellular myelodysplastic syndromes. Blood 2011;117:6876-84.
19. Sakthivel P, Gereke M, Breithaupt A, et al. Attenuation of immune-mediated influenza pneumonia by targeting the inducible co-stimulator (ICOS) molecule on T cells. PLoS One 2014;9:e100970.
20. McGee MC, Zhang T, Magazine N, et al. PD-1 and ICOS counter-regulate tissue resident regulatory T cell development and IL-10 production during flu. Front Immunol 2022;13:984476.
21. Ampuero S, Bahamonde G, Tempio F, et al. IL-7/IL7R axis dysfunction in adults with severe community-acquired pneumonia (CAP): a cross-sectional study. Sci Rep 2022;12:13145.
22. McGee MC, August A, Huang W. BTK/ITK dual inhibitors: Modulating immunopathology and lymphopenia for COVID-19 therapy. J Leukoc Biol 2021;109:49-53.
23. Trisolini R, Lazzari Agli L, Cancellieri A, et al. Bronchoalveolar lavage findings in severe community-acquired pneumonia due to Legionella pneumophila serogroup 1. Respir Med 2004;98:1222-6.
24. Terekhov IV, Bondar' SS, Khadartsev AA. The state of receptor-dependent signal pathways in the agranulocytes from the peripheral blood of the reconvalescent patients following community-acquired pneumonia under the influence of microwave radiation. Vopr Kurortol Fizioter Lech Fiz Kult 2016;93:23-8.
25. Handel AS, Muller WJ, Planet PJ. Metagenomic Next-Generation Sequencing (mNGS): SARS-CoV-2 as an Example of the Technology's Potential Pediatric Infectious Disease Applications. J Pediatric Infect Dis Soc 2021;10:S69-70.
26. Li N, Ma X, Zhou J, et al. Clinical application of metagenomic next-generation sequencing technology in the diagnosis and treatment of pulmonary infection pathogens: A prospective single-center study of 138 patients. J Clin Lab Anal 2022;36:e24498.

(English Language Editor: L. Huleatt)

Cite this article as: Zuo Y, Shen W, Chen G, Liu H, Liu N, Xu T, Pu J. Prediction of target genes in community-acquired pneumonia based on the bioinformatics method. J Thorac Dis 2023;15(5):2694-2707. doi: 10.21037/jtd-23-592