Bioinformatics analysis of new diagnostic and treatment biotargets in pulmonary tuberculosis

Yang Mo[1#], Qin Lu[3,4#], Qi Zhang[5#], Jie Chen[2], Youming Deng[2], Ke Zhang[2], Ran Tao[2], Weidong Liu[2*], Yiming Wang[3,4*]

1.Teaching and Research Section of Clinical Nursing, Xiangya Hospital of Central South University.

2.Department of Essential Surgery, Xiangya Hospital, Central South University, Changsha, Hunan, 410008, P. R. China.

3.State Key Laboratory of Chemo/Bio-Sensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan Provincial Key Laboratory of Biomacromolecular Chemical Biology, Hunan University, Changsha 410082, P. R. China.

4. GeneTalks Biotech Co., Ltd. Changsha, Hunan, 410000, P. R. China.

5. Blood Transfusion Departmen, Zibo Central Hospital , Zibo, China.

#These authors contributed equally to this manuscript.

Running title: CCDC66 improves CRC by MDM4/miR-370.

***Corresponding authors**

Yiming Wang

State Key Laboratory of Chemo/Bio-Sensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan Provincial Key Laboratory of Biomacromolecular Chemical Biology, Hunan University, Changsha 410082, P. R. China.

Email: yimin.wang@genetalks.com

Weidong Liu

Department of Essential Surgery, Xiangya Hospital, Central South University, Changsha, Hunan, 410008, P. R. China.
Email: weidong.liu@csu.edu.cn

Abstract

Background: Tuberculosis is a seriously infectious disease, and has a high morbidity and fatality rate. This study aims to explore the biomarkers of its diagnosis and treatment targets in pulmonary tuberculosis.

Method: Three gene expression profile GSE107731 was screened and downloaded for analysis. Differentially expressed genes (DEGs) were determined by GEO2R and Heml software. Then, gene set enrichment analysis (GSEA) was conducted to analyze GSE107731 date set. Moreover, a protein-protein interaction (PPI) network of the DEGs was designed. In order to find hub genes, we used Molecular Complex Detection plug-ins to further construct sub-networks. Finally, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed according to these hub genes.

Result: We obtained a total of 309 DEGs from 6 samples. And the MCODE networks provided us ten hub genes, including ACTB, RAC2, PTEN, VAV1, CD44, ACTG1, MSN, SELL, STAT5A, STAT5B. GO and KEGG enrichment analyses showed us several key pathways related to tuberculosis.

Conclusion: These pathways and hub genes could deepen the potential mechanism, diagnosis and treatment targets for tuberculosis.

Keywords: Bioinformatics analysis, Pulmonary tuberculosis, Differentially expressed genes, Hub genes, Pathways

Background

Tuberculosis is a respiratory infection resulted by mycobacterium tuberculosis, especially pulmonary tuberculosis infection[1]. Some people may not develop pulmonary tuberculosis after infection[2]. If the resistance is decreased or the allergic reaction of cell mediators is increased, clinical symptoms may be caused, such as cough, expectoration, hemoptysis, chest pain, low fever, night sweats, fatigue, weight loss, etc[3, 4]. Since tuberculosis is easily spread through the air, some patients should be isolated and treated through drugs and surgery[5, 6]. Therefore, the identification of pulmonary tuberculosis-related biomarkers is of great significance to the diagnosis and prognosis of pulmonary tuberculosis[7].

Bioinformatics is an interdisciplinary subject in which mathematical methods are applied to extract relevant biological information from vast amounts of data[8]. Its research focuses on the structure, function and relationship of macromolecules, such as nucleic acids and biological proteins, as well as their metabolism, information transmission and other life activities in the organism[9, 10]. Li L et al. conducted bioinformatics analysis on the two data sets of GSE34608 and GSE83456, and identified a total of 180 DEGs, and found that they were related to the activation of myeloid leukocytes and the production of cytokines[11]. In addition, Zhang T et al. found that CTLA4, GZMB, GZMA and PRF1 genes were highly expressed in TB patients based on bioinformatics methods[12]. More studies have concentrated on verifying TB signaling and metabolic networks, data mining, and associated biomolecular targets, as bioinformatics technology has advanced[13].

In this study, we conducted a comprehensive bioinformatics analysis on 6 samples based on the GSE107731 data set. Specifically, the biological functions of Differentially Expressed Genes (DEGs) were studied, and the hub genes were discovered through the Protein-Protein Interaction (PPI) network in order to better understand the potential molecular mechanisms in pulmonary tuberculosis and explore therapeutic targets[14, 15].

Material and methods

Acquisition of data set

Gene Expression Omnibus (GEO) is an online sequencing database that contains various gene expression data. This time, we entered the "TB" keyword in the database to search, and the family identified the GSE107731 data set. After that, the 6 samples in the data set were downloaded in the form of matrix files for analysis.

Recognition of DEGs

We divided the 6 samples in the GSE107731 data set into 3 pulmonary tuberculosis patients and 3 healthy controls. The DEGs of the two sets of samples were analyzed based on the GEO2R tool, and $P<0.05$ was used as the filter condition. After that, with the help of Heml software (http://hemi.biocuckoo.org/), the cluster distribution of DEGs in 6 samples was drawn for the next step of biological function analysis.

GSEA analysis of GSE107731 data set

Gene Set Enrichment Analysis (GSEA) is a gene set-based enrichment analysis method. Firstly, the genes were sorted, then the gene set was analyzed whether the gene set was enriched, and finally the enrichment score (ES) value of the gene set was calculated. In this study, based on the MSigDB database, we used GSEA to analyze the enrichment of all genes in the two sets of samples in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway. When $P<0.05$, the enrichment results obtained were meaningful.

Construction of PPI network

In order to further explore the genes that affected the pathogenesis of pulmonary tuberculosis, we built a PPI network based on the Search Tool for the Retrieval of Interacting Genes database (STRING; https://string-db.org/), to better understand the functions and relationships of proteins. First, we found Multiple Proteins in the STRING database, inputted the information of DEGs obtained to generate a PPI network, and then analyzed and trimmed the network with the help of Cytoscape software.

Key sub-networks and hub genes

After constructing the PPI network, we used the cytoHubba and Molecular Complex Detection (MCODE) plug-ins in the Cytoscape software to analyze the sub-networks, and further screen out the hub genes according to the degrees of genes in the network for the next step of bioinformatics analysis.

Construction and functional enrichment analysis of PPI network of hub genes

Finally, we built a PPI network based on the hub genes obtained by the cytoHubba algorithm, and analyzed the enrichment of these hub genes in the Gene Ontology (GO) term and KEGG pathways through the Database for Annotation, Visualization and Integrated Discovery database (DAVID; https://david.ncifcrf.gov/tools.jsp). Among them, GO is an important bioinformatics initiative that aims to unify the description of the characteristics of genes and gene products, including cell component (CC), molecular function (MF), and biological process (BP). KEGG is a systematic analysis of gene function, genome information database, including genome, biochemical reactions, biochemical substances, diseases and drugs, and the most commonly-used pathway information. Based on this, when the enrichment results satisfied $P<0.05$, data were considered to be meaningful.

Results

Identification of DEGs

The GSE107731 data set contained two different samples. After analyzing these samples based on GEO2R, under the condition of $P<0.05$, we obtained a total of 309 DEGs in the cluster distribution of 6 samples (Figure 1).

Results of GSEA

To explore the biological functions of the genes in the samples, we used GSEA to enrich the analysis of all genes in the KEGG pathway. According to the results of abdominal muscles in Figure 2A-F, there were 6 pathways in the abdominal muscles of the gene set in the sample, namely Dorso ventral axis formation, Drug metabolism other enzymes, Other glycan degradation, Fc gamma R mediated phagocytosis, Autoimmune thyroid disease and Leishmania infection.

PPI network of DEGs and hub genes

The constructed PPI network had 309 DEGs, including 189 nodes and 304 edges (Figure 3). Later, in order to identify the hub genes from the network, we used 12 algorithms in the cytoHubba plug-in to calculate the hub genes sub-networks. On the basis of degree value, a total of 10 hub genes were obtained, which were ACTB (degree=33), RAC2 (degree=20), PTEN (degree=19), VAV1 (degree=18), CD44 (degree=15), ACTG1 (degree=14), MSN (degree=13), SELL (degree=12), STAT5A (degree=12), STAT5B (degree=11) (Figure 4A-L).

PPI network and enrichment analysis of hub genes

Figure 5A-5D was a PPI network constructed based on 10 hub genes, including 10 nodes and 34 edges. The results of enrichment analysis exhibited that these hub genes were related to Regulation of cell size, Regulation of cellular component size, Taurine metabolic process, Fc-gamma receptor signaling pathway and Fc-gamma receptor signaling pathway involved in phagocytosis (BP), Kinase binding, Protease binding, Protein kinase binding, Phosphatidylinositol trisphosphate phosphatase activity (MF), Focal adhesion, Cell-substrate junction, Actin filament, Cytoskeleton, Polymeric cytoskeletal fiber (CC), etc. Finally, Figure 5E also displayed the top 10 KEGG pathways enriched by these hub genes, namely Leukocyte transendothelial migration, Focal adhesion, Proteoglycans in cancer, Regulation of actin cytoskeleton, Yersinia infection, Rap1 signaling pathway, Viral myocarditis, Adherens junction, Measles, Fluid shear stress and atherosclerosis. The above research results showed us the function and role of these hub genes in TB progression.

Discussion

Symptoms like expectoration, night sweats and hemoptysis all commonly exist in pulmonary TB[16]. However, asymptomatic pulmonary tuberculosis exhibits various clinical and radiological features that resemble a number of other diseases[17]. At present, if pulmonary tuberculosis disease can be detected early and receive standardized anti-tuberculosis treatment in time, tuberculosis can be cured radically[18]. This study has expanded our knowledge on the molecular mechanism in pulmonary tuberculosis and determined hub genes that can be used for the early diagnosis of tuberculosis.

Herein, the expression profile datasets GSE107731 were downloaded from the GEO database. Convenient for research, the samples were divided into two kinds of groups: 3 groups of pulmonary tuberculosis patients vs 3 group normal. We totally screened 309 DEGs from 6 samples. Next, we selected 10 hub genes from these DEGs through PPI network. These hub genes also regulated numerous biological pathways in pulmonary tuberculosis[19], such as Regulation of cell size, Taurine metabolic process, Fc-gamma receptor signaling pathway involved in phagocytosis and Fc-gamma receptor signaling pathway (BP), Kinase binding, Protease binding, Protein kinase binding, Phosphatidylinositol trisphosphate phosphatase activity (MF), Focal adhesion, Cell-substrate junction, Actin filament, Cytoskeleton, Polymeric cytoskeletal fiber (CC), KEGG pathways enriched by these hub genes, including Leukocyte transendothelial migration, Focal adhesion, Proteoglycans in cancer, Yersinia infection, Rap1 signaling pathway, Viral myocarditis, Adherens junction etc[20-22]. Further research on these findings is still needed.

Moreover, GSEA demonstrated drug metabolism other enzymes and other glycan degradation pathway were crucial parts in the pulmonary tuberculosis. In the context of infectious diseases such as tuberculosis, experiments by Tadahiro Kumagai et al. proved that mycobacterium tuberculosis infection directly affected protein glycosylation in mouse models[23]. Mycobacterium tuberculosis infection can increase serum IgM levels and induce changes in glycosylation, which can be used as a biological target for disease diagnosis. Lipoarabinomannan (LAM) has a significant impact on the innate immune system[24]. Glycan-binding receptors initiate a signal response, which usually causes macrophages and dendritic cells to activate multiple antibacterial mechanisms. The immunological response to LAM and the possible use of LAM's mannose-capped arabinan moiety are candidate biomarkers for anti-tuberculosis vaccines[25]. Throughout the process of tuberculosis, more attention should be paid to autoimmune thyroid disease and leishmania infection pathways, because they affect the phagocytosis of macrophages[26]. Phagocytosis is the oldest and one of the most basic defense mechanisms of organisms[27]. There is no specificity for the object to be eliminated, which is called non-specific immunity in immunology. The study by Z A Malik et al. showed that the phagocytosis of complement receptor (CR)-mediated mycobacterium tuberculosis by macrophages was minimal, indicating that mycobacterium tuberculosis interfered with macrophages to kill harmful microorganisms[28].

It is worth noting that we establish a sub-network from the PPI network and identified hub genes. They are ACTB, RAC2, PTEN, VAV1, CD44, ACTG1, MSN, SELL, STAT5A, STAT5B respectively. Some of these genes are associated with tuberculosis, like RAC2, PTEN, CD44, STAT5A and STAT5B. RAS-related C3 botulinum toxin substrate 2 (RAC2) is a member of the RHO subclass of RAS superfamily GTPases[29]. However, RAC2 mutations can cause RAC2 dysfunction. The experiments of Megan E Arrington et al. showed that promoting RAC2 overactivation, changing GEF specificity and impairing GAP function, but retaining key effector interactions, could lead to RAC2 mutations and promote immune dysfunction[30]. PTEN, a lipid phosphatase, is a tumor suppressor that is often altered in human malignancies. The PTEN-PI3K pathway has been shown to control a variety of cellular functions in addition to tumor suppression, including cell resistance to infection. The PTEN signal infects non-cancer and cancer cells by regulating a variety of intracellular mycobacterial pathogens. The pathway regulated by PTEN plays a key role in pathogen infection, and PTEN is also our biological target for studying tuberculosis. Interestingly, CD44 is an adhesion molecule involved in inflammation and connects to the actin cytoskeleton. The article by Jaklien C Leemans et al. clarified that CD44 was a new Mphi binding site for mycobacterium tuberculosis, which played a role in protective immunity of mycobacterial phagocytosis, recruitment of Mphi and tuberculosis[31]. The proteins encoded by STAT5A and STAT5B genes are members of the STAT transcription factor family. In the experiment of Yaoqin Yuan et al., STAT5A and STAT5B were used for qRT-PCR detection[32]. The results proved to help reduce the uptake of BCG in macrophages and regulate the inflammatory response of macrophages. This indicates that members of the STAT family are potential targets for tuberculosis treatment.

In conclusion, we perform a systematical bioinformatics analysis in the tumorigenesis of pulmonary tuberculosis. We obtained and studied pertinent hub genes and pathways. Knowledge on the molecular mechanism, diagnosis and treatment targets for pulmonary tuberculosis will be deepened.

Author Contribution

Conception and design: Yang Mo, Qin Lu and Qi Zhang

Development of methodology: Jie Chen

Sample collection: Youming Deng

Analysis and interpretation of data: Ke Zhang and Ran tao

Writing, review, and/or revision of the manuscript: Weidong Liu, and Yimin Wang

Conflict of interest

None.

Fund

None.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the data during the current study were available from the corresponding author on reasonable request.

Figure 1. Heat map. The cluster distribution of DEGs in 6 samples.

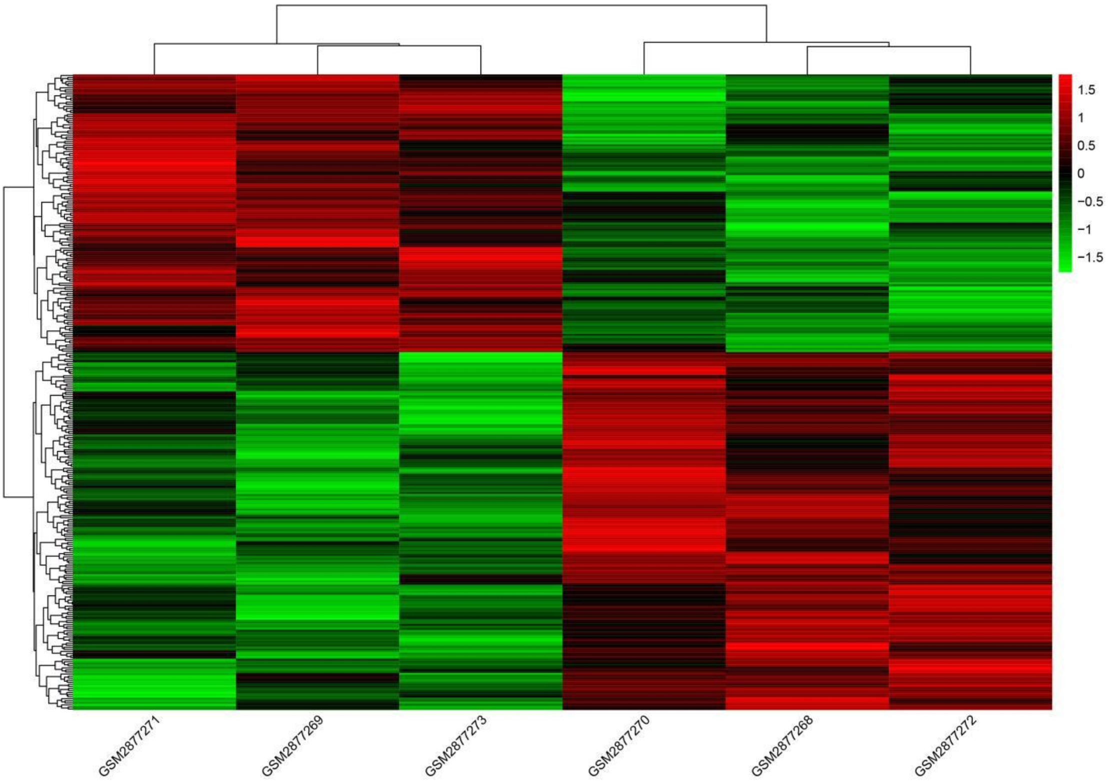


Figure 2. GSEA analysis of KEGG pathway. (A) Dorso ventral axis formation. (B) Drug metabolism other enzymes. (C) Other glycan degradation. (D) Fc gamma R mediated phagocytosis. (E) Autoimmune thyroid disease. (F) Leishmania infection.

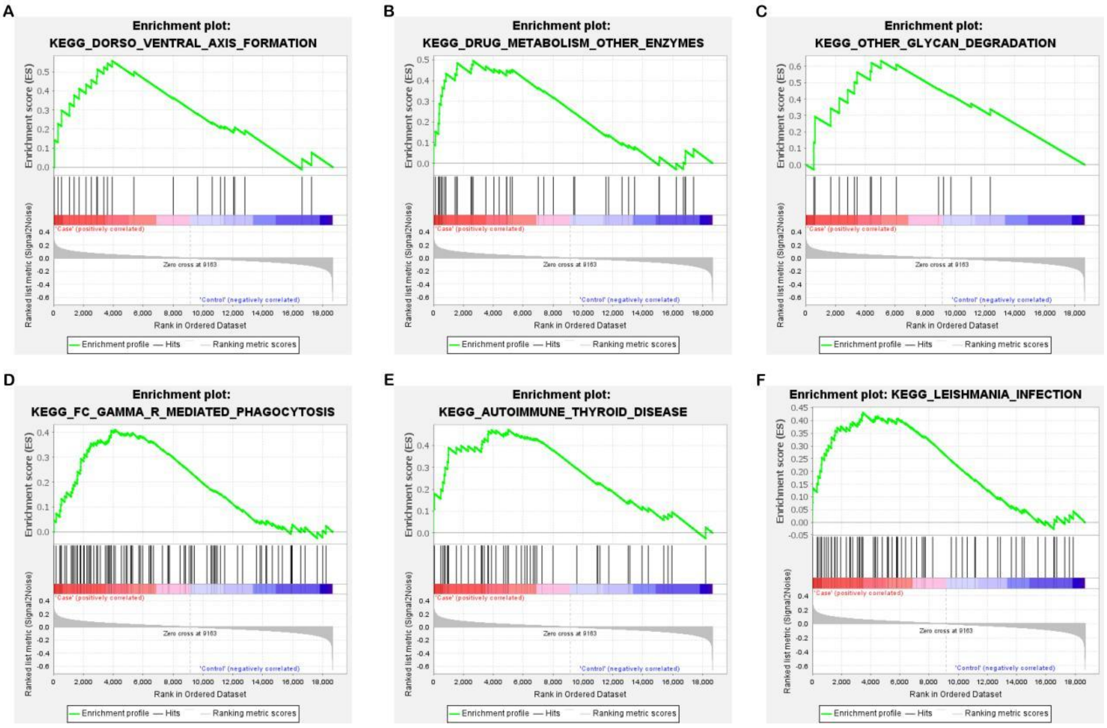


Figure 3. PPI network. The red nodes are genes, and the edges are the connectivity between genes.

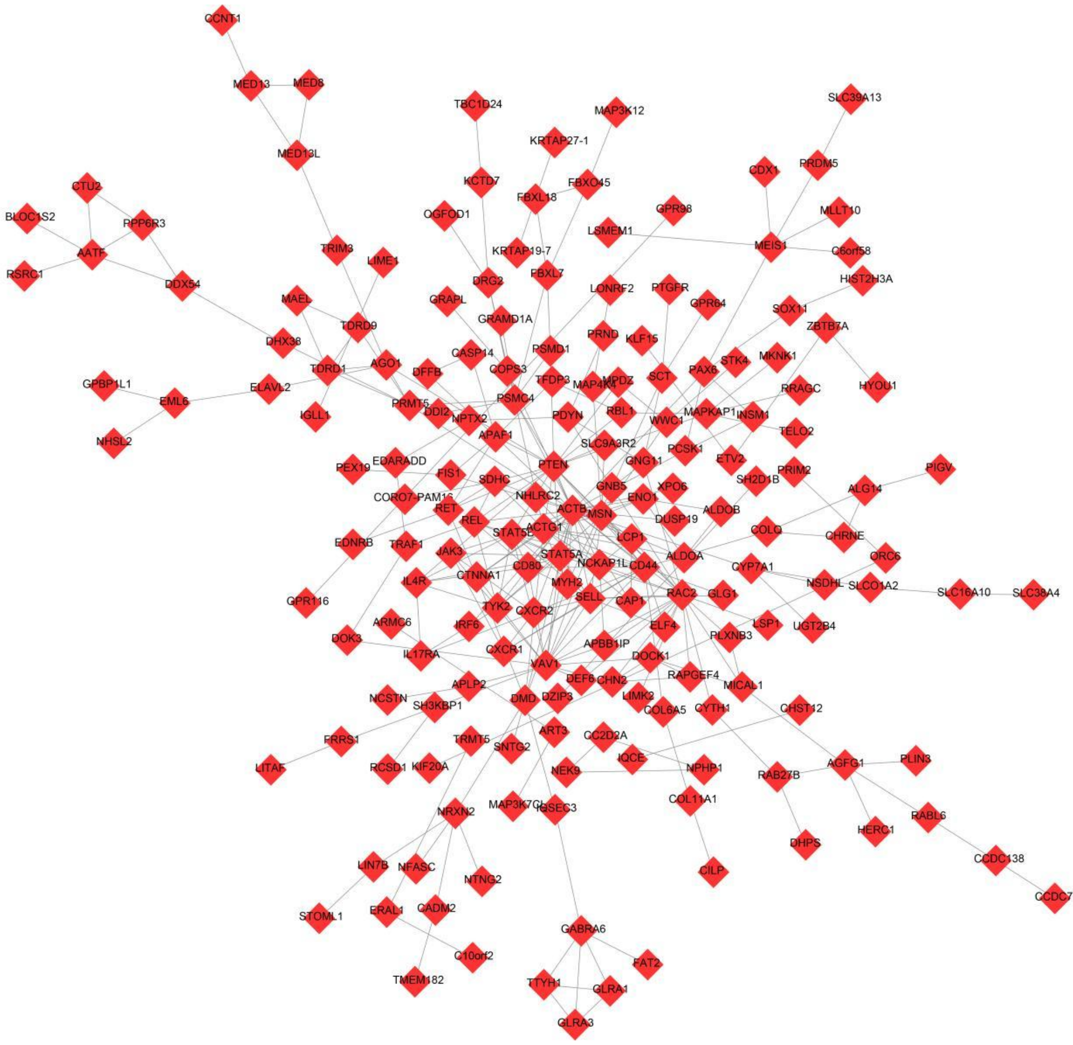


Figure 4. High degree hub genes in TB calculated based on cytoHubba. (A) MCC. (B) DMNC. (C) MNC. (D) DEGREE. (E) EPC. (F) Bottleneck. (G) Eccentricity. (H) Closeness. (I) Radiality. (J) Betweenness. (K) Stress. (L) Clustering-Coefficient.

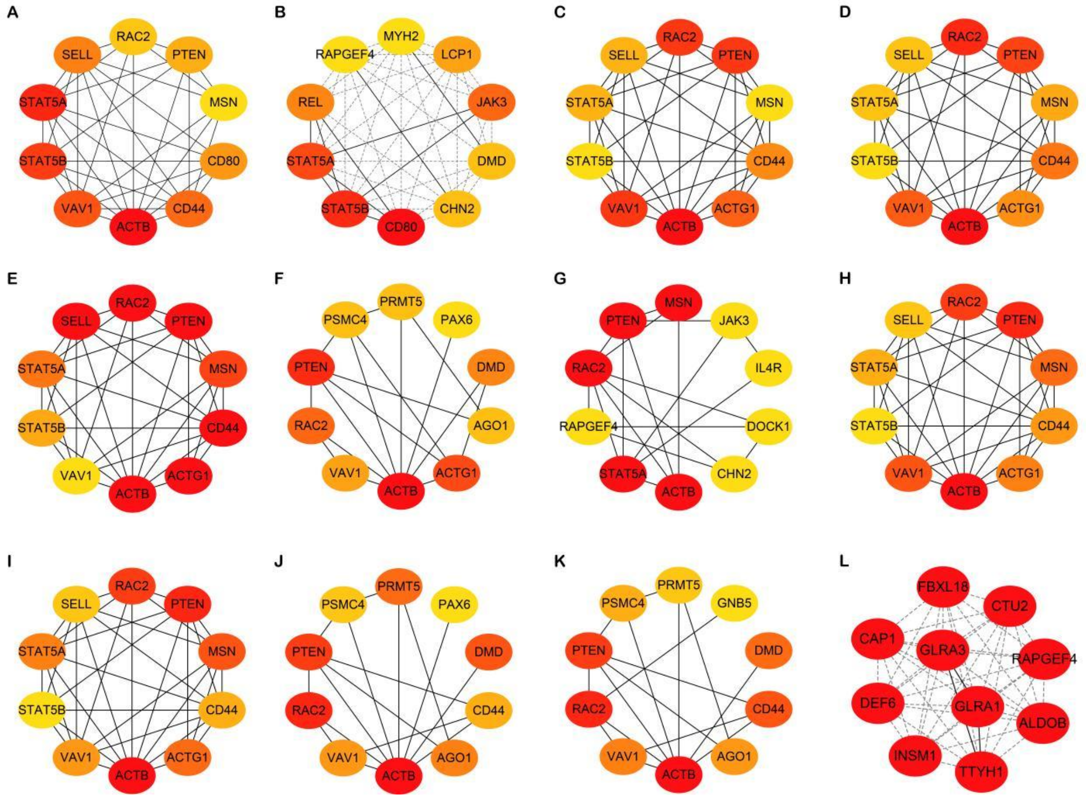
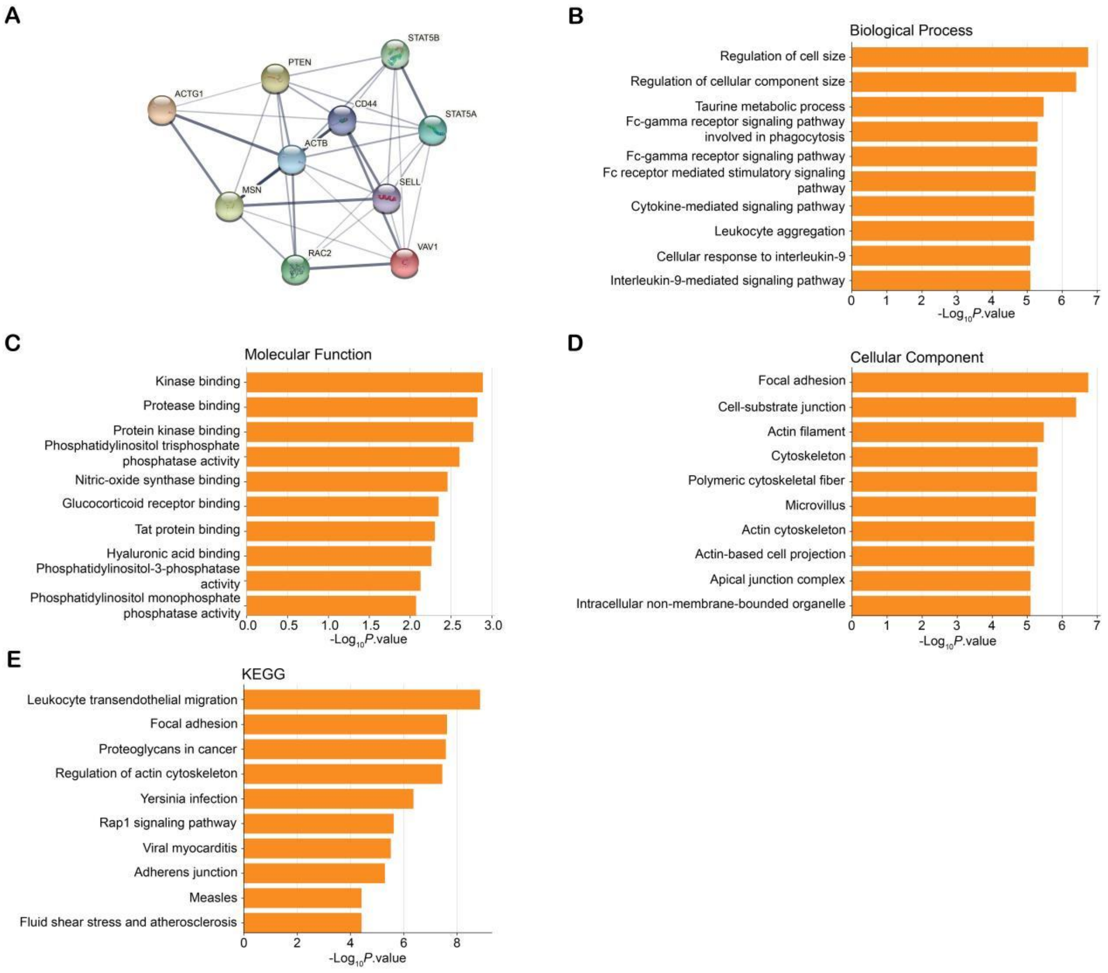


Figure 5. PPI network and functional enrichment analysis of hub genes. (A)PPI network, and the highest confidence is 0.788. (B) BP. (C) MF. (D) CC. (E) Top 10 enriched KEGG pathways.



Reference

1. Cardona, P.J., *Pathogenesis of tuberculosis and other mycobacteriosis.* Enferm Infecc Microbiol Clin (Engl Ed), 2018. **36**(1): p. 38-46.

2. Koenig, S.P. and J. Furin, *Update in Tuberculosis/Pulmonary Infections 2015.* Am J Respir Crit Care Med, 2016. **194**(2): p. 142-6.

3. Chaudhary, P., B. Chaudhary, and C.K. Munjewar, *Parotid tuberculosis.* Indian J Tuberc, 2017. **64**(3): p. 161-166.

4. Finkelman, F.D., M.V. Khodoun, and R. Strait, *Human IgE-independent systemic anaphylaxis.* J Allergy Clin Immunol, 2016. **137**(6): p. 1674-1680.

5. Brett, K., C. Dulong, and M. Severn, in *Drug-Resistant Tuberculosis: A Review of the Guidelines*. 2020: Ottawa (ON).

6. Jun, H.J., et al., *Nontuberculous mycobacteria isolated during the treatment of pulmonary tuberculosis.* Respir Med, 2009. **103**(12): p. 1936-40.

7. Morrison, H. and H. McShane, *Local Pulmonary Immunological Biomarkers in Tuberculosis.* Front Immunol, 2021. **12**: p. 640916.

8. Wooller, S.K., et al., *Bioinformatics in translational drug discovery.* Biosci Rep, 2017. **37**(4).

9. Chen, C., H. Huang, and C.H. Wu, *Protein Bioinformatics Databases and Resources.* Methods Mol Biol, 2017. **1558**: p. 3-39.

10. Alkhnbashi, O.S., et al., *CRISPR-Cas bioinformatics.* Methods, 2020. **172**: p. 3-11.

11. Li, L., et al., *Screening and identification of key biomarkers in hepatocellular carcinoma: Evidence from bioinformatic analysis.* Oncol Rep, 2017. **38**(5): p. 2607-2618.

12. Zhang, T., G. Rao, and X. Gao, *Identification of Hub Genes in Tuberculosis via Bioinformatics Analysis.* Comput Math Methods Med, 2021. **2021**: p. 8159879.

13. Walzl, G., et al., *Tuberculosis: advances and challenges in development of new diagnostics and biomarkers.* Lancet Infect Dis, 2018. **18**(7): p. e199-e210.

14. Athanasios, A., et al., *Protein-Protein Interaction (PPI) Network: Recent Advances in Drug Discovery.* Curr Drug Metab, 2017. **18**(1): p. 5-10.

15. Khan, S.A., et al., *Differentially and Co-expressed Genes in Embryo, Germ-Line and Somatic Tissues of Tribolium castaneum.* G3 (Bethesda), 2019. **9**(7): p. 2363-2373.

16. Pai, M., M.P. Nicol, and C.C. Boehme, *Tuberculosis Diagnostics: State of the Art and Future Directions.* Microbiol Spectr, 2016. **4**(5).

17. Akram, S.M. and J. Koirala, *Histoplasmosis*, in *StatPearls*. 2021: Treasure Island (FL).

18. Cardona, P.J., M. Catala, and C. Prats, *Origin of tuberculosis in the Paleolithic predicts unprecedented population growth and female resistance.* Sci Rep, 2020. **10**(1): p. 42.

19. Cao, H., et al., *Hub genes and gene functions associated with postmenopausal osteoporosis predicted by an integrated method.* Exp Ther Med, 2019. **17**(2): p. 1262-1267.

20. Li, L., et al., *Gene network in pulmonary tuberculosis based on bioinformatic analysis.* BMC Infect Dis, 2020. **20**(1): p. 612.

21. Zhao, M., et al., *Identification of Biomarkers for Sarcoidosis and Tuberculosis of the Lung Using Systematic and Integrated Analysis.* Med Sci Monit, 2020. **26**: p. e925438.

22. Jin, B., et al., *Identifying hub genes and dysregulated pathways in hepatocellular carcinoma.* Eur Rev Med Pharmacol Sci, 2015. **19**(4): p. 592-601.

23. Kumagai, T., et al., *Serum IgM Glycosylation Associated with Tuberculosis Infection in Mice.* mSphere, 2019. **4**(2).

24. Correia-Neves, M., et al., *Lipoarabinomannan in Active and Passive Protection Against Tuberculosis.* Front Immunol, 2019. **10**: p. 1968.

25. Wang, L., et al., *Synthesis and Immunological Comparison of Differently Linked Lipoarabinomannan Oligosaccharide-Monophosphoryl Lipid A Conjugates as Antituberculosis Vaccines.* J Org Chem, 2017. **82**(23): p. 12085-12096.

26. Khan, A., et al., *Macrophage heterogeneity and plasticity in tuberculosis.* J Leukoc Biol, 2019. **106**(2): p. 275-282.

27. Brown, G.C. and J.J. Neher, *Microglial phagocytosis of live neurons.* Nat Rev Neurosci, 2014. **15**(4): p. 209-16.

28. Malik, Z.A., G.M. Denning, and D.J. Kusner, *Inhibition of Ca(2+) signaling by Mycobacterium tuberculosis is associated with reduced phagosome-lysosome fusion and increased survival within human macrophages.* J Exp Med, 2000. **191**(2): p. 287-302.

29. Hu, W., et al., *Overexpression of Ras-Related C3 Botulinum Toxin Substrate 2 Radiosensitizes Melanoma Cells In Vitro and In Vivo.* Oxid Med Cell Longev, 2019. **2019**: p. 5254798.

30. Arrington, M.E., et al., *The molecular basis for immune dysregulation by the hyperactivated E62K mutant of the GTPase RAC2.* J Biol Chem, 2020. **295**(34): p. 12130-12142.

31. Leemans, J.C., et al., *CD44 is a macrophage binding site for Mycobacterium tuberculosis that mediates macrophage recruitment and protective immunity against tuberculosis.* J Clin Invest, 2003. **111**(5): p. 681-9.

32. Yuan, Y., et al., *Upregulation of miR-196b-5p attenuates BCG uptake via targeting SOCS3 and activating STAT3 in macrophages from patients with long-term cigarette smoking-related active pulmonary tuberculosis.* J Transl Med, 2018. **16**(1): p. 284.