Identification of differentially expressed genes and hub genes of human hosts with tuberculosis through an integrated bioinformatics strategy

Peng Yue
Kunming Medical University
Yan Dong
Kunming Medical University
Xin Xu
Kunming Medical University
Yu Zhang
Kunming Medical University
Jing Kong
Kunming Medical University
Jingjing Chen
Kunming Medical University
Yuxin Fan
Kunming Medical University
Meixiao Liu
Kunming Medical University
Yuan He
Kunming Medical University
Wenjing Cao
Kunming Medical University
Shiyuan Wen
Kunming Medical University
Binxue Li
Kunming Medical University
Lisha Luo
Kunming Medical University
Taigui Chen
Kunming Medical University
Lianbao Li
Kunming Medical University
Feng Wang

Kunming Medical University
Guozhong Zhou
Kunming Medical University
Suyi Luo
Kunming Medical University
Aihua Liu
Kunming Medical University
Fukai Bao (baofukai@kmmu.edu.cn)
Kunming Medical University

Research Article

Keywords: Tuberculosis, Mycobacterium tuberculosis, GEO dataset, functional enrichment, protein–protein interaction, miRNA–hub gene network, bioinformatics

Posted Date: May 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-466207/v1

Abstract

Background: Tuberculosis is a chronic infectious disease caused by *Mycobacterium tuberculosis*. Until now, molecular mechanisms underlying the occurrence, development and prognosis of tuberculosis have not been fully characterized. The aim of the study was to identify hub genes involved in tuberculosis.

Methods: We used four microarray datasets (GSE51029, GSE52819, GSE54992, and GSE65517) from the Gene Expression Omnibus (GEO) and GEO2R software to identify differentially expressed genes (DEGs) between samples from humans infected with *M. tuberculosis* and a healthy control group (using cutoffs of LogFC > 1 and *p* value < 0.05). DEGs shared by the four microarray datasets were further identified. Next, we carried out functional enrichment analysis using the Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG); Then, the host hub genes with a relatively high number of connections to other DEGs, were identified by Cytoscape. Other bioinformatics methods are also performed, including protein–protein interaction (PPI) network analysis and construction of miRNA–hub gene networks and transcription factors (TF)–hub gene networks. Finally, the expression of the host hub genes was verified using the reverse transcription polymerase chain reaction(RT–PCR).

Results: A total of 46 DEGs were identified. GO analysis showed that the biological functions of DEGs were mainly in immune response regulation, cytokine/chemokine activity, and receptor ligand activity. DEGs were significantly enriched in membrane rafts, the mitochondrial outer membrane, cytoplasmic vesicle cavities, and nuclear chromatin. KEGG enrichment analysis showed involvement of the genes in the NOD-like receptor and toll-like receptor signaling pathways. Five highly differentially expressed hub genes – STAT1, TLR7, CXCL8, CCR2, and CCL20 – were identified. Finally, based on NetworkAnalyst's database, we constructed miRNA–hub gene networks and TF–hub gene networks.

Conclusions: In summary, bioinformatics analyses were used to identify DEGs to find potential biomarkers that may be associated with tuberculosis. This study provides a set of candidate DEGs and five important hub genes that can be potential for the early detection, prognostic determination, risk assessment, and targeted therapy of tuberculosis.

1. Background

Tuberculosis (TB) is a chronic infectious disease caused by *Mycobacterium tuberculosis*, spread through the inhalation of droplets from the coughs or sneezes of an infected individual.TB can involved in many organs, but pulmonary TB is the most common infection. According to the World Health Organization (WHO), an estimated 10 million people worldwide were infected with TB in 2020, of which 7.1 million were newly diagnosed with TB(1). The incidence of TB varies from less than five to more than 500 cases per 100,000 people per year, and TB remains a deadly disease even in developing countries with well-established healthcare systems.

Infection with *M. tuberculosis* causes clinical signs and symptoms when host defense is reduced or cell-mediated allergy is increased(2). Respiratory symptoms include cough, sputum, hemoptysis, chest pain,

and varying degrees of chest tightness or shortness of breath. However, these symptoms are relatively nonspecific, and cannot be used to definitively diagnose TB. Sputum smear microscopy, bacterial culture and *M. tuberculosis* isolation are the most traditiona and classical diagnostic tools for TB. However, these tools need a lot of time, and their accuracy is not high, it is difficult to realize the early diagnosis and effective treatment of TB patients(3). The tuberculin skin test (TST) and the interferon (IFN) –γ release assay (IGRA)are widely used in the diagnosis of *M. tuberculosis* infection(4). However, TST and IGRA are essentially unable to distinguish between active and latent TB infections(5). As an intracellular pathogen, *M. tuberculosis* largely depends on its ability to destroy the host's macrophage innate immune defense system(6). Therefore, there is an urgent need for potential biological markers for efficient diagnosis and treatment, and bioinformatics research on macrophages stimulated by *M. tuberculosis* is particularly important.

Microarray technology has rapidly developed in recent years, and is widely used to more accurately diagnose various diseases, compare gene expression levels, predict disease progression, and provide prognosis evaluation(7, 8). As a result, a large amount of genes expression microarray data have been published on public database platforms, including genomic data on TB; these databases could be integrated to identify the molecular mechanisms of disease. This novel approach could significantly improve molecular disease prediction and reveal opportunities for drug-based molecular targeting and molecular therapy(9). It could not only provide new insights into the molecular and cellular processes involved in the pathogenesis of TB, but also establish much–needed rapid, sensitive, and effective methods to better diagnose and treat TB(10).

With the development of genomics technology, there is a large amount of data in the field of TB research(11). So far, many related studies have explored potential biomarkers of TB using bioinformatics, but the efficiency and accuracy of diagnosis and treatment of the disease remain unsatisfactory. To further explore sensitive and specific TB biomarkers, we screened the DEGs between the *M. tuberculosis* infection groups and the healthy control groups in four separate GEO datasets, and then performed GO function enrichment analysis, KEGG pathway enrichment analysis and PPI network construction and module analysis. The results of this study may help to explore potential targets for the diagnosis and treatment of TB.

2. Materials And Methods

2.1 Microarray dataset

The GEO database (http://www.ncbi.nlm.nih.gov/geo) is a free public genomics database containing a variety of data including microarray and next-generation sequencing data. We used the following keywords and medical subject heading (MeSH) terms to specifically identify human tuberculosis datasets in the GEO database: ("tuberculosis" [MeSH Terms] OR tuberculosis[All Fields]) AND "Homo sapiens" [Organism]. The inclusion criteria of gene expression profile are as follows: i)the datasets must include samples from both peripheral blood mononuclear cells infected with *M. tuberculosis* and non-

infected controls; ii) the sample size must be six or greater; iii each sample must be supported by a extremely large number of data.

Based on the above search results, we obtained four microarray datasets from the GEO database(GSE51029,GSE52819 GSE54992 and GSE65517), the platform of GSE51029 is GPL4133 Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Feature Number version), the platform of GSE52819 is GPL6244 [HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version],the platform of GSE54992 is GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array, the platform of GSE65517 is GPL10558 Illumina HumanHT-12 V4.0 expression beadchip. GSE51029 dataset includes 27 samples infected with *M. tuberculosis* and 27 normal samples. GSE52819 includes 3 samples infected with *M. tuberculosis* and 3 normal samples. GSE54992 includes 9 samples infected with *M. tuberculosis* and 6 normal samples. GSE65517 includes 3 samples infected with *M. tuberculosis* and 3 normal samples. The microarray datasets information is shown in Table 1.

Table 1
Details of the Gene Expression Omnibus (GEO) microarray datasets used in this study.

GEO profile	Source	Case	Control	Platform	Annotation platform
GSE51029	Tuberculosis	27	27	GPL4133	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
GSE52819	Tuberculosis	3	3	GPL6244	Affymetrix Human Gene 1.0 ST Array
GSE54992	Tuberculosis	9	6	GPL570	Affymetrix Human Genome U133 Plus 2.0 Array
GSE65517	Tuberculosis	3	3	GPL10558	Illumina Human HT-12 V4.0 expression beadchip

2.2 Identifying differentially expressed genes

The interactive network tool GEO2R is used to analyze the gene expression data of the microarray and find DEGs(12). In this study, the selected datasets of *M. tuberculosis* infection groups and healthy control groups were first analyzed by GEO2R. Subsequently, the analysis results are downloaded in Microsoft Excel format, and the genes that meet the following conditions: |log FC| > 1 and p values < 0.05,was considered DEGs(13). Finally, we used the enrichment analysis tool FunRich (version 3.1.3) to show the intersection of DEGs. In addition, we used the R language tool (version 4.0.3) and R Studio to draw a heatmap and a correlation circle map for the DEGs. The gene expression matrix data used to draw these maps were derived from the GSE51029 dataset from the GEO database.

2.3 Function and pathway enrichment analysis of DEGs

We use Metascape online software for GO analysis and KEGG pathway enrichment analysis(14), and to further explore the primary biological functions of the identified DEGs through functional enrichment

analysis based on GO and KEGG databases(15). The purpose of GO analysis is to identify the key biological characteristics of genes, gene products and sequences, including biological processes (BP), cell components (CC) and molecular functions (MF)(16). KEGG pathway enrichment analysis was performed to provide a complete set of biologically interpreted genome sequence and protein interaction network information(17).The CC, BP and MF categories and the KEGG pathway were classified and presented in the form of bubble charts. These bubble charts were drawn based on the p value using the ggplot2 function in the R software package, using a cutoff value for statistical significance of $p < 0.05$.

2.4 Protein–protein interaction (PPI) network construction and module analysis

The STRING database (http: // www. string-db. org/) is an online tool used to identify and predict physical and functional interactions between genes or proteins(18). These interactions include physical and functional associations, data mainly come from computational predictions, high-throughput experiments, automatic text mining and co-expression networks(19). We mapped the DEGs identified in our analysis to the PPI network from the STRING database with a threshold interaction score > 0.4 to build a PPI analysis network for these DEGs. Analysis of DEGs in the context of protein interactions can help clarify the biochemical complexes or signal transduction components that control biological output(20), and PPI analysis is very important to explain the underlying molecular mechanisms of key cell activities in pathogenicity(10).

Next, we used Cytoscape software (version 3.7.2) to visualize the PPI network of DEGs. Each node in the network represents a gene, protein or molecule. The connections between nodes represent the interaction of these biological molecules and can be used to identify genes that are differentially expressed Interactions between encoded proteins and signal pathway relationships(9). Subsequently, the modules of the PPI network were screened through Metascape online tool and the MCODE plugin in Cytoscape software (21).The default parameters were as follows: degree cutoff = 2, node score cutoff = 0.2, k-score = 2, and maximum. depth = 100. As another plugin of Cytoscape(22), Cytohubba studies important nodes in the network through 11 methods, of which MCC has shown satisfactory performance, and the first five genes are selected as the key hub genes.

2.5 Construction of miRNA–hub gene networks and TF–hub gene networks

NetworkAnalyst (https: //www. networkanalyst. ca/) is a comprehensive network visualization analytical platform for gene expression analysis(23). In order to construct miRNA or TF regulatory network that interacts with hub genes in the post–transcription stage(24, 25), we applied NetworkAnalyst to integrate miRNA databases(26). In the study, the targeted miRNA–hub genes were defined according to the positive results of ≥ 3 miRNA–target predicting databases, including TargetScan, miRanda, PicTar, and PITA. The targeted TF–hub genes were defined according to the positive results of databases ENCODE (http://cistrome.org/BETA/)(27, 28). Finally, we visualized target miRNA–hub gene and TF–hub gene networks by employing Cytoscape software.

2.6 Confirmation of gene expression using RT-PCR

We selected the immune cell line THP-1 to confirm gene expression in a RT-PCR assay, as it has previously been successfully used to investigate TB immune defense, antigen presentation, and phagocytosis(29). During the stimulation phase of cell culture, the standard strain *M. tuberculosis* H37Rv was used to infect the macrophages induced and differentiated from human THP-1 monocytes *in vitro*(30), Trizol reagent was used to extract total RNA from cells at 24h, 48h, and 72h, respectively.

We used the PrimeScript RT kit (Takara, Dalian, China) according to the manufacturer's protocol to reverse-transcribe total RNA into complementary DNA (cDNA) strands, and used the company's SYBR® Premix Ex Taq II (2×) to perform RT-PCR using the CFX96 PCR detection system (BioRad, California, USA). The configuration system and sample addition operations are performed on ice, and two replicate wells are made for detecting the target genes and internal reference gene of each sample. The reaction volume of 25 μL contains 12.5 μL SYBR® Premix Ex Taq II (Tli RNaseH Plus) (2×), 1 μL forward primer, 1 μL reverse primer, 1 μL cDNA template, and 8.5 μL RNase-free dH2O. The amplification conditions were as follows: 95°C for 15 s, then 40 cycles, in which the denaturation process at 95°C lasts for 5 s, and the annealing process at 58.5°C (β-actin) lasts for 30 s. SYBR® TB Green Premix PCR mix and primers specific to our target gene are used to effectively amplify the target region. The $2^{-\triangle\triangle CT}$ method was used to calculate gene relative expression and perform statistical analysis(31, 32). The primer sequences were shown in Table 2, and the β-actin gene was used as an internal reference gene.

Table 2
Primer sequences used in reverse transcription polymerase chain reaction (RT-PCR).

Gene	Forward primer	Reverse primer	Purification method
STAT1	ATGCTGGCACCAGAACGAATGAG	TCACCACAACGGGCAGAGAGG	PAGE
TLR7	ACCAACTGACCACTGTCCCTGAG	TCGCAACTGGAAGGCATCTTGTAG	PAGE
CCL20	AACAGCACTCCCAAAGAACTGG	GCAGAGGTGGAGTAGCAGCA	PAGE
CCR2	CCAACGAGAGCGGTGAAGAAGTC	CGAGTAGAGCGGAGGCAGGAG	PAGE
CXCL8	ACTTTCAGAGACAGCAGAGCACAC	CACACAGTGAGATGGTTCCTTCCG	PAGE
β-actin	TGGCATCCACGAAACTACCT	CAATGCCAGGGTACATGGTG	PAGE

2.7 Statistical analysis

Mean ± standard deviation (SD) were calculated for all gene expression data. GraphPad PRISM version 8.0 (GraphPad Software, San Diego,CA, USA) was used to perform statistical calculations and prepare graphs. An unpaired, two-tailed Student's t-test was used to compare gene expression between targets and references, with a cutoff for statistical significance of $p < 0.05$.

3. Results

3.1 Identification of candidate DEGs from microarray expression datasets

The four microarray expression datasets of GSE51029, GSE52819, GSE54992 and GSE65517 were obtained from the GEO database. The above data sets were uploaded to GEO2R and standardized (Fig. 1) to screen DEGs between the *M. tuberculosis* infection groups and the healthy control groups, and create a volcano map of the distribution of these DEGs in four datasets. The differential expression of multiple genes in the 2 sets of sample datas in each array is shown in Fig. 2.

In GSE51029 dataset, 5182 DEGs were obtained, including 1024 up-regulated genes and 4156 down-regulated genes. In GSE52819 dataset, 1959 DEGs obtained, including 1044 up-regulated genes and 915 down-regulated genes. In GSE54992 dataset, 7461 DEGs obtained, including 3918 up-regulated genes and 3543 down-regulated genes. In GSE65517 dataset, 2604 DEGs obtained, including 1337 up-regulated genes and 1267 down-regulated genes.

According to the criteria of |log FC|> 1 and p values < 0.05, we used FunRich software to display the intersection of DEGs in the four microarray expression datasets, a total of 46 overlapping genes were found (Fig. 3A), which were regarded as candidate DEGs and used for further analysis. We use the pheatmap software package in R Studio to visualize the heatmap of 46 DEGs (Fig. 3B). It is worth mentioning that by using the corrplot and circlize software packages in R Studio to visualize 46 DEGs, the generated correlation circle map can show the correlation of multiple genes in one picture (Fig. 3C). The outer circle represents genes, and the line between the two genes represents the correlation coefficient. Positive correlation is shown in red, and negative correlation is shown in green. The darker the color, the more significant the correlation.

3.2 Enrichment analysis of GO and KEGG pathways

Enrichment analysis is the core of most existing gene annotation portals(33). In the enrichment analysis process, the input gene list is compared with thousands of gene sets, which are defined by their participation in specific biological processes, protein localization, enzyme functions, pathway members, and other characteristics. We used RSQLite, org.Hs.eg.db, clusterProiler and other software packages in R software to analyze and visualize the GO function and KEGG pathway enrichment of the 46 identified DEGs. The results are shown in Fig. 4.

The top 20 most important items of enrichment analysis are illustrated in the form of bubble diagrams. The size and color of the bubbles indicate the number of DEGs and the importance of enrichment in the enrichment analysis of GO and KEGG pathways, respectively. The first 5 important terms of GO enrichment analysis indicate that in the BP category, DEGs participate in the defense response to the virus, the regulation of the immune effect process, the cellular response of interferon-γ, the regulation of the innate immune response and the chemokine-mediated signal pathway (Fig. 4A). For the MF category, DEGs are related to cytokine receptors, cytokine activity, G protein-coupled receptor binding, receptor ligand activity and signal receptor activator activity (Fig. 4B). For the CC category, DEGs are significantly enriched in membrane rafts, mitochondrial outer membrane, cytoplasmic vesicle cavities, endocytic vesicles and nuclear chromatin (Fig. 4C).

For KEGG pathway enrichment analysis, the first several important KEGG pathways of DEGs include NOD-like receptor signaling pathway, influenza A, chemokine signaling pathway, COVID- 19, toll-like receptor signaling pathway and nuclear factor (NF)-kappa B signaling pathway. In order to show in detail the position and importance of DEGs in the pathway, based on the KEGG enrichment analysis, the pathview software package in R Studio was further used to visualize the 46 DEGs identified. The pathway diagrams are stored as the additional files.

3.3 PPI network construction analysis and hub gene selection

PPI network analysis has been regarded as a useful tool for exploring biological responses in health and disease(34). According to the analysis we conducted using the STRING database and Cytoscape software, the 46 candidate DEGs were placed in a PPI network complex, including 46 nodes and 83 edges. The PPI enrichment *p* value was less than $1.0e^{-16}$, and the confidence score was greater than 0.4. Finally, Cytoscape is used to visualize the PPI network of these DEGs(35)(Fig. 5). After deleting isolated and partially connected nodes, a complex PPI network was successfully constructed (Fig. 5A).

The online software Metascape applies the mature complex recognition algorithm MCODE, which can automatically extract protein complexes embedded in large-scale networks(36). It is worth mentioning that Metascape has obvious advantages in PPI network module analysis and GO enrichment analysis(37), and can directly obtain relevant data. We used Cytoscape to visualize the PPI module analysis results generated by Metascape. The MCODE plugin of Metascape filters the module network in the PPI network again (Fig. 5B-C), and the options are set as default parameters. Host hub genes are defined as genes that play a vital role in diverse biological processes, and usually regulate the activity of other genes(38). We used the Cytohubba plugin, which uses maximal clique centrality (MCC) to study important nodes in the network and select the most relevant genes as the target hub genes(39). The top five hub genes, which had the highest degree of interaction with other genes in the PPI, were identified (Fig. 5D), these included signal transducer and activator of transcription 1-α/β (STAT1), toll-like receptor 7 (TLR7), C-X-C motif chemokine ligand 8 (CXCL8), C-C chemokine receptor type 2 (CCR2), and C-C motif chemokine 20 (CCL20).

3.4 Construction and analysis of the miRNA–hub gene network and TF–hub gene networks

NetworkAnalyst is used to screen targeted miRNA and TF for hub genes(40). For these five identified hub genes, the top three miRNA targeted DEGs are TLR7 regulated by 39 miRNAs, CXCL8 regulated by 36 miRNAs, and STAT1 regulated by 19 miRNAs. The miRNA that controls the largest number of hub genes (three genes) was found to be hsa-mir-335-5p (Fig. 6A), and other important miRNAs are shown in Table 3. In the TF hub gene network analysis, when the selected 5 hub genes were imported into the ENCODE of the TF database, only two of the five genes could be sufficiently analyzed, namely CCL20 regulated by 39 transcription factors and CCR2 regulated by 2 transcription factors (Fig. 6B).

Table 3
Micro RNAs (miRNAs) associated with hub genes identified in this study.

miRNA	Hub genes targeted by miRNA	Gene count	Betweenness Centrality
hsa-mir-335-5p	CXCL8,CCR2,CCL20	3	0.21933622
hsa-mir-93-5p	CXCL8,TLR7	2	0.0620336
hsa-mir-106a-5p	CXCL8,TLR7	2	0.0620336
hsa-mir-203a-3p	CXCL8,STAT1	2	0.05025596
hsa-mir-146a-5p	CXCL8,STAT1	2	0.05025596
hsa-mir-150-5p	TLR7,STAT1	2	0.11976912
hsa-mir-155-5p	CXCL8,STAT1	2	0.05025596
hsa-mir-302c-3p	CXCL8,TLR7	2	0.0620336
hsa-mir-302d-3p	CXCL8,TLR7	2	0.0620336
hsa-mir-520b	CXCL8,TLR7	2	0.0620336

3.5 Verification of hub gene expression levels using RT–PCR

In order to verify the accuracy of the prediction results, this study used RT–PCR to detect the mRNA expression levels of the 5 host hub genes at 24h, 48h, and 72h after *M. tuberculosis* infection of the cells. At least 3 biological replicates were performed for each experiment. The results showed that the expressions of hub genes is basically consistent with the heatmap results drawn by the gene expression matrix (GSE51029_series_matrix) data. Specifically, compared with the PBS groups, gene expression, as indicated by mRNA levels, of STAT1, CCL20, and CXCL8 in the *M. tuberculosis*–infected groups were all upregulated at the three time points (Fig. 7A–C), while TLR7 and CCR2 expression was downregulated at all three time points (Fig. 7D–E).

4. Discussion

TB is an infectious disease that seriously harms human health. It is caused by *M. tuberculosis* that is parasitic in macrophages(41). As the human body's first line of immune defense, macrophages can kill *M. tuberculosis* through phagocytosis, oxidative stress, acidification and antigen presentation(42). Due to the lack of specific auxiliary examination indicators, it is difficult to realize the early diagnosis and effective treatment of TB patients with the traditional TB diagnosis scheme.

The method of bioinformatics is helpful to analyze the expression of key genes to reveal the potential molecular mechanism of the biological behavior of TB, and to provide novel views for elucidating the pathogenesis of TB. Microarray technology allows us to explore the host genetic changes and gene

expression related to TB, and has proven to be a useful method for identifying new biomarkers in other diseases(43).In this study, four GEO microarray data sets (GSE51029, GSE52819, GSE54992 and GSE65517) were integrated to identify DEGs between PBMC of TB patients and healthy persons to offset the false positive rate in the analysis of independent datasets. According to the criteria of |log FC|> 1 and *p* values < 0.05, this study used FunRich software to display the intersection of DEGs in the four microarray expression datas, and found a total of 46 overlapping DEGs that can be considered as candidates. The GO function and KEGG pathway enrichment analysis, PPI network analysis and hub gene selection, the construction of miRNA–hub genes network and target TF–hub genes network, and the verification of the expression of hub genes at the mRNA level were carried out successively.

According to the results of GO enrichment analysis, for the BP category, DEGs participate in the defense response to the virus, the regulation of the immune effect process, the cellular response of IFN–γ, the regulation of the innate immune response and the chemokine-mediated signal pathway. Previous studies have shown that during infection, macrophages encounter *M. tuberculosis* before being stimulated by IFN–γ produced by T-helper 1 (Th1) cells(44). However, IFN–γ stimulation is necessary for the complete activation of antibacterial and antigen presentation functions in macrophages(45). For the MF category, DEGs were found to be associated with cytokine/chemokine receptor binding, cytokine/chemokine activity, G protein coupled receptor binding, receptor ligand activity, and signal receptor activator activity. *M. tuberculosis* is known to induce host proinflammatory mediators that play an important role in disease control(46), including chemokines, which are small molecular weight proteins involved in immune regulation and inflammation(41). For the CC category, DEGs were found to be significantly enriched in membrane rafts, the mitochondrial outer membrane, cytoplasmic vesicle cavities, endocytic vesicles and nuclear chromatin. KEGG enrichment analysis showed that NOD–like and toll-like receptor signaling pathways were associated with *M. tuberculosis* infection. Innate immune cells are known to use various pattern recognition receptors, such as toll–like receptors (TLRs), C-type lectin receptors, and NOD–like receptors to respond to pathogen components when performing a variety of biological functions(47, 48). Previous research using eperimental models of TB have emphasized the importance of TLRs in the prevention of *M. tuberculosis* infection.(49). In addition, antigen recognition of NOD–2 (nucleotide-binding oligomerization domain 2), a member of the NOD–like receptor family, is also crucial in conferring immunity against viruses or bacteria, which may include *M. tuberculosis*(50, 51). This suggests that the coordinated triggering of TLRs and NOD-2 may lead to a stronger and lasting immune response, thereby limiting the growth of *M. tuberculosis*(52), which would be consistent with our results showing enrichment of these pathways during *M. tuberculosis* infection.

By constructing a PPI network and analyzing it with MCODE and Cytohubba in Cytoscape, five hub genes were identified, including STAT1, TLR7, CXCL8, CCR2 and CCL20. Previous studies have reported that in the early stage of TB infection, STAT1 can promote downstream apoptotic factors to activate transcription through phosphorylation(53). At the same time, STAT1 plays an important role in the polarization of macrophages to the M1 type, involved in the immune response to viruses and bacteria, including *M. tuberculosis*. Polarized M1 macrophages have been shown to eliminate *M. tuberculosis* infection more effectively than Polarized M2 macrophages(54). It is also reported that after ssRNA

upregulates TLR7, the number of *M. tuberculosis* in macrophages is significantly reduced, and the macrophage viability is significantly increased, indicating that TLR7 can effectively inhibit the growth of *M. tuberculosis* and increase the viability of macrophages(42). Kane et al. found that fibroblasts have a previously unrecognized role in regulating TB inflammation via a CXCL8-dependent contribution to immune cell recruitment and mycobacterial killing in granulomas(55). In the study of Dunlap et al., the mouse model provided evidence that the CCR2 axis is essential for protective immunity against the emerging *M. tuberculosis* lineage infection(56). Another report showed that CCL20 is overexpressed in monocytes infected by *M. tuberculosis* and inhibits the production of reactive oxygen species (ROS) (57).Therefore, prior research provides biologically plausible mechanisms by which these genes may be involved in immune responses to *M. tuberculosis* infection, supporting our results.

To study the molecular mechanisms of potential hub gene disorders, it is necessary to search for potential miRNAs through bioinformatics methods. The miRNA is an endogenous non–coding RNA molecule with a length of 18–22 nt that targets the 3'UTR region of a gene. It can regulate gene expression at the post-transcriptional level to degrade or inhibit the translation of target genes(58). MiRNAs are known to regulate protein translation inhibition or targeted mRNA cleavage(59). More and more evidences have showed that miRNAs are closely related to the occurrence and development of cancer and other major diseases. In our analysis, the three DEGs most associated with miRNA regulation were CXCL8, TLR7, and STAT1. At the same time, we observed 10 miRNAs, and found that their targeting involves at least 2 hub genes. Among these, hsa-mir-335-5p was found to be associated with the greatest number of genes; however, there are relatively few previous studies on this miRNA. One study found that the gain or loss of hsa-miR-335-3p function can lead to changes in the expression of GATA4 and NKX2–5 markers during the cardiac differentiation of human embryonic stem cells(60). In addition, hsa-miR-335-3p has been identified as an upstream regulator of two modules related to the recurrence of osteosarcoma patients(61). The results of bioinformatics analysis found that hsa-mir-335-5p has high potential value used as a new biomarker. Our study also established a TF–hub genes regulatory network to further explore the molecular regulatory mechanisms underlying TB(62). TFs are the primary regulators of gene expression and are associated with pathogenesis in TB. In our study, we also found several TFs that interact closely with hub genes, including max-like protein X (MLX), transcription factor DP 1 (TFDP1), retinoid X receptor alpha (RXRA), zinc finger protein 197 (ZNF197), glucocorticoid modulatory element binding protein 2 (GMEB2), and tripartite motif containing 22 (TRIM22). The complex interactions between TFs and hub genes have made a huge contribution to the development of the disease. Our analysis, which identifies miRNAs and TFs associated with newly identified hub genes, provides potential candidates for the development of therapeutic targets and exploration of the biological mechanism of TB in future research.

5. Conclusions

In summary, this study identified several candidate DEGs, including five key hub genes, associated with *M. tuberculosis* infection, some of which may be potentially useful biomarkers for TB. These results may provide a basis for screening candidate therapeutic agents and biomarkers for diagnosis of TB and could

inform future research on molecular mechanisms involved in this disease. However, further *in vitro* and *in vivo* studies are needed to confirm the predicted functions of these DEGs in TB-related physiological and pathological processes. Experimental models in studies exploring the molecular mechanisms of TB could then potentially be constructed based on these genes and may enable improvements in early detection, prognosis, disease risk estimation, and targeted therapies in TB.

Declarations

Ethics approval and consent to participants

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the [GEO] repository, [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi TO DATASETS ALONG WITH THE GSE51029、GSE52819、GSE54992、GSE65517]

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the Natural Foundation of Yunnan Province [No. 2019FE001 (-002) and 2017FE467 (-001)] and the National Natural Science Foundation of China (No. 32060180, 81860644, 81560596, and 31560051). The funding institutions had no involvement in the design of the study or review of the manuscript.

Authors' contributions

FKB, AHL, and PY conceived and designed the experiments. PY, YD, XX, YZ, JK, and YH developed the methodology. PY, YD, JJC, YXF, MXL, and WJC performed all experiments. PY, SYW, BXL, LSL, TGC, LBL, FW, and GZZ analyzed and discussed the data. PY and YD wrote the manuscript. FKB, AHL, and PY edited and revised the manuscript. All authors read and approved the manuscript.

Acknowledgments

Not applicable.

References

1. Organization WH. Global tuberculosis report 2020. 2020.
2. Fan S, Zhou G, Shang P, Meng L. Clinical Study of 660 Cases of Pulmonary Tuberculosis. Harbin Medical Journal. 2014;34(1):1-11.
3. Siddiqi K, Lambert M-L, Walley J. Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: the current evidence. The Lancet Infectious Diseases. 2003;3(5):288-96.
4. Won E-J, Choi J-H, Cho Y-N, Jin H-M, Kee HJ, Park Y-W, et al. Biomarkers for discrimination between latent tuberculosis infection and active tuberculosis disease. Journal of Infection. 2017;74(3):281-93.
5. Sia IG, Wieland ML, editors. Current concepts in the management of tuberculosis. Mayo Clinic Proceedings; 2011.
6. Kumar M, Sahu SK, Kumar R, Subuddhi A, Maji RK, Jana K, et al. MicroRNA let-7 modulates the immune response to Mycobacterium tuberculosis infection via control of A20, an inhibitor of the NF-κB pathway. Cell host & microbe. 2015;17(3):345-56.
7. Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing. 2017;50:124-34.
8. Ramaswamyreddy SH, Smitha T. Microarray-based gene expression profiling for early detection of oral squamous cell carcinoma. Journal of oral and maxillofacial pathology. 2018;22(3):293.
9. Yang X, Zhu S, Li L, Zhang L, Xian S, Wang Y, et al. Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. Onco Targets Ther. 2018;11:1457-74.
10. Xie L, Chao X, Teng T, Li Q, Xie J. Identification of Potential Biomarkers and Related Transcription Factors in Peripheral Blood of Tuberculosis Patients. Int J Environ Res Public Health. 2020;17(19):6993.
11. Qin XB, Zhang WJ, Zou L, Huang PJ, Sun BJ. Identification potential biomarkers in pulmonary tuberculosis and latent infection based on bioinformatics analysis. BMC Infect Dis. 2016;16(1):500.
12. Dumas J, Gargano M, Dancik GM. An online tool for biomarker analysis in Gene Expression Omnibus (GEO) datasets. Bioinform; 2016. p. 5292.
13. Xu Z, Zhou Y, Cao Y, Dinh TLA, Wan J, Zhao M. Identification of candidate biomarkers and analysis of prognostic values in ovarian cancer by integrated bioinformatics analysis. Medical Oncology. 2016;33(11):130.
14. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications. 2019;10(1):1-10.
15. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457-62.
16. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006;34(Database issue):D322-6.

17. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.
18. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41(Database issue):D808-15.
19. Wang H, Zhu H, Zhu W, Xu Y, Wang N, Han B, et al. Bioinformatic Analysis Identifies Potential Key Genes in the Pathogenesis of Turner Syndrome. Front Endocrinol (Lausanne). 2020;11:104.
20. Pizzuti C, Rombo SE. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics. 2014;30(10):1343-52.
21. Bandettini WP, Kellman P, Mancini C, Booker OJ, Vasu S, Leung SW, et al. MultiContrast Delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study. Journal of Cardiovascular Magnetic Resonance. 2012;14(1):1-10.
22. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol. 2014;8 Suppl 4(Suppl 4):S11.
23. Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 2019;47(W1):W234-w41.
24. Soifer HS, Rossi JJ, Sætrom P. MicroRNAs in disease and potential therapeutic applications. Molecular therapy. 2007;15(12):2070-9.
25. Baldwin AS. Series introduction: the transcription factor NF-κB and human disease. The Journal of clinical investigation. 2001;107(1):3-6.
26. Xia J, Gill EE, Hancock RE. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nature protocols. 2015;10(6):823-44.
27. Yang D, He Y, Wu B, Deng Y, Wang N, Li M, et al. Integrated bioinformatics analysis for the screening of hub genes and therapeutic drugs in ovarian cancer. Journal of ovarian research. 2020;13(1):10.
28. Yang W, Zhao X, Han Y, Duan L, Lu X, Wang X, et al. Identification of hub genes and therapeutic drugs in esophageal squamous cell carcinoma based on integrated bioinformatics strategy. Cancer cell international. 2019;19(1):1-15.
29. Zhang YW, Lin Y, Yu HY, Tian RN, Li F. Characteristic genes in THP-1 derived macrophages infected with Mycobacterium tuberculosis H37Rv strain identified by integrating bioinformatics methods. International journal of molecular medicine. 2019;44(4):1243-54.
30. Feng Z, Bai X, Wang T, Garcia C, Bai A, Li L, et al. Differential responses by human macrophages to infection with Mycobacterium tuberculosis and non-tuberculous mycobacteria. Frontiers in microbiology. 2020;11:116.
31. Ding Z, Sun L, Bi Y, Zhang Y, Yue P, Xu X, et al. Integrative Transcriptome and Proteome Analyses Provide New Insights Into the Interaction Between Live Borrelia burgdorferi and Frontal Cortex

Explants of the Rhesus Brain. Journal of Neuropathology & Experimental Neurology. 2020;79(5):518-29.

32. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. methods. 2001;25(4):402-8.
33. Li H, Long J, Xie F, Kang K, Shi Y, Xu W, et al. Transcriptomic analysis and identification of prognostic biomarkers in cholangiocarcinoma. Oncology reports. 2019;42(5):1833-42.
34. Vella D, Marini S, Vitali F, Di Silvestre D, Mauri G, Bellazzi R. MTGO: PPI network analysis via topological and functional module identification. Scientific reports. 2018;8(1):1-13.
35. Feng H, Gu Z-Y, Li Q, Liu Q-H, Yang X-Y, Zhang J-J. Identification of significant genes with poor prognosis in ovarian cancer via bioinformatical analysis. Journal of ovarian research. 2019;12(1):1-9.
36. Liang J, Wu M, Bai C, Ma C, Fang P, Hou W, et al. Network Pharmacology Approach to Explore the Potential Mechanisms of Jieduan-Niwan Formula Treating Acute-on-Chronic Liver Failure. Evidence-Based Complementary and Alternative Medicine. 2020;2020.
37. Li W, Wang S, Qiu C, Liu Z, Zhou Q, Kong D, et al. Comprehensive bioinformatics analysis of acquired progesterone resistance in endometrial cancer cell line. Journal of translational medicine. 2019;17(1):1-17.
38. Zhang YM, Meng LB, Yu SJ, Ma DX. Identification of potential crucial genes in monocytes for atherosclerosis using bioinformatics analysis. J Int Med Res. 2020;48(4):300060520909277.
39. Guo C, Li Z. Bioinformatics Analysis of Key Genes and Pathways Associated with Thrombosis in Essential Thrombocythemia. Medical science monitor: international medical journal of experimental and clinical research. 2019;25:9262.
40. Zhou R, Liu D, Zhu J, Zhang T. Common gene signatures and key pathways in hypopharyngeal and esophageal squamous cell carcinoma: Evidence from bioinformatic analysis. Medicine. 2020;99(42).
41. Lyon SM, Rossman MD. Pulmonary tuberculosis. Tuberculosis and Nontuberculous Mycobacterial Infections. 2017:283-98.
42. Bao M, Yi Z, Fu Y. Activation of TLR7 inhibition of Mycobacterium tuberculosis survival by autophagy in RAW 264.7 macrophages. Journal of cellular biochemistry. 2017;118(12):4222-9.
43. Li L, Lei Q, Zhang S, Kong L, Qin B. Screening and identification of key biomarkers in hepatocellular carcinoma: evidence from bioinformatic analysis. Oncology reports. 2017;38(5):2607-18.
44. Brzezinska M, Szulc I, Brzostek A, Klink M, Kielbik M, Sulowska Z, et al. The role of 3-ketosteroid 1 (2)-dehydrogenase in the pathogenicity of Mycobacterium tuberculosis. BMC microbiology. 2013;13(1):1-12.
45. Raja A. Immunology of tuberculosis. Indian Journal of Medical Research. 2004;120(4):213-32.
46. Ansari AW, Kamarulzaman A, Schmidt RE. Multifaceted impact of host C–C chemokine CCL2 in the immuno-pathogenesis of HIV-1/M. tuberculosis co-infection. Frontiers in immunology. 2013;4:312.

47. Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. Cell. 2006;124(4):783-801.
48. Akira S, Takeda K, Kaisho T. Toll-like receptors: critical proteins linking innate and acquired immunity. Nature immunology. 2001;2(8):675-80.
49. Fremond CM, Yeremeev V, Nicolle DM, Jacobs M, Quesniaux VF, Ryffel B. Fatal Mycobacterium tuberculosis infection despite adaptive immune response in the absence of MyD88. The Journal of clinical investigation. 2004;114(12):1790-9.
50. Pandey AK, Yang Y, Jiang Z, Fortune SM, Coulombe F, Behr MA, et al. NOD2, RIP2 and IRF5 play a critical role in the type I interferon response to Mycobacterium tuberculosis. PLoS Pathog. 2009;5(7):e1000500.
51. Lupfer C, Thomas PG, Kanneganti T-D. Nucleotide oligomerization and binding domain 2-dependent dendritic cell activation is necessary for innate immunity and optimal CD8+ T cell responses to influenza A virus infection. Journal of virology. 2014;88(16):8946-55.
52. Khan N, Pahari S, Vidyarthi A, Aqdas M, Agrewala JN. NOD-2 and TLR-4 signaling reinforces the efficacy of dendritic cells and reduces the dose of TB drugs against Mycobacterium tuberculosis. Journal of innate immunity. 2016;8(3):228-42.
53. Yao K, Chen Q, Wu Y, Liu F, Chen X, Zhang Y. Unphosphorylated STAT1 represses apoptosis in macrophages during Mycobacterium tuberculosis infection. Journal of cell science. 2017;130(10):1740-51.
54. Lim Y-J, Yi M-H, Choi J-A, Lee J, Han J-Y, Jo S-H, et al. Roles of endoplasmic reticulum stress-mediated apoptosis in M1-polarized macrophages during mycobacterial infections. Scientific reports. 2016;6(1):1-11.
55. O'Kane CM, Boyle JJ, Horncastle DE, Elkington PT, Friedland JS. Monocyte-dependent fibroblast CXCL8 secretion occurs in tuberculosis and limits survival of mycobacteria within macrophages. The Journal of Immunology. 2007;178(6):3767-76.
56. Dunlap MD, Howard N, Das S, Scott N, Ahmed M, Prince O, et al. A novel role for C–C motif chemokine receptor 2 during infection with hypervirulent Mycobacterium tuberculosis. Mucosal immunology. 2018;11(6):1727-42.
57. Rivero-Lezcano OM, González-Cortés C, Reyes-Ruvalcaba D, Diez-Tascón C. CCL20 is overexpressed in Mycobacterium tuberculosis-infected monocytes and inhibits the production of reactive oxygen species (ROS). Clinical & Experimental Immunology. 2010;162(2):289-97.
58. Sun K-T, Chen MY, Tu M-G, Wang I-K, Chang S-S, Li C-Y. MicroRNA-20a regulates autophagy related protein-ATG16L1 in hypoxia-induced osteoclast differentiation. Bone. 2015;73:145-53.
59. Bartel DP. MicroRNAs: target recognition and regulatory functions. cell. 2009;136(2):215-33.
60. Kay M, Soltani BM, Aghdaei FH, Ansari H, Baharvand H. Hsa-miR-335 regulates cardiac mesoderm and progenitor cell differentiation. Stem cell research & therapy. 2019;10(1):1-13.
61. Chen Y, Chen Q, Zou J, Zhang Y, Bi Z. Construction and analysis of a ceRNA-ceRNA network reveals two potential prognostic modules regulated by hsa-miR-335-5p in osteosarcoma. International

journal of molecular medicine. 2018;42(3):1237-46.

62. Li T, Gao X, Han L, Yu J, Li H. Identification of hub genes with prognostic values in gastric cancer by bioinformatics analysis. World journal of surgical oncology. 2018;16(1):1-12.

Figures

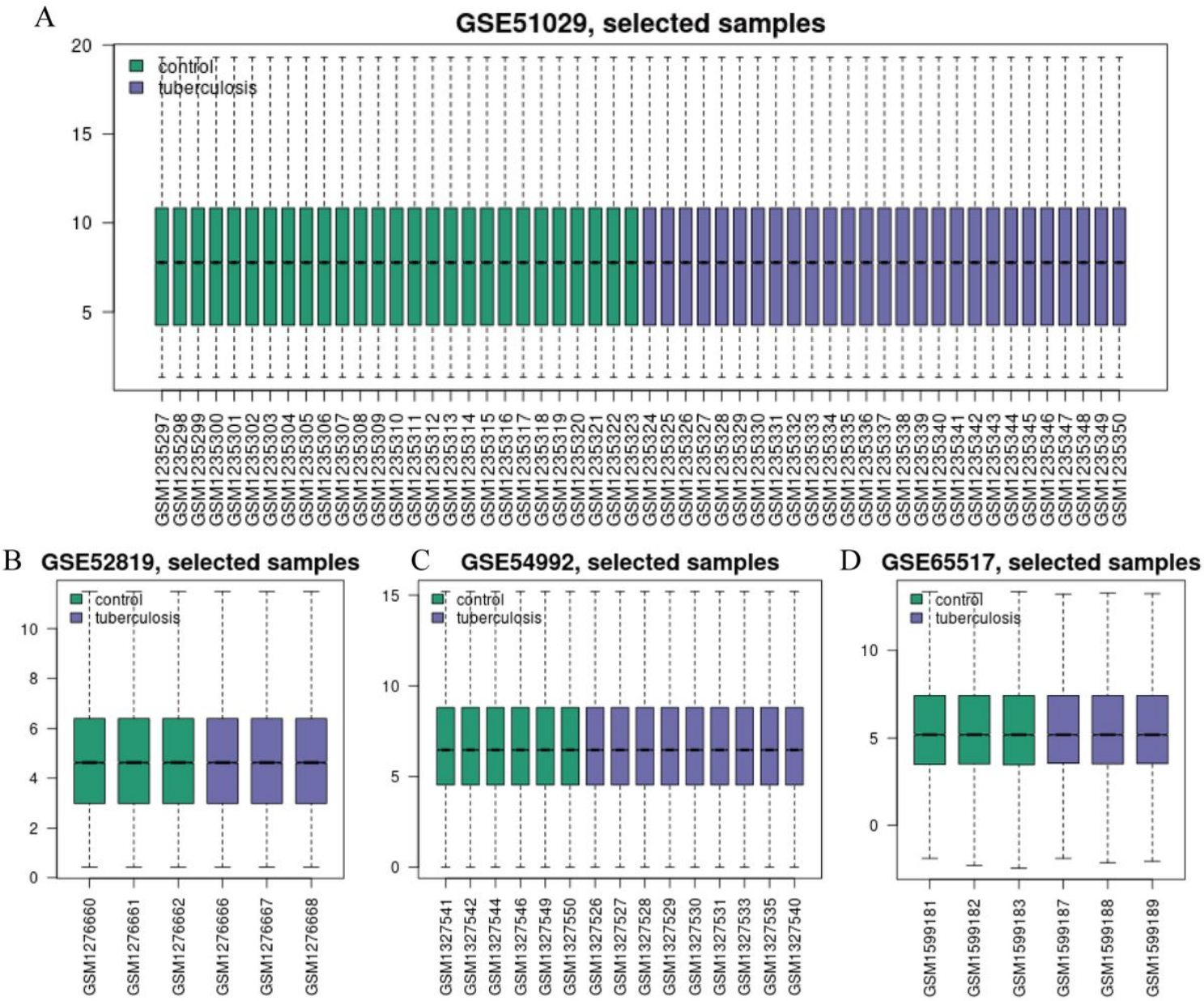

Figure 1

Standardization of gene expression. Notes: (A) The standardization of GSE51029 data, (B) the standardization of GSE52819 data, (C) the standardization of GSE54992 data,and (D) the standardization of GSE65517 data.The green bars represent the normalized data for the healthy control groups, and the blue bars represent the normalized data for the M. tuberculosis-infected groups .

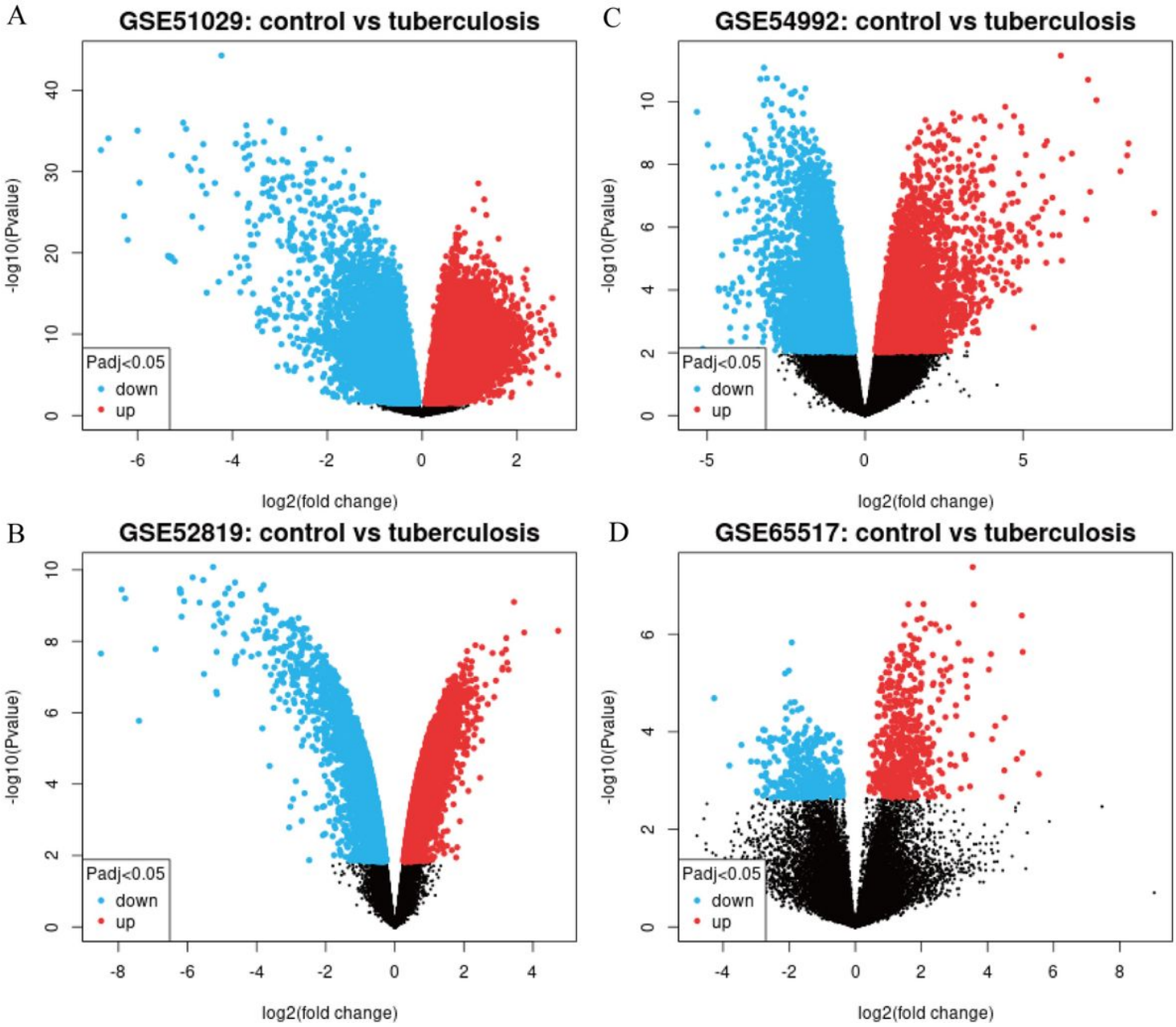

Figure 2

Differential expression trends of data between M. tuberculosis and control samples. Notes: (A) GSE51029 data, (B) GSE52819 data, (C) GSE54992 data, and (D) GSE65517 data. Red and blue points represent significantly up- and downregulated genes (log fold change (LogFC) > 1.0 and corrected p-value < 0.05); black points represent genes with no significant difference in expression between the two samples in each microarray.

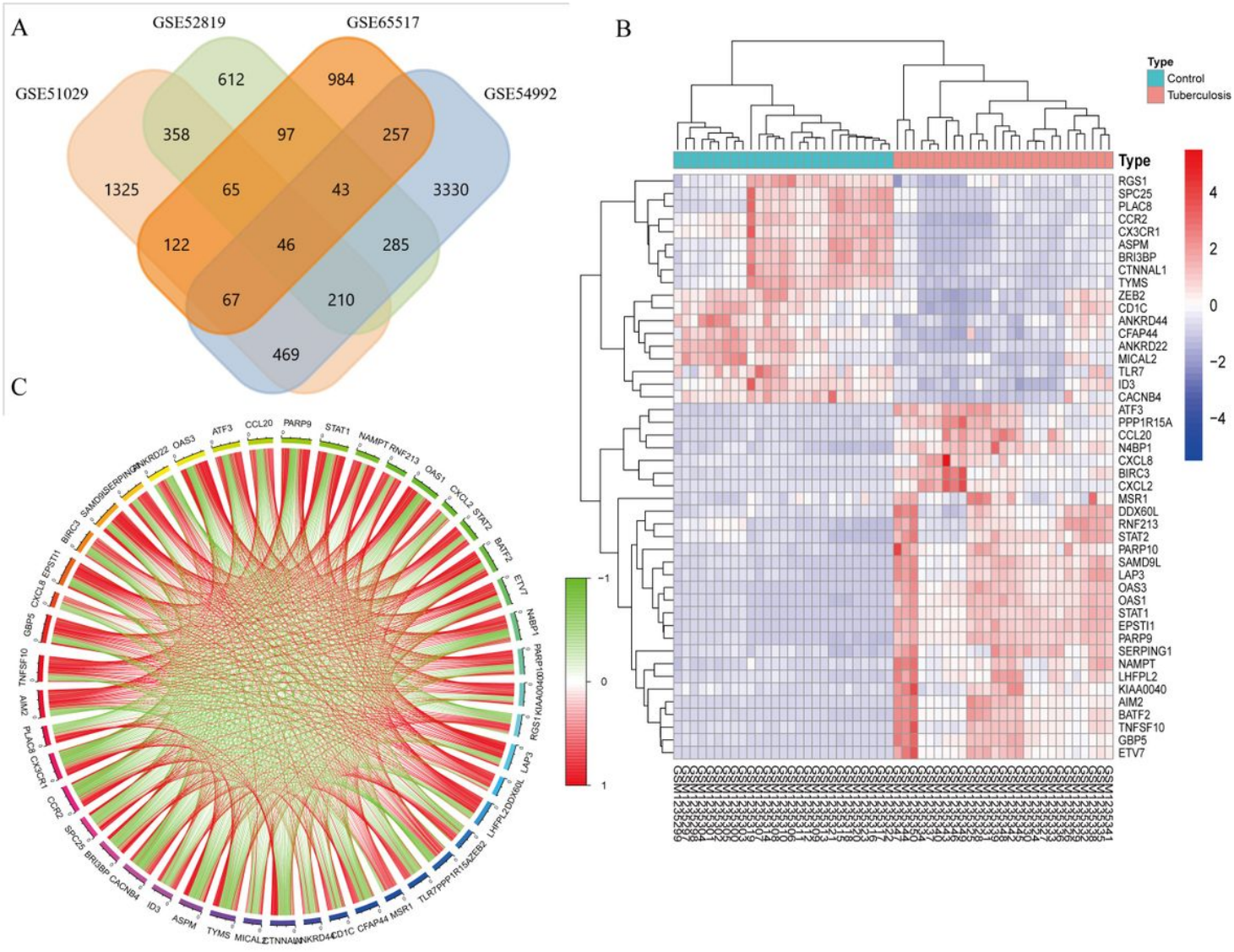

Figure 3

Selection of candidate DEGs, hierarchical clustering heatmap and corCircos. Notes: (A) Venn diagram for overlapping DEGs based on the four datasets, namely, GSE51029,GSE52819, GSE54992, and GSE65517. (B) Hierarchical clustering heatmap using data derived from the dataset GSE51029; red indicates that the expression of genes is relatively up regulated, blue indicates that the expression of genes is relatively down regulated. (C) corCircos shows the relationship of multiple genes in a single picture. The positive correlation is shown in red and the negative correlation is shown in green.

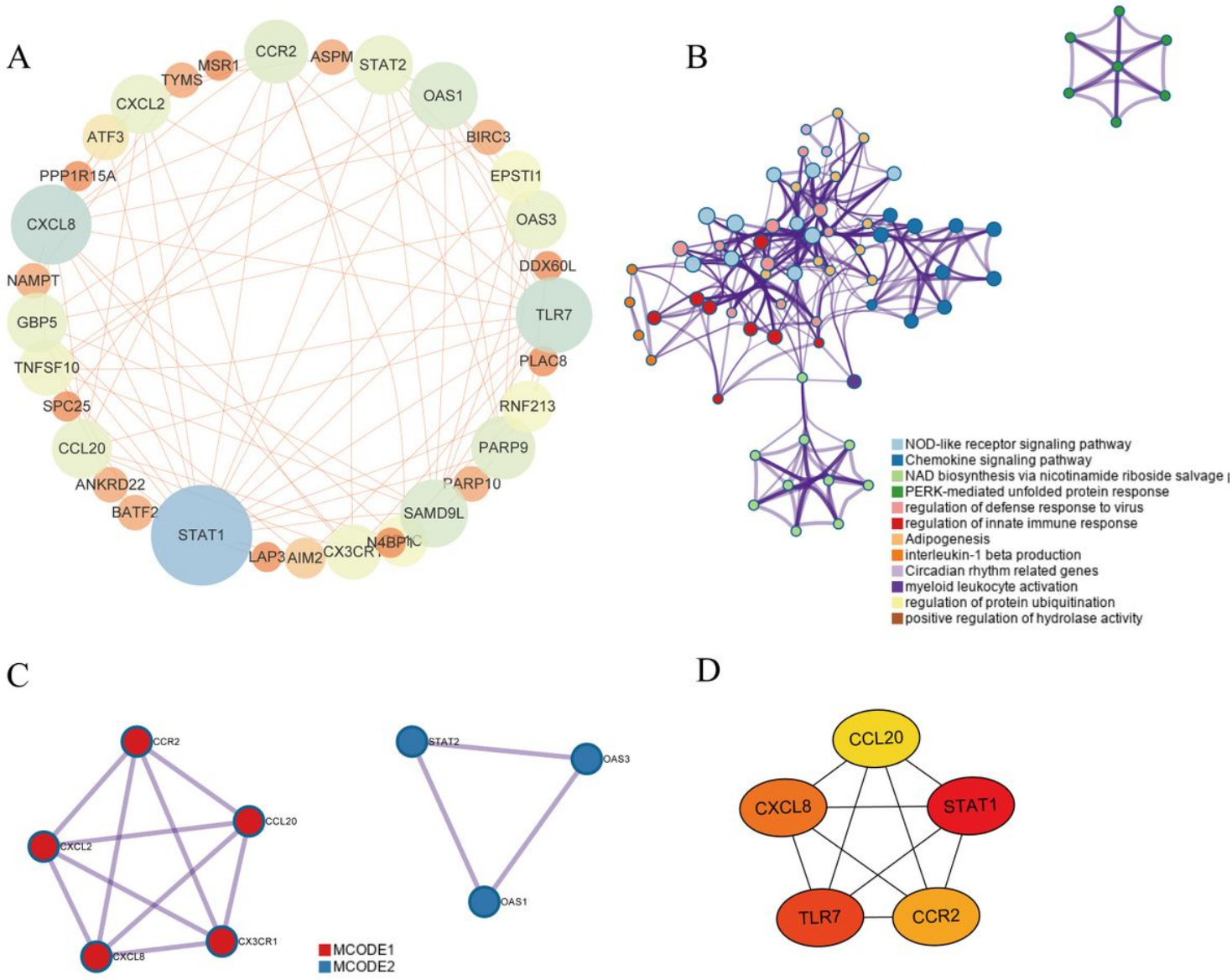

Figure 5

Protein-protein interaction (PPI) network of DEGs and identification of hub genes. Notes: (A) The PPI network of DEGs was visualized by Cytoscape in our analysis; (B) network of enriched terms for differentially expressed proteins, with each node representing an enriched term, colored according to its cluster ID; (C) MCODE components identified in the PPI interaction network, with each node representing an enriched gene, with and different colors representing different MCODE components; (D) subnetwork of top five hub genes from the PPI network, with red, orange, and yellow colours representing higher, medium, and lower degrees of connectivity, respectively.

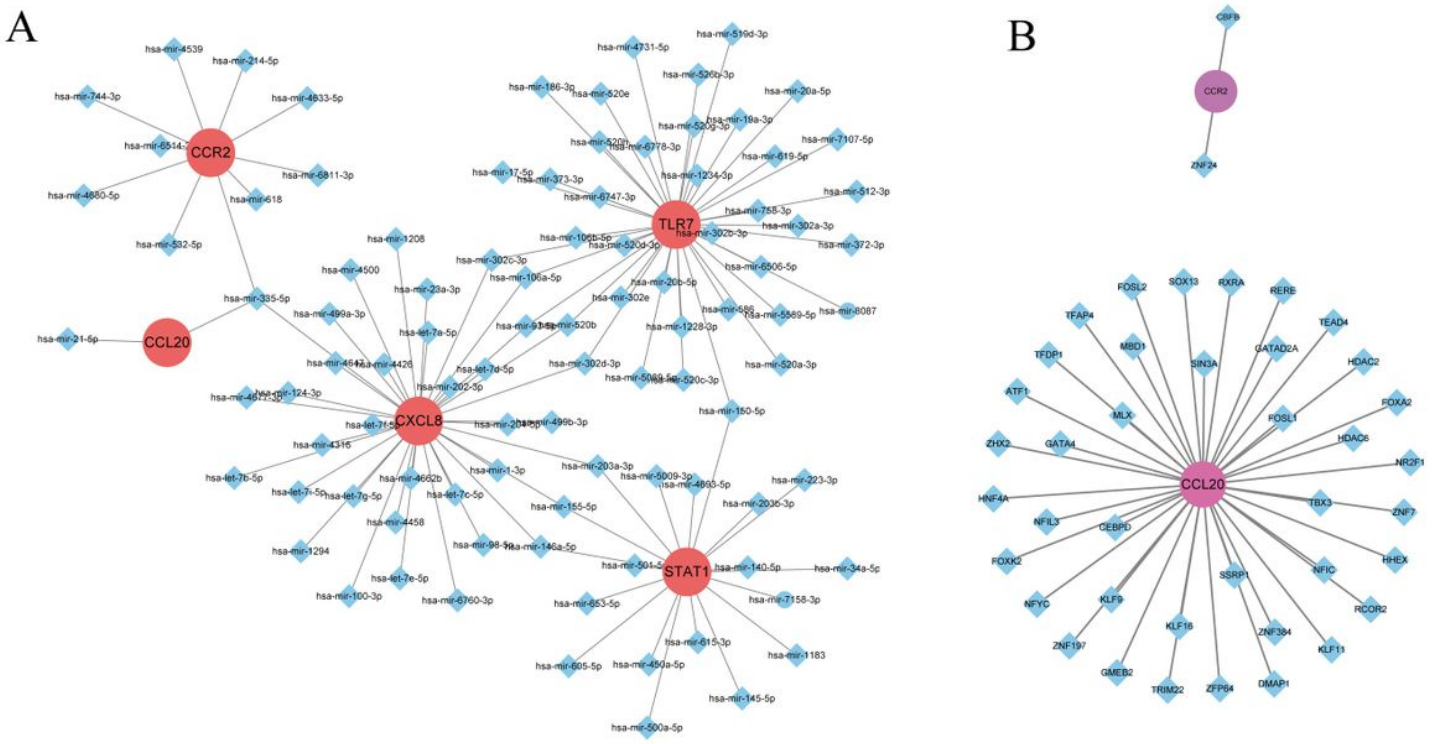

Figure 6

Regulatory networks of miRNA-hub genes and TF-hub genes. Notes: (A) miRNA–hub genes (B) transcription factor (TF)–hub genes regulatory networks. Red or purple circular nodes represent individual hub genes, and blue diamond nodes represent the miRNAs and TFs in A and B, respectively.

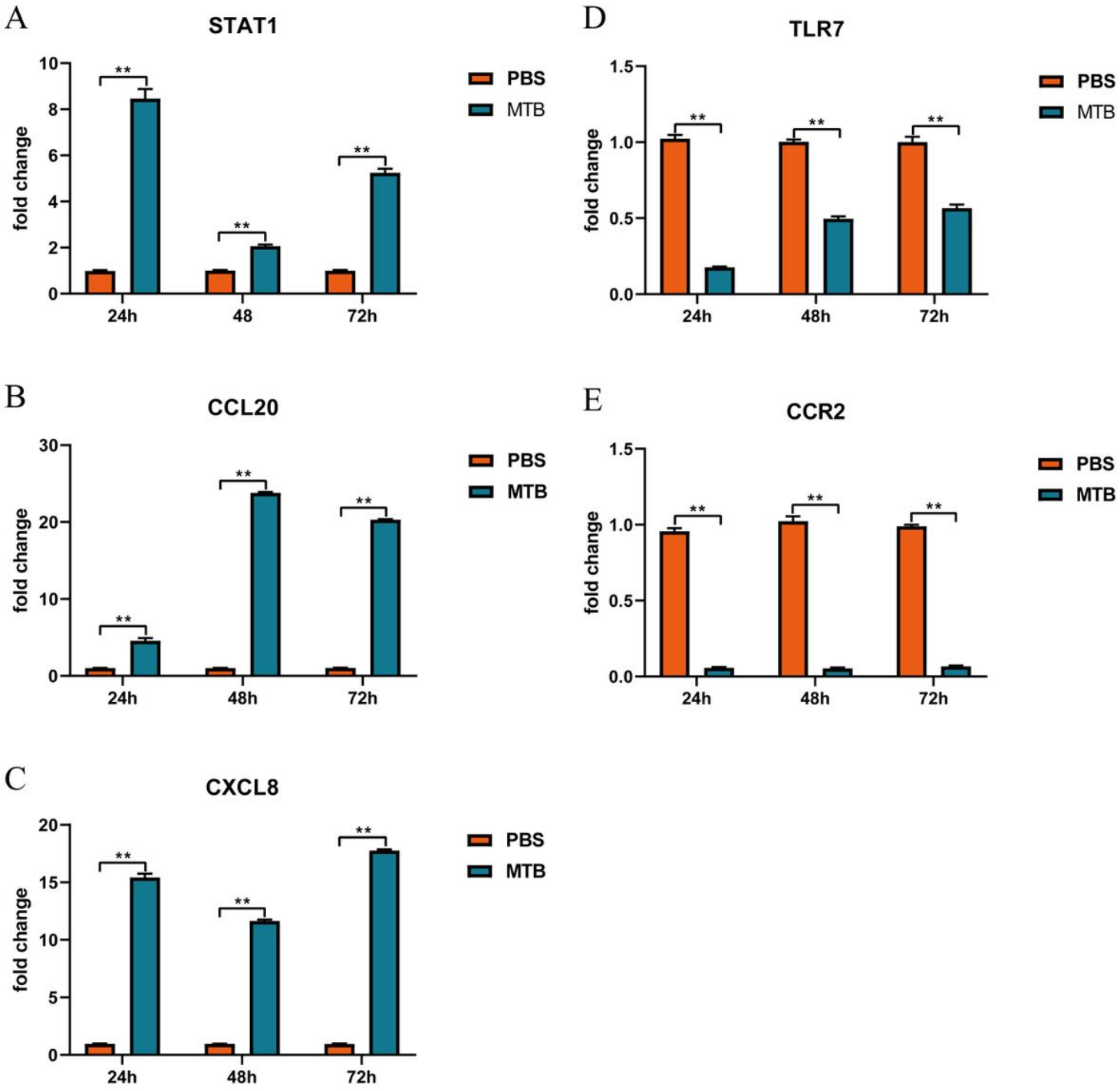

Figure 7

The expression levels of relevant gene mRNA analyzed using RT–PCR comparing M. tuberculosis-infected groups with the control groups. Notes: (A-C) STAT1,CCL20 and CXCL8 was demonstrated upregulated in M. tuberculosis-infected groups compared to the control groups at 24、48、72 hours. (D-E)TLR7 and CCR2 was demonstrated downregulated in M. tuberculosis-infected groups compared to the control groups at 24, 48, 72 hours. Data represent mean 6±SD(n=3). Asterisks denote: **p<0.01indicates significance. Yellow color represents M. tuberculosis-infected groups, and blue color represents the control groups.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFileKEGGpathwaydiagrams.pdf