Research Paper

Integrative genomics analysis identifies promising SNPs and genes implicated in tuberculosis risk based on multiple omics datasets

Mengqiu Xu[1], Jingjing Li[2], Zhaoying Xiao[1], Jiongpo Lou[1], Xinrong Pan[1], Yunlong Ma[3,4]

[1]Department of Infectious Diseases, Shengzhou People's Hospital, The First Affiliated Hospital of Zhejiang University Shengzhou Branch, Shengshou 312400, Zhejiang, China
[2]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Zhejiang University School of Medicine, Hangzhou 310003, Zhejiang, China
[3]Institute of Biomedical Big Data, Wenzhou Medical University, Wenzhou 325027, Zhejiang, China
[4]School of Biomedical Engineering, School of Ophthalmology and Optometry and Eye Hospital, Wenzhou Medical University, Wenzhou 325027, Zhejiang, China

Correspondence to: Yunlong Ma; **email:** glb-biotech@zju.edu.cn

Keywords: genetic variants, GWAS, risk genes, tuberculosis, gene expression
Received: April 14, 2020 **Accepted:** July 7, 2020 **Published:** October 13, 2020

ABSTRACT

More than 10 GWASs have reported numerous genetic loci associated with tuberculosis (TB). However, the functional effects of genetic variants on TB remains largely unknown. In the present study, by combining a reported GWAS summary dataset (N = 452,264) with 3 independent eQTL datasets (N = 2,242) and other omics datasets downloaded from public databases, we conducted an integrative genomics analysis to highlight SNPs and genes implicated in TB risk. Based on independent biological and technical validations, we prioritized 26 candidate genes with eSNPs significantly associated with gene expression and TB susceptibility simultaneously; such as, *CDC16* (rs7987202, rs9590408, and rs948182) and *RCN3* (rs2946863, rs2878342, and rs3810194). Based on the network-based enrichment analysis, we found these 26 highlighted genes were jointly connected to exert effects on TB susceptibility. The co-expression patterns among these 26 genes were remarkably changed according to *Mycobacterium tuberculosis* (MTB) infection status. Based on 4 independent gene expression datasets, 21 of 26 genes (80.77%) showed significantly differential expressions between TB group and control group in mesenchymal stem cells, mice blood and lung tissues, as well as human alveolar macrophages. Together, we provide robust evidence to support 26 highlighted genes as important candidates for TB.

INTRODUCTION

Tuberculosis (TB), a communicable respiratory disease, is major threat to human health in the world, especially in low and middle income countries in Asia [1–3]. There are approximately 10.4 million new cases and 1.7 million deaths worldwide in 2016 [4]. Although the advanced developments in diagnosis and treatment, accurate diagnosis of TB is still difficult and the healthcare and economic burdens of TB remain high. Complicated interactions among host, pathogen, and environmental factors contributed to the development of TB, of which the symptoms contain severe persistent coughing, fever, hemoptysis, chest pain and weight loss [5]. Family and twin studies [6–8] have reported that host genetic components play important roles in contributing risk to TB. Thereby, substantial interests in identifying the genetic components implicated in the aetiology of TB are growing.

In previous decades, TB has been a focus of many candidate gene-based and genome-wide association studies (GWAS). For candidate gene-based association studies on TB, which are dependent on a prior hypothesis that we know the knowledge of the functions of candidate genes, numerous genes with pressing single nucleotide polymorphisms (SNPs) have been identified to be associated with TB [9–15]. For example, genetic variations in *TLR* genes have reported to show associations with TB and clinical outcomes in previous studies [9–11]. With the advances of next-generation sequencing or microarray technology, the approach of GWAS based on powerful hypothesis-free methodology has been extensively applied to investigate the genetic architectures of complex diseases including TB and identify thousands of common risk SNPs. Since the first GWAS on TB was reported in the year of 2010 [16], subsequently many GWASs [17–25] have demonstrated associations between numerous common SNPs and TB among European and other ancestry populations. For example, there were 4 common SNPs identified to be significantly associated with TB via GWASs in Russian or African populations [16–18]. Nevertheless, despite intensifying GWAS studies have been conducted, much of the heritability of TB remains missing.

The vast majority of GWAS-identified significant or suggestive SNPs associated with complex diseases were located in non-coding genomic regions [8, 26]. Consistently, most of previously identified susceptibility variants associated with TB were mapped into non-coding regions [27]. Thus, it is plausible to infer that these GWAS-identified variants may have regulatory effects on influencing the expression level of specific gene instead of altering the function of its protein. A recent multi-cohort study [28] demonstrated that aberrant expression signature of a three-gene set (*GBP5*, *DUSP3*, and *KLF2*) is highly diagnostic for active TB. Furthermore, an accruing number of studies have concentrated on exploration of susceptibility genes whose aberrant expression are associated with diseases and traits of medical importance in humans due to pleiotropy [28–32]. For example, by using an integrative analysis of GWAS summary-level, mQTL and eQTL data, our team [33] previously found 34 important genes including *PRKCZ*, *ARHGEF3*, and *CDKN1A* with various critical SNPs contribute risk to the comorbidity of schizophrenia and smoking behaviors. Many novel risk genes identified by numerous integrative genomics studies were hard to be detected by a GWAS alone.

To the best of our knowledge, there was no systematical integrative genomics analysis on TB conducted to reveal the genome-wide regulatory effects of SNPs on gene expression. In the present study, we applied a two-stage designed analysis to identified risk SNPs, genes and pathways for TB. We first used the Sherlock integrative analysis to identify cis- and trans-regulatory effects of SNPs on expression abundance of interested genes via incorporating a large-scale GWAS summary dataset (N = 452,264) with a blood-based eQTL dataset (N = 1,490). Then, using the Sherlock analysis with same parameters, we adopted two independent eQTL datasets based on blood (N = 369) and lung tissue (N = 383) to replicate the results in the discovery stage. Furthermore, we employed a series of bioinformatics analyses including MAGMA analysis, *in silico* permutation analysis, pathways/diseases-based enrichment analysis, network-based enrichment analysis, DGIdb enrichment analysis, and co-expression analysis based on multi-omics data to highlight TB-associated risk genes with strong evidence.

RESULTS

Identification of TB-associated genes in the discovery stage

In the discovery stage, we conducted a Sherlock Bayesian integrative analysis by incorporating GWAS summary statistics (Dataset #1; N = 452,264) with eQTL data (Dataset #3; N = 1,490) to identify aberrant expressed genes with eSNPs implicated in TB risk (Figure 1). There were a number of 694 genes identified to be significantly associated with TB risk (Gene set #1, Simulated P ≤ 0.05; Figure 2A and Supplementary Table 1). For example, the top-ranked significant genes were *SIPA1L1* (Simulated P = 1.26×10^{-5}), *GSTA2* (Simulated P = 1.61×10^{-4}), *TIGD6* (Simulated P = 3.02×10^{-4}), *TSPYL4* (Simulated P = 4.22×10^{-4}), and *POLG2* (Simulated P = 4.68×10^{-4}). Interestingly, among these identified significant genes, 4 genes of *C2CD2*, *HLA-DRB6*, *HLA-DQB1*, and *LPCAT2* have been reported to be associated with TB in earlier studies (Supplementary Figure 1 and Supplementary Table 1). In addition, there existed 7 genes documented to be associated with respiratory relevant diseases, such as asthma and chronic obstructive pulmonary disease (Supplementary Figure 1 and Supplementary Table 1); and 38 genes identified to be associated with lung function and related diseases, such as lung cancer and adenocarcinoma (Supplementary Figure 1 and Supplementary Table 1).

To annotate the molecular functions and biological pathways of these 694 identified genes, we performed a functional enrichment analysis by using the KOBAS tool. As for pathway enrichment analysis, 305 pathways were significantly enriched by these TB-associated genes (FDR ≤ 0.05; Figure 2B and Supplementary Table 2). For example, the pathways of metabolism (FDR = 1.78×10^{-28}), immune system (FDR = 2.12×10^{-21}), metabolic pathways (FDR = 7.75×10^{-19}), and tuberculosis (FDR = 4.44×10^{-5}). Furthermore, 231

GO-terms (FDR ≤ 0.05; Figure 2C and Supplementary Table 3) and 50 diseases-terms (FDR ≤ 0.05; Figure 2D and Supplementary Table 4) were significantly overrepresented by these TB-relevant genes.

Validation of TB-associated genes in the replication stage

Furthermore, we utilized two independent eQTL datasets (Datasets #4 and #5) to carry out the Sherlock Bayesian analysis with same parameters for validation. Based on these two independent datasets, we identified 311 significant genes for Dataset #4 based on whole blood samples (Gene set #2, Simulated P ≤ 0.05; Supplementary Table 5) and 405 significant genes for Dataset #5 based on lung tissues (Gene set #3, Simulated P ≤ 0.05; Supplementary Table 6). Among these genes, 3 genes of *ESPPRB*, *GLRX5*, and *LRPAP1* have been reported to be linked with TB in earlier studies (Supplementary Figure 2). 30 and 18 genes have been documented to be associated with lung-related diseases (Supplementary Table 5) and respiratory-related diseases (Supplementary Table 6), separately. Interestingly, there existed 7 genes showing associations with both lung-related diseases and respiratory disease (Supplementary Figure 2). Compared with genes identified in the discovery stage (Gene set #1), we found 26 genes were significantly replicated by the Sherlock analysis of both datasets in the replication stage (Gene sets #2 and #3) (Figure 2A and Table 1). Most of these 26 highlighted genes were highly expressed in human lung tissue (Supplementary Figures 3–15).

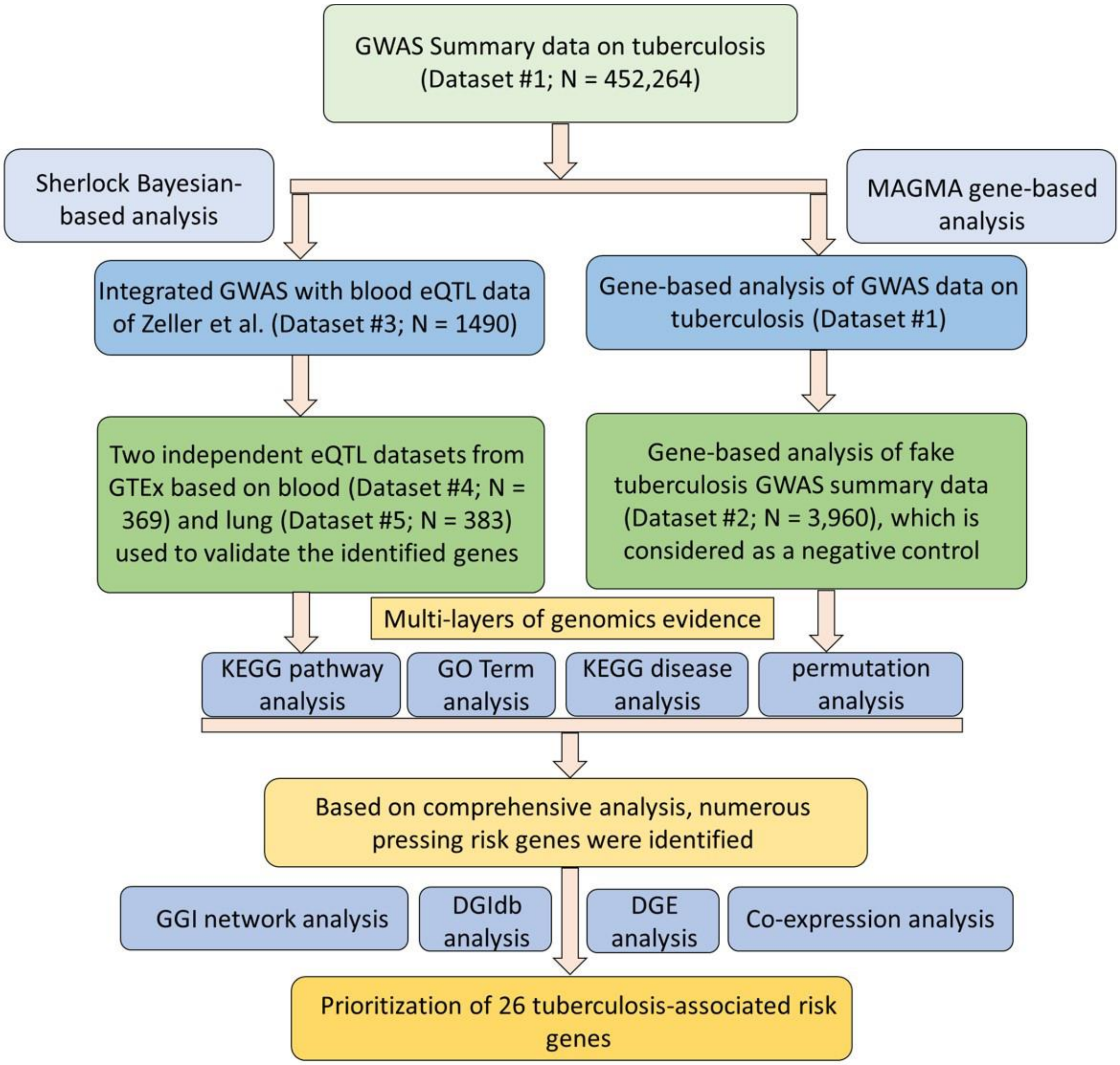

GWAS Summary data on tuberculosis (Dataset #1; N = 452,264)
Sherlock Bayesian-based analysis
MAGMA gene-based analysis
Integrated GWAS with blood eQTL data of Zeller et al. (Dataset #3; N = 1490)
Gene-based analysis of GWAS data on tuberculosis (Dataset #1)
Two independent eQTL datasets from GTEx based on blood (Dataset #4; N = 369) and lung (Dataset #5; N = 383) used to validate the identified genes
Gene-based analysis of fake tuberculosis GWAS summary data (Dataset #2; N = 3,960), which is considered as a negative control
Multi-layers of genomics evidence
KEGG pathway analysis
GO Term analysis
KEGG disease analysis
permutation analysis
Based on comprehensive analysis, numerous pressing risk genes were identified
GGI network analysis
DGIdb analysis
DGE analysis
Co-expression analysis
Prioritization of 26 tuberculosis-associated risk genes

Figure 1. Workflow of current comprehensive genomics analysis.

For the functional enrichment analyses of these two gene sets, 40 pathways, 50 GO-terms, and 7 diseases-terms (FDR ≤ 0.05; Supplementary Tables 7–9) were significantly overrepresented by Gene set #2, as well as 29 pathways, 51 GO-terms, and 13 diseases-terms (FDR ≤ 0.05; and Supplementary Tables 10–12) were significantly enriched by Gene set #3. Furthermore, we found 10 common pathways, 18 common GO-terms, and 2 common enriched diseases (FDR ≤ 0.05; Figure 2B–2D, Tables 2, 3, and Supplementary Table 13) were significantly enriched by all the 3 independent gene sets.

MAGMA-based gene analysis for technical replication

By performing MAGMA gene-level analysis of TB-based GWAS, we identified 1,017 genes were significantly or suggestively associated with TB (Gene set #4, MAGMA-based P ≤ 0.05; Supplementary Table 14). Among them, 128 genes have been documented to be associated with TB or at least one of other respiratory related traits or diseases in the database of GWAS Catalog (Supplementary Figure 16 and Supplementary Table 14). Compared with 3 independent Sherlock-identified gene sets, 18 of 26 common genes were significantly replicated by using MAGMA analysis (Figure 3A and Table 1). As a negative control, genes identified from MAGMA analysis on fake TB (Gene set #5) have obviously lower overlap with Sherlock-identified common genes than those with genes from MAGMA analysis on TB (Figure 3B and Table 1). In addition, we used the MAGMA tool to perform a pathway enrichment analysis based on the KEGG pathway resource. We found that 19 pathways showed significant or suggestive enrichment (P < 0.05). Of them,

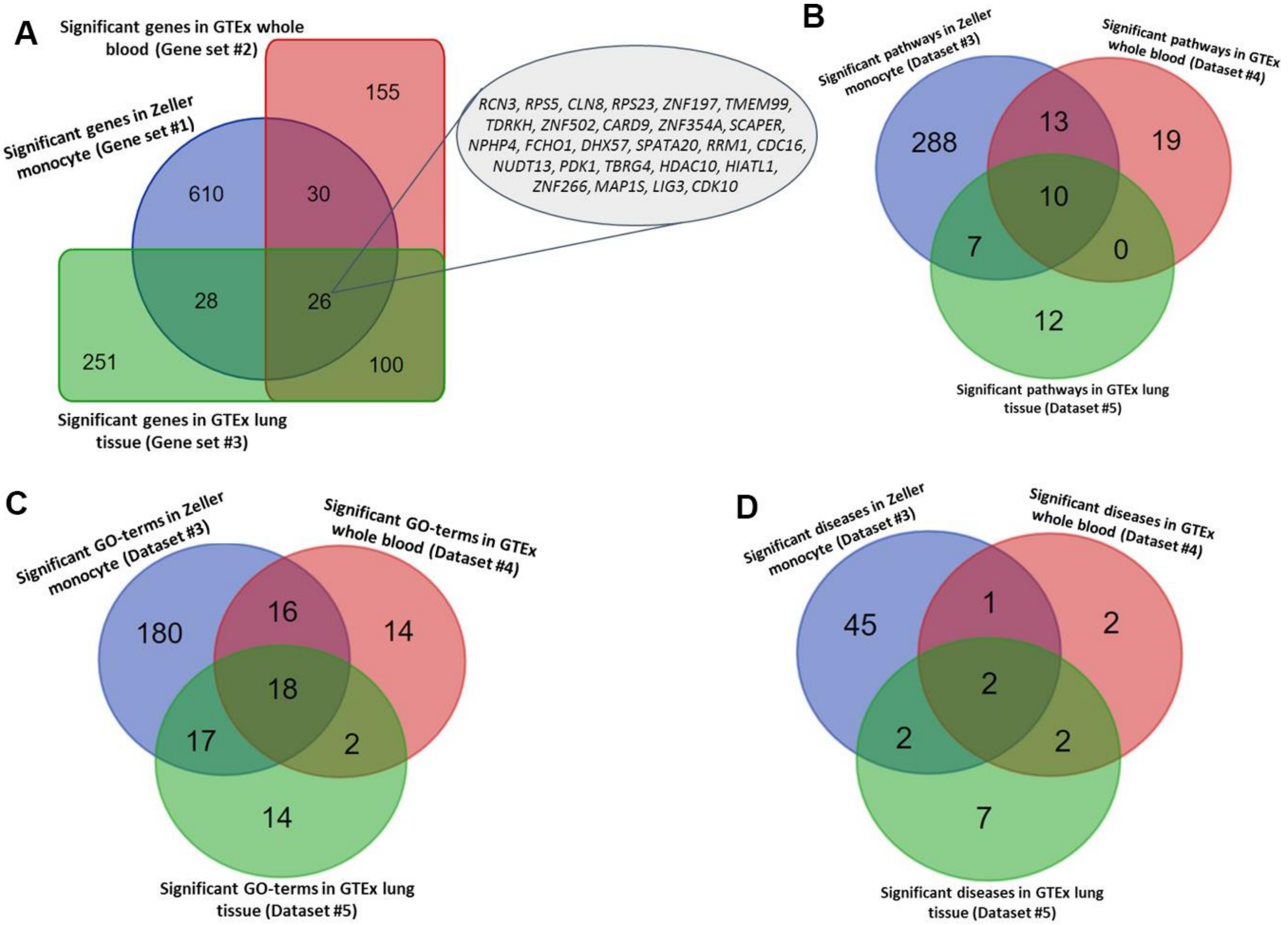

A
Significant genes in GTEx whole blood (Gene set #2)
Significant genes in Zeller monocyte (Gene set #1)
155
610
30
28
26
251
100
Significant genes in GTEx lung tissue (Gene set #3)
RCN3, RPS5, CLN8, RPS23, ZNF197, TMEM99, TDRKH, ZNF502, CARD9, ZNF354A, SCAPER, NPHP4, FCHO1, DHX57, SPATA20, RRM1, CDC16, NUDT13, PDK1, TBRG4, HDAC10, HIATL1, ZNF266, MAP1S, LIG3, CDK10
B
Significant pathways in Zeller monocyte (Dataset #3)
Significant pathways in GTEx whole blood (Dataset #4)
288
13
19
10
7
0
12
Significant pathways in GTEx lung tissue (Dataset #5)
C
Significant GO-terms in Zeller monocyte (Dataset #3)
Significant GO-terms in GTEx whole blood (Dataset #4)
180
16
14
18
17
2
14
Significant GO-terms in GTEx lung tissue (Dataset #5)
D
Significant diseases in Zeller monocyte (Dataset #3)
Significant diseases in GTEx whole blood (Dataset #4)
45
1
2
2
2
2
7
Significant diseases in GTEx lung tissue (Dataset #5)

Figure 2. Identified tuberculosis-related risk genes, pathways, and GO-terms. (**A**) Common significant genes identified from the Sherlock analysis based on Gene sets #1, #2, and #3. (**B**) Common significant pathways enriched by genes identified from the Sherlock analysis cross 3 gene sets (i.e., Gene sets #1, #2, and #3). (**C**) Common significant GO-terms enriched by genes identified from the Sherlock analysis cross 3 gene sets (i.e., Gene sets #1, #2, and #3). (**D**) Common significant KEGG or NHGRI GWAS Catalog diseases enriched by genes identified from the Sherlock analysis cross 3 gene sets (i.e., Gene sets #1, #2, and #3).

Table 1. Sherlock-based Bayesian genomics analysis identifies 26 candidate genes associated with tuberculosis risk.

Gene	Simulated P values in Gene set #1	Simulated P values in Gene set #2	Simulated P values in Gene set #3	MAGMA-based P values in Gene set #4	MAGMA-based P values in Gene set #5*	T-test P values in GSE133803	Anova P values in GSE1440943	Anova P values in GSE1440944	Anova P values in GSE139825
CDC16	6.21E-3	1.20E-2	1.38E-2	5.45E-3	NA	5.63E-4	8.16E-4	8.17E-2	8.34E-02
HIATL1	1.51E-2	2.05E-2	1.16E-2	0.12	0.59	1.83E-7	7.33E-4	2.17E-2	0.11
RCN3	2.01E-2	1.40E-2	7.14E-3	4.41E-3	0.81	8.73E-3	0.13	1.56E-2	0.78
FCHO1	2.93E-2	1.31E-2	1.64E-2	3.53E-2	0.80	0.27	4.01E-2	9.03E-7	7.12E-03
CDK10	3.08E-2	3.35E-2	3.70E-2	2.62E-2	NA	0.49	8.70E-3	8.92E-5	7.93E-02
SCAPER	3.60E-2	1.95E-2	1.62E-2	1.38E-2	0.20	8.84E-4	0.14	3.34E-2	0.65
LIG3	3.98E-2	1.66E-2	3.73E-2	2.91E-2	0.28	1.86E-3	8.88E-2	2.23E-2	8.15E-02
RRM1	4.82E-2	2.67E-2	2.04E-2	0.49	0.29	3.24E-3	1.29E-2	5.23E-4	0.11
PDK1	3.79E-3	2.54E-2	1.87E-3	5.61E-3	3.51E-3	0.83	1.14E-2	7.84E-4	3.77E-02
TMEM99	5.18E-3	1.86E-2	3.02E-3	1.36E-3	0.47	2.00E-3	NA	NA	0.11
SPATA20	7.80E-3	4.13E-3	3.51E-3	4.55E-4	0.39	1.98E-3	0.23	0.38	2.23E-02
TDRKH	8.18E-3	1.01E-2	8.50E-3	1.17E-2	0.15	0.83	0.14	1.45E-3	7.42E-02
NPHP4	1.15E-2	3.69E-2	2.98E-2	2.01E-2	0.35	0.32	0.42	8.73E-2	0.15
CLN8	2.10E-2	1.13E-2	1.19E-2	1.46E-2	0.40	8.18E-6	0.10	0.17	1.04E-02
DHX57	3.05E-2	1.48E-2	8.48E-3	1.19E-2	0.20	0.48	2.80E-2	0.10	0.18
RPS5	3.71E-2	4.65E-2	4.41E-2	0.19	0.93	2.11E-4	3.73E-4	8.54E-2	2.71E-04
MAP1S	4.03E-2	8.39E-3	6.70E-3	1.30E-2	6.0E-2	1.01E-2	NA	NA	0.78
HDAC10	2.34E-3	2.36E-2	2.53E-2	0.21	NA	0.42	0.15	4.23E-2	0.89
TBRG4	1.67E-2	4.53E-2	3.66E-2	0.25	0.80	0.11	0.11	2.99E-3	0.28
CARD9	1.73E-2	3.86E-2	1.74E-2	0.13	NA	5.48E-2	NA	NA	0.21
ZNF354A	1.74E-2	3.75E-2	4.14E-2	2.74E-2	0.98	0.29	NA	NA	3.80E-02
ZNF266	3.66E-2	3.94E-2	3.09E-2	1.09E-2	0.41	0.11	NA	NA	0.18
ZNF502	4.23E-2	2.18E-2	1.99E-2	1.23E-2	0.53	0.14	NA	NA	0.14
ZNF197	4.32E-2	1.81E-2	2.57E-2	9.53E-4	0.71	5.40E-2	NA	NA	6.69E-03
NUDT13	3.27E-2	3.78E-2	3.57E-2	6.0E-2	0.78	0.28	0.41	0.56	0.22
RPS23	7.88E-7	2.22E-2	1.66E-2	0.54	2.26E-2	5.45E-5	0.24	0.82	0.49

Note: NA means not available, which were largely due to that the expression levels of these genes very lower or the qualities were not feasible.
*Gene set #5 is generated from MAGMA analysis on fake tuberculosis as a negative control.

15 pathways were enriched by genes identified from Sherlock analysis in the discovery stage (P < 0.05, Supplementary Table 15).

Consistently, by using permutation analyses, genes identified from the discovery stage (Gene set #1) were significantly higher overlapped with identified genes from Gene sets #2, #3, and #4 in the replication stage than that of 100,000 times of random selections (Permuted P = 0, 0, 0 separately; Figure 3C–3E). Furthermore, there was no difference in overlap between genes from Gene set #1 with genes from Gene set #5 and genes from random selections (Permuted P = 0.32; Figure 3F). Additionally, to further determine whether these identified TB-associated genes were due to genetic determinants rather than false discoveries, we compared the results from MAGMA analysis on TB (Gene set #4) and fake TB (Gene set #5) with significant genes identified from 3 times of independent Sherlock analyses (Gene sets #1, #2, and #3) at 3 distinct P value thresholds (i.e., P = 0.05, 0.01, or 0.001), respectively. Consistently, we found that the overlapped gene rates between Sherlock-identified genes and MAGMA-identified genes were remarkably higher than that with MAGMA analysis on fake TB across 3 different thresholds (Figure 4A–4C). Together, these results further confirm that our identified genes are potentially convincing candidate genes for TB.

GGI network constructed by 26 highlighted TB-risk genes

Based on independent biological and technical replications, we highlighted 26 genes as important candidates conferring susceptibility to TB. Based on these 26 genes, we performed a GGI network enrichment analysis. Figure 5 demonstrates that most of

Table 2. 10 common pathways enriched by tuberculosis-associated genes across 3 identified gene sets.

Pathway ID	Common pathways	Gene set #1		Gene set #2		Gene set #3	
		Proportion of risk genes	FDR	Proportion of risk genes	FDR	Proportion of risk genes	FDR
R-HSA-1430728	Metabolism	4.96%	1.78E-28	1.35%	1.19E-4	1.69%	3.16E-5
R-HSA-74160	Gene expression (Transcription)	4.97%	1.06E-19	1.17%	1.90E-2	1.80%	1.24E-4
R-HSA-392499	Metabolism of proteins	3.93%	3.94E-16	1.29%	3.72E-4	1.64%	5.19E-5
R-HSA-73857	RNA Polymerase II Transcription	4.71%	8.33E-16	1.22%	1.90E-2	1.67%	1.37E-3
R-HSA-212436	Generic Transcription Pathway	4.78%	7.63E-15	1.17%	3.79E-2	1.84%	4.24E-4
R-HSA-597592	Post-translational protein modification	4.32%	4.60E-14	1.42%	1.09E-3	1.63%	1.34E-3
R-HSA-5653656	Vesicle-mediated transport	5.38%	1.36E-10	1.64%	1.68E-2	1.79%	3.09E-2
R-HSA-1643685	Disease	4.29%	3.11E-10	1.81%	1.33E-4	1.43%	4.22E-2
R-HSA-199991	Membrane Trafficking	5.23%	2.17E-9	1.74%	1.10E-2	1.74%	4.22E-2
R-HSA-382551	Transport of small molecules	3.33%	7.01E-4	1.53%	2.19E-2	1.67%	3.98E-2

Note: Proportion of risk genes: these identified risk genes accounted for the proportion of all genes in each pathway enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Table 3. 18 common GO-terms enriched by tuberculosis-associated genes across 3 identified gene sets.

GO-terms ID	GO-terms	Gene set #1		Gene set #2		Gene set #3	
		Proportion of risk genes	FDR	Proportion of risk genes	FDR	Proportion of risk genes	FDR
GO:0005622	Intracellular	4.44%	4.09E-24	1.26%	2.89E-4	1.53%	1.07E-4
GO:0110165	Cellular Anatomical Entity	3.77%	2.12E-21	0.91%	2.42E-2	1.47%	3.36E-5
GO:0044237	Cellular Metabolic Process	3.80%	4.45E-15	1.04%	1.97E-2	1.18%	3.51E-2
GO:0043227	Membrane-Bounded Organelle	3.79%	4.45E-15	1.33%	1.74E-4	1.63%	5.19E-5
GO:0043229	Intracellular Organelle	3.76%	1.97E-13	1.77%	8.74E-8	1.61%	1.36E-4
GO:0005488	Binding	3.47%	2.19E-13	1.38%	4.34E-5	1.56%	5.19E-5
GO:0005737	Cytoplasm	3.90%	5.71E-13	1.46%	1.63E-4	1.40%	8.80E-3
GO:0005515	Protein Binding	3.85%	5.71E-13	1.24%	3.39E-3	1.90%	1.41E-5
GO:1901363	Heterocyclic Compound Binding	5.17%	6.56E-12	1.60%	6.97E-3	2.09%	9.91E-4
GO:0019222	Regulation Of Metabolic Process	3.83%	5.78E-8	1.68%	4.24E-4	2.15%	5.19E-5
GO:0016787	Hydrolase Activity	4.72%	4.28E-7	1.52%	4.74E-2	1.85%	3.47E-2
GO:0031982	Vesicle	4.20%	3.60E-6	1.95%	1.57E-3	1.95%	1.14E-2
GO:0008152	Metabolic Process	2.45%	2.16E-5	1.03%	6.97E-3	1.19%	1.14E-2
GO:0005654	Nucleoplasm	4.43%	5.31E-5	2.11%	6.08E-3	2.11%	2.70E-2
GO:0000166	Nucleotide Binding	4.51%	3.62E-4	1.86%	4.95E-2	3.45%	1.24E-4
GO:0003723	Rna Binding	5.16%	1.42E-2	3.87%	6.80E-3	4.52%	3.95E-3
GO:1901265	Nucleoside Phosphate Binding	3.35%	2.00E-2	2.32%	6.61E-3	2.32%	2.70E-2
GO:1990904	Ribonucleoprotein Complex	6.33%	3.85E-2	5.06%	2.19E-2	6.33%	9.07E-3

Note: Proportion of risk genes: these identified risk genes accounted for the proportion of all genes in each pathway enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

these highlighted genes were highly connected with each other. The majority of interactions in the constructed network were depended on co-expression, which accounted for 71.52% of interactions (Supplementary Table 16 and Supplementary Figure 17). For example, the hub gene of *RPS5* had co-expression evidence with *NPHP4* and *PDK1*. Furthermore, the hub gene of *RPS23* showed a genetic interaction with *SCAPER*, as well as the *SCAPER* gene interacted with *TDRKH* based on evidence of genetic interactions. It should be noted that 5 TB-associated genes of *CLN8*, *TMEM99*, *CARD9*, *SPATA20*, and *DHX57* had no interactions with other genes in this constructed network (Figure 5).

Differential gene expression analysis of these 26 highlighted genes

By utilizing the expression data of GSE133803, we performed a DGE analysis of these 26 highlighted genes and found 12 genes were significantly expressed between MTB-infected cells and controls (Figure 6A, Table 1, and Supplementary Table 17); for example, *CDC16* (P = 5.63 × 10^{-4}), *RPS5* (P = 2.11 × 10^{-4}), *HIATL1* (P = 1.83 × 10^{-7}), and *RPS23* (P = 5.45 × 10^{-5}). 2 genes of *CARD9* (P = 0.055) and *ZNF197* (P = 0.054) were identified to be suggestively significant (Supplementary Table 17). In light of most of interactions among genes were derived from co-expression (71.52%) in our GGI network analysis, we further conducted a Pearson correlation analysis to uncover whether the co-expression patterns of these highlighted genes altered or not between MTB-infected cells and uninfected cells. We detected that there was remarkable differences in co-expression patterns among 26 highlighted genes between MTB-infected cells and uninfected cells (Figure 6B, 6C and Supplementary Tables 18, 19). For example, the positive correlation coefficient of *RPS23* with *NUDT13* was decreased from 0.99 in uninfected cells to 0.34 in MTB-infected cells. Furthermore, the correlation coefficient between *RCN3* and *CLN8* was changed from 0.41 in uninfected cells to -0.92 in MTB-infected cells.

By analyzing the GSE1440943 dataset based on blood samples, 8 significant genes and 1 suggestive genes showed differential expressions between MTB-infected mice with 5 different time points and uninfected mice (Table 1, Figure 7 and Supplementary Figure 18). Furthermore, we analyzed the GSE1440944 dataset

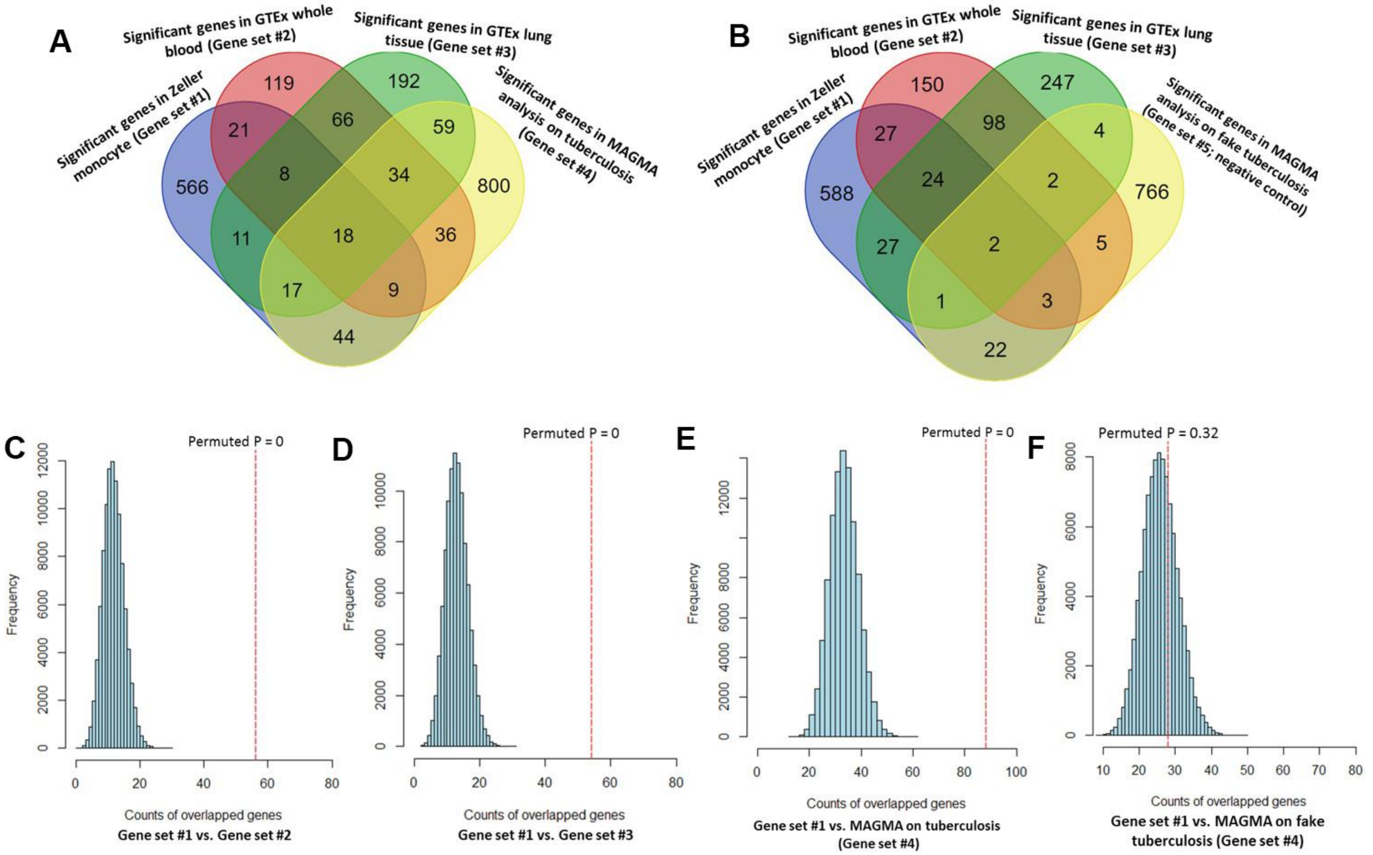


Figure 3. Consistent evidence support Sherlock-identified genes implicated in tuberculosis (TB). (**A**) Venn diagram shows that common genes between Sherlock-identified genes of Gene sets #1, #2, and #3 and MAGMA-identified genes on TB (Gene set #4). (**B**) Venn diagram shows that common genes between Sherlock-identified genes of Gene sets #1, #2, and #3 and MAGMA-identified genes on fake TB (Gene set #5). (**C**–**F**) Computer-based permutation analysis; (**C**) for the overlap between Gene set #1 and Gene set #2; (**D**) for the overlap between Gene set #1 and Gene set #3; (**E**) for the overlap between Gene set #1 and Gene set #4; (**F**) for the overlap between Gene set #1 and Gene set #5.

based on lung tissues and identified 11 significant genes and 3 suggestive genes have differential expressions between MTB-infected with 5 different time points and uninfected mice (Table 1, Figure 7 and Supplementary Figure 19). There existed a consistent finding of significant genes between both datasets (Table 1 and Figure 7). For example, 2 genes of *FCHO1* and *RPS5* showed significantly higher expression in MTB-infected mice at 5 time points than in uninfected mice in both blood (Figure 7A: Anova P = 0.04; and Figure 7C: Anova P = 3.73 × 10^{-4}) and lung samples (Figure 7B: Anova P = 9.03 × 10^{-7} and Figure 7D: Anova P = 0.085). Consistently, by using the dataset of GSE139825 based on human alveolar macrophages, 7 significant genes (Anova P < 0.05; Supplementary Figure 20) and 4 suggestive genes (Anova P < 0.1; Supplementary Figure 21) showed differential expressions between TB group and control group. For example, *RPS5* (Anova P = 2.74 × 10^{-4}) and *FCHO1* (Anova P = 7.12 × 10^{-3}).

Identification of risk eSNPs among these 26 highlighted TB-risk genes

For each highlighted gene, there were multiple eSNPs showing significant association with the expression of this gene and TB risk simultaneously (Supplementary Table 20). To name a few, with respect to the gene of *CDC16*, 2 cis-regulatory eSNPs of rs7987202 (P_{eQTL} = 4.70 × 10^{-13} and P_{GWAS} = 2.53 × 10^{-3}) and rs9590408 (P_{eQTL} = 3.79 × 10^{-49} and P_{GWAS} = 2.02 × 10^{-3}) and 1 trans-regulatory eSNPs of rs948182 (P_{eQTL} = 4.13 × 10^{-6} and P_{GWAS} = 2.01 × 10^{-2}) were identified. 1 eSNP of rs3118766 (P_{eQTL} = 5.45 × 10^{-7} and P_{GWAS} = 7.32 × 10^{-4}) has cis-regulatory effect on *HIATL1* gene. 3 eSNPs of rs2946863 (P_{eQTL} = 3.26 × 10^{-7} and P_{GWAS} = 6.42 × 10^{-3}), rs2878342 (P_{eQTL} = 2.70 × 10^{-12} and P_{GWAS} = 3.82 × 10^{-3}), rs3810194 (P_{eQTL} = 6.65 × 10^{-6} and P_{GWAS} = 1.43 × 10^{-2}) have cis-regulatory functions on *RCN3* gene. Furthermore, with regard to *FCHO1* gene, 3 cis-eSNPs (rs4280376: P_{eQTL} = 1.95 × 10^{-10} and P_{GWAS} = 5.86 × 10^{-2}, rs4808683: P_{eQTL} = 9.98 × 10^{-15} and P_{GWAS} = 3.39 × 10^{-3}, rs8107550: P_{eQTL} = 2.85 × 10^{-6} and P_{GWAS} = 4.40 × 10^{-3}) and 1 trans-eSNP (rs1058348: P_{eQTL} = 3.24 × 10^{-7} and P_{GWAS} = 2.78 × 10^{-2}) were identified.

DISCUSSION

TB is an infectious disease and remains a leading public health problem in developing world and an increasing threat in developed countries [1–3]. There were

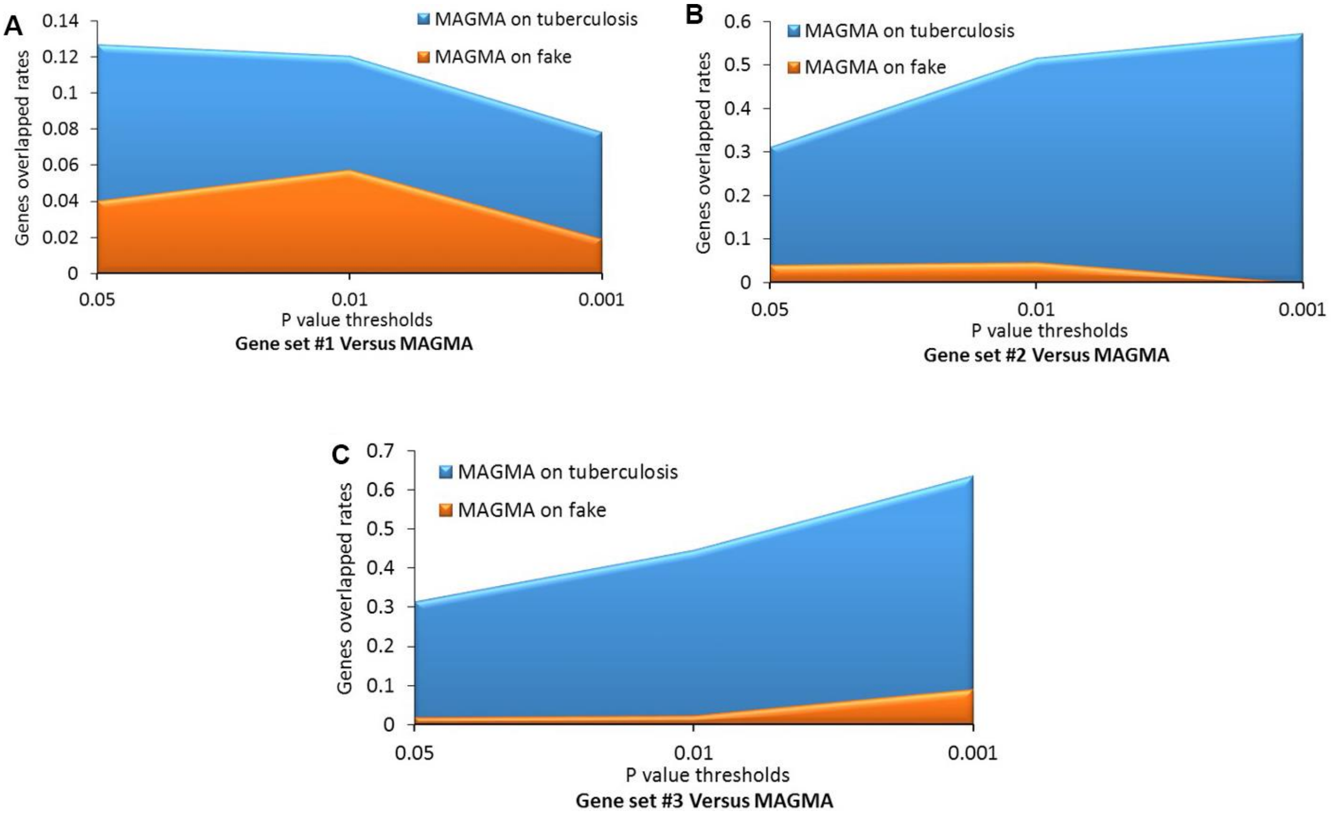


Figure 4. Comparative analysis of genes identified from the Sherlock analysis with that from the MAGMA analysis of tuberculosis (TB) and fake TB. (**A**) Gene set #1 versus MAGMA; (**B**) Gene set #2 versus MAGMA; (**C**) Gene set #3 versus MAGMA.

approximately one third of the world populations estimated to be infected with the TB pathogen, *Mycobacterium tuberculosis*, but only about 10% of infected individuals eventually become active TB patients [3], suggesting genetic heterogeneity potentially contribute differential susceptibility to infection. Consistently, host genetic factors having important roles in determining susceptibility to *Mycobacterium tuberculosis* are well-indicated by twin, family linkage, candidate gene analyses, and mouse models [6–8, 34, 35]. Hitherto, more than 10 GWASs on TB have been reported [17–25], and many TB-associated genetic loci have been identified and documented in the NHGRI GWAS Catalog [36]. Nevertheless, some identified genetic variants were hard to be replicated [37, 38], which could be attributed to the genetic heterogeneity of samples used, underpowered GWASs, or small effect sizes of variants. Lack of replications lead to these GWAS-identified SNPs have not translated into clinical practice so far. Thus, there exists a strong interest in improving our understanding of the pathophysiological mechanisms of genetic components on TB with the use of advanced genetics- and genomics-based methods.

For the method of GWAS, it has been widely used to identify genetic loci conveying risk to complex diseases [39]. With the use of GWAS, a growing and large number of SNPs have been documented to be of significant associations with hundreds of phenotypes [36, 40, 41]. However, due to the stringent correction for multiple testing of GWAS, many SNPs with small-to-moderate effects which not reach a genome-wide significance but have important functional roles were largely neglected. In light of many SNP-SNP pairs have highly LD accompanied with similar level of significance when calculate the P-values, thus to pinpoint the exact causative variants of these GWAS-identified associations is still a big challenge. Generally,

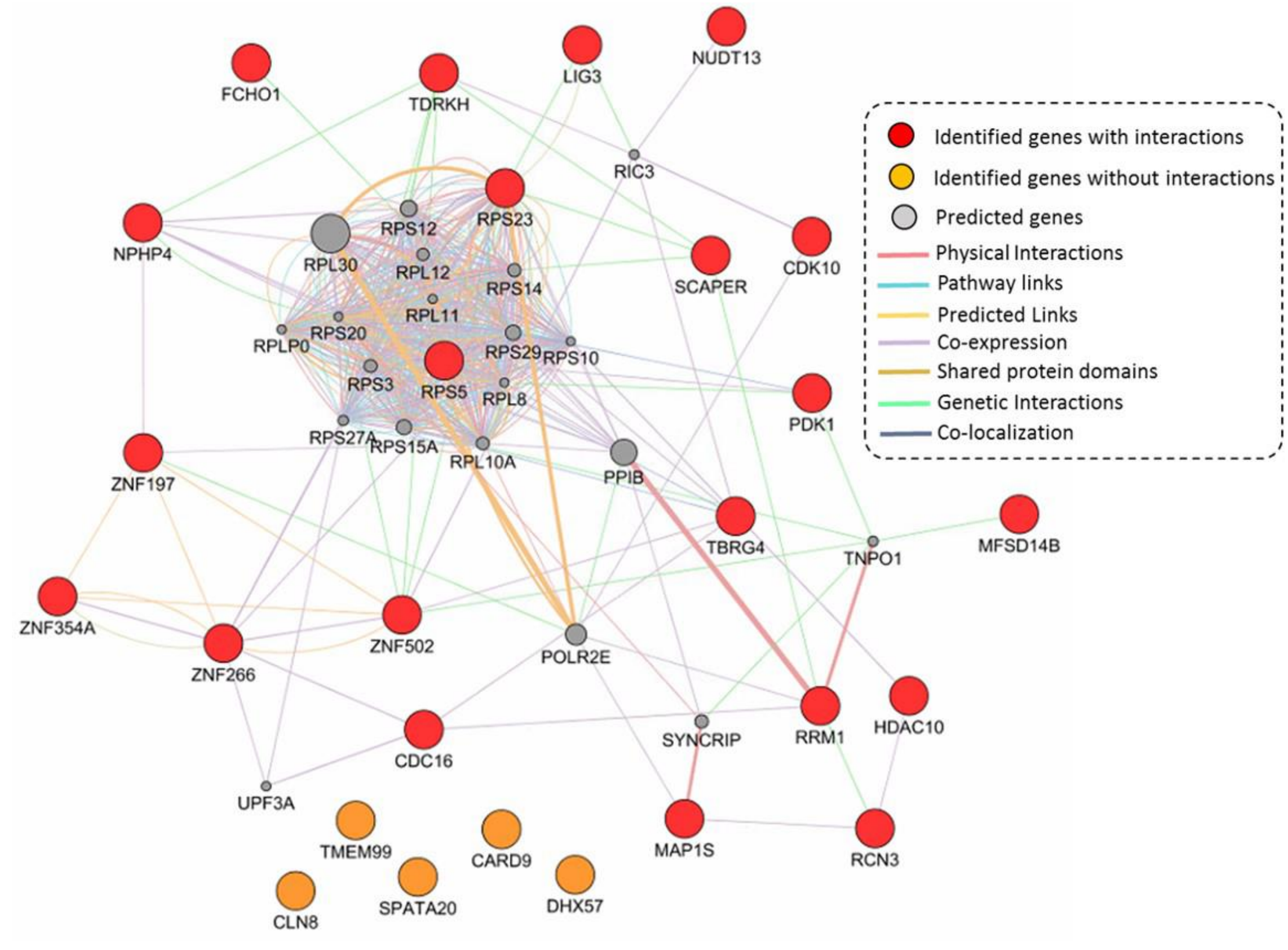


Figure 5. Constructed GGI network by using identified 26 TB-associated genes. These 21 identified genes with interactions are marked with red color, 5 identified genes without interactions are marked with orange color, and 20 predicted genes are colored with gray color.

a large proportion of identified risk SNPs were annotated into noncoding regions of genome in GWASs on complex diseases including TB [8, 26, 28], indicating these SNPs may influence the gene expression levels by cis- and/or trans-regulatory mechanisms to involve in TB risk. Considerable work on exploring the links between genetic variants and RNA expression is interested and warranted. For our current study, we conducted an integrative genomics analysis by combining multi-layers of omics data, including genomics, eQTL, RNA expression, eSNPs, and gene-gene interactions, to identify more susceptible SNPs, genes, and pathways implicated in the etiology of TB risk.

We first performed a Sherlock-based Bayesian analysis through incorporating a large-scale GWAS summary dataset on TB with a discovery eQTL dataset to identify susceptible genes and eSNPs. At this discovery stage, a number of 694 significant genes were identified to be associated with TB. Of note, we noticed that 49 genes of 694 significant genes have been documented to be associated with TB, lung-related or respiratory-related diseases in earlier studies. For example, 4 genes of *C2CD2* [20], *HLA-DRB6* [42], *LPCAT2* [43], and *HLA-DQB1* [42, 44] were associated with TB risk, and *RUNX* [45] showed association with asthma or allergic disease risk. In addition, *RUFY1*, *DEPDC7*, and *IRF4* were reported to be involved in lung cancer [46]. To validate

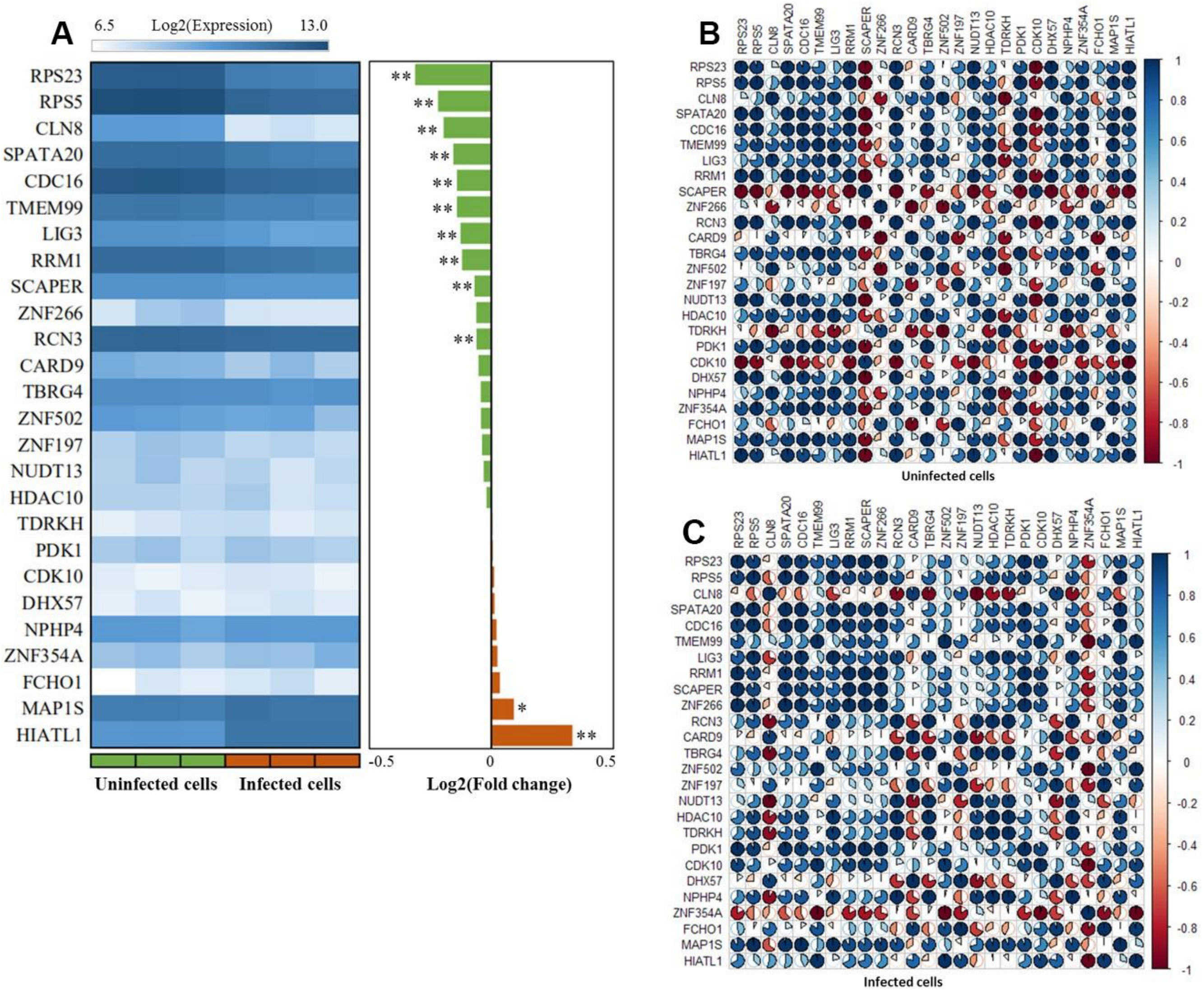


Figure 6. The expression patterns of these 26 risk genes between infected cells and uninfected cells based on the GSE133803 dataset. (**A**) Heatmap showing the expression levels of 26 risk genes between infected cells and uninfected cells; * represents the t-test P value < 0.05, ** represents the t-test P value < 0.01; (**B**) The co-expression patterns of 26 risk genes based on the Pearson correlation analysis in uninfected cells; (**C**) The co-expression patterns of 26 risk genes based on the Pearson correlation analysis in infected cells.

the findings in the discovery stage, we conducted Sherlock analysis based on 2 independent eQTL datasets and found that there were 26 genes significantly replicated. Of note, 1 common gene of *CARD9* was previously identified to be associated with lung function (FVC) [47, 48]. Additionally, based on these significantly identified genes in both discovery and replication stage, we found 10 important biological pathways implicated in TB risk, providing a mechanistic clue for performing molecular studies for TB. Based on multiple layers of protein and genomics evidence deposited in public databases, we found these 26 genes were highly connective with each other in the constructed network, indicating these genes jointly impact on TB susceptibility. Noteworthy, all these 26 genes encompassed at least one eSNPs which are significantly associated with both expression of gene and TB risk. Meanwhile, we also utilized MAGMA analysis of GWAS on TB as an independent technical validation. Interestingly, 18 of 26 (69.23%) common genes were significantly replicated in MAGMA analysis.

Since there existed a high proportion of co-expression links among these 26 genes in our constructed network, we inferred that the co-expression patterns might be changed according to the different disease status of TB. In line with our speculation, the co-expression patterns among 26 genes were prominently altered between MTB-infected and uninfected cells. By performing the DGE analysis based on 4 independent expression datasets, we found that 21 of 26 genes had significantly differential expressions between TB group and control

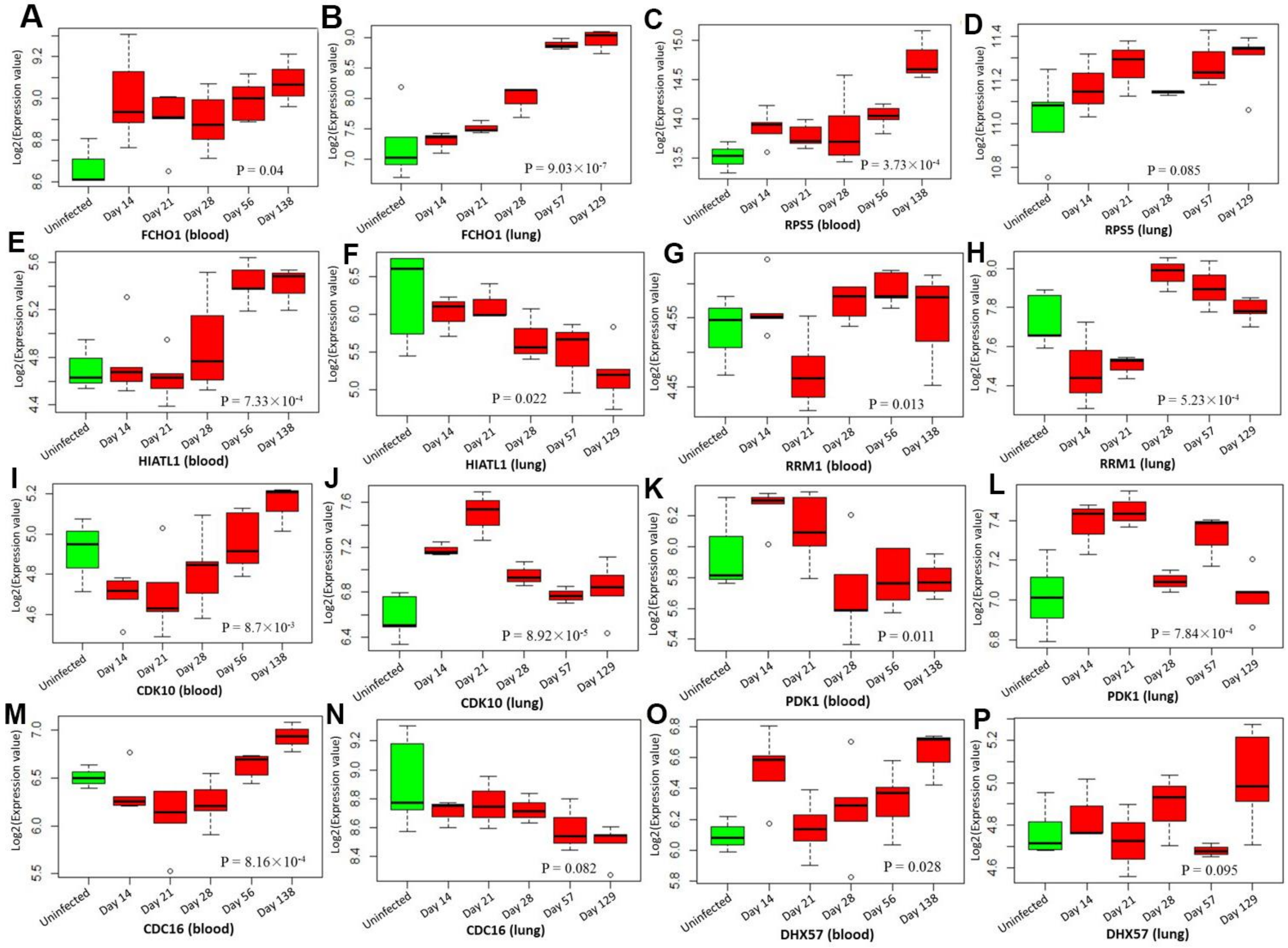


Figure 7. Boxplots show the differential expression levels of tuberculosis-risk genes between uninfected mice and infected mice with 5 distinct time points based on two GSE1440943 (blood) and GSE1440944 (lung) datasets. (**A**) *FCHO1* for blood; (**B**) *FCHO1* for lung; (**C**) *RPS5* for blood; (**D**) *RPS5* for lung; (**E**) *HIATL1* for blood; (**F**) *HIATL1* for lung; (**G**) *RRM1* for blood; (**H**) *RRM1* for lung; (**I**) *CDK10* for blood; (**J**) *CDK10* for lung; (**K**) *PDK1* for blood; (**L**) *PDK1* for lung; (**M**) *CDC16* for blood; (**N**) *CDC16* for lung; (**O**) *DHX57* for blood; (**P**) *DHX57* for lung. P values were generated by Anova test.

group in mesenchymal stem cells, mice blood and lung tissues, as well as human alveolar macrophages; such as, *CDC16*, *HIATL1*, *RCN3*, *FCHO1*, and *RPS5*. These results are consistent with the primary assumption of the Sherlock-based Bayesian inference algorithm that aberrant expression of genes are more likely to convey risk to complex diseases [29]. For the original GWAS reported by Canela-Xandri and coworkers [49], there was no SNP reaching genome-wide significance to be associated with TB. Due to the strict genome-wide significance threshold applied by the GWAS, numerous susceptible genes and SNPs with small-to-moderate effects on TB being neglected. As the method of reported studies [50–53], based on the two-stage designed integrative genomics analysis, we highlighted 26 genes with multiple eSNPs as important candidates for revealing the pathogenesis of TB risk.

The protein of CDC16, encoded by the highlighted gene of *CDC16*, is a protein ubiquitin ligase and is one of components of the multiprotein APC complex. CDC16 has been reported as a binding partner of chitooligosaccharide deacetylase homolog (YDJC) in breast cancer cells [54]. Overexpression of CDC16 enhanced the ubiquitination of YDJC in an orthotopic mouse model [55]. Kim and coworkers reported that suppression of YDJC or boosting of CDC16 interaction with YDJC might be implicated in the progression of lung cancer [56]. Previous studies have reported that TB is considered as a potential risk factor for the development of lung cancer [57, 58]. In our current analysis, there were 3 eSNPs (rs7987202, rs9590408, and rs948182) with cis- or trans-regulatory effects in *CDC16* gene identified to be associated with TB risk. As for the highlighted gene of *RCN3*, it encodes reticulocalbin 3 (Rcn3), which is an endoplasmic reticulum lumen protein mapped in the secretory pathway. Jin and colleagues [59] showed that Rcn3 protein has an indispensable physiological role in the maturation of perinatal lung and neonatal respiratory adaption by using an Rcn3 knockout mouse model. Furthermore, they demonstrated that upregulated expression of Rcn3 in maturating alveolar epithelial type II cells (AECIIs) seems to have a contribution to the survival and wound healing of AECIIs, indicating Rcn3 has a critical part in mediating pulmonary injury remodeling [60]. Hou and coworkers [61] suggested that there is a potential association between the depletion of Rcn3 protein and development of non-small cell lung cancer. We noticed 3 eSNPs of rs2946863, rs2878342, and rs3810194 in *RCN3* were associated with TB risk in our integrative genomics analysis.

Some limitations of our current analysis need to comment. Although we employed multiple omics datasets, there were other datasets missed. For example, in our current study, gene expression datasets were mainly based on blood samples. Only two datasets of eQTL Dataset #5 and GSE1449044 were derived from mice lung tissue. More molecular studies for exploring the functions of genes identified from our current analysis are warrant to assess tissues that could be more related to the etiology of TB, for example, human lung tissue. Furthermore, due to the heterogeneity of different datasets, we applied different correction methods for multiple testing at each individual dataset; such as, simulated P value < 0.05 for Sherlock Bayesian analysis, false discovery rate (FDR) < 0.05 for pathway enrichment analysis, and empirical P value < 0.05 for 100,000 times of *in silico* permutation analysis. Additionally, association signals of eSNPs from current integrative genomics analysis were obtained in the European population. We did not determine whether the associations exist in other ancestries. Future studies are warrant to evaluate the regulatory effects of eSNPs using genotype and expression data from other ethnic populations. In addition, although a total of 452,264 samples were included for our genomics analysis, it should be noted that our chosen controls might contain persons have latent infection or they are the susceptible host that have never been exposed to TB, which might result in the power loss for genome-wide association analysis of this dataset.

In conclusion, in the present study, we conducted a systematically integrative genomics analysis to identify TB-associated risk SNPs, susceptible genes, and biological pathways. By incorporating GWAS summary statistics with eQTL data, we offered a reasonable explanation of the regulatory functions of intronic SNPs for TB. With the use of detailed topology data on gene-gene and gene-drug information, we highlighted 26 candidate genes for TB susceptibility, which were difficult to be identified by any single GWAS. More molecular experiments are warranted to be performed for identification of the biological mechanisms of these prioritized genes implicated in the aetiology of developing TB.

MATERIALS AND METHODS

Sherlock-based integrative genomics analysis

To exploit whether abnormal expression of gene with susceptible SNPs implicated in the etiology of TB risk, we performed a Sherlock-based integrative genomics analysis to integrate GWAS summary-based SNP information with eQTL [29]. The Sherlock integrative analysis based on a Bayesian algorithm is intended to cluster multiple lower-confidence SNPs from GWAS with expression QTL data to reveal authentic susceptible genes involved in complex diseases. In our

Sherlock analysis, SNP rs IDs and P values extracted from GWAS summary-level statistics were utilized as an input list. The definition of expression-associated SNPs (i.e., eSNPs) are that SNPs show significant associations with TB risk and meanwhile have cis- or trans-regulatory effects on expression levels of interested genes. There exists 3 potential scenarios: 1) A positive score would be recorded based on a specific eSNP shows a significant association with TB; 2) A negative score would be recorded based on a specific eSNP shows a non-significant association with TB; 3) No score would be recorded based on an SNP was not eSNP but shows a significant association with TB. The summed score of a specific gene was based on the number of eSNPs with integrative evidence from both GWAS and eQTL data. The logarithm of the Bayes Factor (LBF) is generated as a crucial indicator to predict TB-associated functionally-important genes. The significance of Sherlock Bayesian algorithm is assessed by using a simulation analysis, and $P < 0.05$ is considered to be significant.

Dataset #1 for GWAS summary statistics on TB

The Dataset #1, the large-scale GWAS summary dataset on TB [49], was downloaded from the UK-Biobank database (Fields: 20002; Field codes: 1440). There were 452,264 subjects with 2,219 patients included in the GWAS. The Affymetrix UK BiLEVE Axiom array and the Affymetrix UK Biobank Axiom array were utilized for obtaining the genotypes of all subjects. There were 62,394 genotyped variants passed quality control. Moreover, based on the UK10K [62], 1,000 Genome [63], and Haplotype Reference Consortium [64] projects as genomics references, all genotyped variants were used for imputation to extend more variants. In the current investigation, we defined two filtering criteria for choosing high quality variants: 1) if variants are genotyped, these variants with minor allele frequency (MAF) > 10^{-4} are included; 2) if variants are imputed, these variants with MAF > 10^{-4} and imputation score > 0.9 are included. After strictly filtering, a number of 13,805,935 SNPs are qualified for subsequent genomics integrative analysis.

Dataset #2 for GWAS dataset on fake TB

To ensure identified TB-risk genes were due to genetic determinants instead of random events, we constructed a fake TB-based GWAS through using a reported GWAS dataset (N = 3,960) [65]. We used the function of RANDBETWEEN in the Microsoft Excel to randomly generate and assign the phenotype of TB or control to these 3,960 individuals. In view of there is no true genetic effect of fake TB, the sample size of constructed GWAS is not a big issue. Thus, we used this constructed GWAS dataset as a negative control to re-perform genomics analysis by using the software of PLINK v1.07 based on the addictive genetic model.

Dataset #3 for eQTL dataset reported by Zeller and coworkers

Here we downloaded the monocyte eQTL data reported by Zeller and colleagues [66], which is used as a discovery dataset for the Sherlock Bayesian genomics analysis. For this eQTL dataset, 1,490 subjects with DNA and RNA samples were enrolled from the Gutenberg Heart Study (GHS). The Affymetrix Genome-wide Human SNP Array 6.0 was utilized to obtain the genotypes of subjects, and the Illumina HT-12 v3 BeadChip was utilized to obtain RNA expression abundances. After stringently excluding, a number of 675,350 SNPs and 12,808 genes were qualified for eQTL analysis and subsequent Sherlock analysis. For more detailed characteristics on this dataset, please refers to the original study [66].

Datasets #4 and #5 for eQTL datasets from the GTEx database

Furthermore, we used two eQTL datasets on whole blood (Dataset #4; N = 369) and lung tissue (Dataset #5; N = 383) from the resource of Genotype-Tissue Expression project (GTEx v7) as an independent replication to conduct Sherlock analysis with same parameters. As for the resource of GTEx [67–69], nearly 1,000 subjects with 54 non-diseased tissues were utilized to collect samples for whole genome sequencing, whole exome sequencing, and RNA sequencing, which can be used for integrative genomics analysis to explore the relationship between genetic variants and expression levels of interested genes across multiple tissues. Multi-layers of omics data including gene expression and QTL data can be obtained through the GTEx Portal (https://www.gtexportal.org/home/).

Gene-based analysis by using MAGMA tool

To further replicate the findings identified from the Sherlock analysis, we conducted a gene-based analysis of GWAS on TB by applying the Multi-marker Analysis of GenoMic Annotation (MAGMA) [70]. Here, we used GWAS-relevant SNP rs IDs and SNP P values as an input list for MAGMA analysis. To improve the mapping of SNPs across different files and reference data, we used the SNP synonym file encompassing lists of synonymous SNP rs IDs that refer to the same SNP on the basis of the resource of dbSNP database release 151. By using multiple regression method, we attempted to discover multi-variant aggregated genetic effects by incorporating SNP-SNP linkage disequilibrium (LD) information, which is reference to the 1,000 Genomes European Panel Phase 3. The definition of the SNP set of each gene is that the SNP located in the gene body or within extended +/-20 kb downstream or upstream of the gene, and the locations of SNPs are based on the Human Genome Build 37. In addition, based on the KEGG pathway resource, we used the MGMA tool to conduct a pathway-based enrichment analysis.

In silico permutation analysis

By using the Sherlock Bayesian and MAGMA analysis, 5 gene sets were identified to be associated with TB risk; namely Gene set #1 from discovery stage (Dataset #3), Gene set #2 from replication stage (Dataset #4), Gene set #3 from replication stage (Dataset #5), Gene set #4 from MAGMA analysis on TB (Dataset #1), and Gene set #5 from MAGMA analysis on fake TB (Dataset #2). Based on these 5 gene sets, we carried out serial *in silico* permutation analyses with 100,000 times of random trial [71]. In first step of this permutation analysis, the number of overlapped genes between Gene set #1 with other gene sets ($N_{\text{observation}}$) were counted separately. Second, the background genes of each gene set was treated as a gene pool, which could be used for random selections. The number of background genes (N_{total}) were 5,786, 7,452, 18,318, and 17,565 for Gene sets #2, #3, #4, and #5, respectively. By randomly picking the same number as the significant genes in Gene sets #2, #3, #4, and #5 from background genes (N_{total}) respectively, via 100,000 times of repeat, we calibrated the count of genes overlapped with these significant genes of Gene set #1(N_{random}). Finally, we calculated the number of times $N_{\text{random}} \leq N_{\text{observation}}$ and divided by 100,000 to obtain an empirical permuted P value. P value less than 0.05 is considered to be of significance. The density plot of each analysis was generated by using the R platform.

Functional enrichment analysis by using KOBAS tool

We carried out functional enrichment analyses with the use of the web-access tool of KOBAS version 3.0 [72]. The tool of KOBAS (http://kobas.cbi.pku.edu.cn/kobas3), which is depended on the machine learning-based called Combined Gene set analysis incorporating Prioritization and Sensitivity (CGPS) [73], is designed to analyze protein or gene functional annotation and functional gene set enrichment. With respect to gene set enrichment analysis, the method of KOBAS can accept either gene list or gene expression data as a submitted file. In our current analysis, we used identified TB-associated genes from 3 times of Sherlock analyses (i.e., Gene sets #1, #2, and #3) as 3 lists of submitted genes for the KOBAS tool to calculate significantly enriched gene sets, including gene set related name, enrichment score, raw P values and corrected P values. There were 3 types of databases used in our analyses: 1) Biological pathways: Reactome pathway, KEGG pathway, PANTHER pathway, and BioCyc pathway; 2) Gene Ontology (GO) terms; 3) Diseases: OMIM, NHGRI GWAS Catalog, and KEGG disease. The statistical significance is corrected by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

GeneMANIA-based GGI network analysis of risk genes

We used the bioinformatics tool of GeneMANIA [74] to conduct a gene-gene interaction (GGI) network-based analysis for identifying collective interaction patterns of the identified TB-associated genes and predicted genes with similar functions or co-expressions. We used these highlighted risk genes to query the large database of documented genomics and proteomics data. By using a guilt-by-association approach, the GeneMANIA tool based on multi-layers of supportive evidence including co-expression links, shared protein domains, genetic interactions, pathway links, co-localization, physical interactions, and predicted links, is designed to quickly and effectively predict the molecular functions and biological interactions of submitted genes. The GGI network is visualized by using the Cytoscape network visualization and analysis platform [75].

Differential expression patterns of identified genes

To determine whether abnormal alterations in RNA expression levels of highlighted TB-risk genes, we downloaded 4 independent gene expression datasets from the database of the NCBI's Gene Expression Omnibus (GEO). The accession numbers of 4 expression datasets were GSE133803, GSE140943, GSE140944, and GSE139825. For GSE133803, the dataset was designed to analyze the mesenchymal stem cell gene expression level upon *Mycobacterium tuberculosis* (MTB) infection. RNA samples were obtained from MTB-infected mesenchymal stem cells (N = 3) and compared with that of uninfected mesenchymal stem cells (N = 3). The Illumina Human HT-12 V4.0 expression BeadChip was used to obtain the genome-wide gene expression profiles for all samples.

As for two datasets of GSE1440943 and GSE1440944, they were designed to characterize global transcriptional responses to MTB infection in different mouse models. The samples of GSE1440943 were based on blood samples obtained from BALB mice infected with low dose of MTB H37Rv, collected at 4 distinct time points Day 14 (N = 5), Day 21 (N = 5), Day 56 (N = 5), and Day 138 (N = 3) after MTB infection and uninfected control mice (N = 3). Similarly, the samples of GSE1449044 were based on lung tissues obtained from BALB mice infected with low dose of MTB H37Rv, collected at 4 distinct time points Day 14 (N = 3), Day 21 (N = 3), Day 56 (N = 3), and Day 129 (N = 5) after MTB infection and uninfected controls (N = 5). The genome-wide gene expression signatures of both GSE1440943 and GSE1440944 were assessed by using the Illumina MouseWG-6 v2.0 expression BeadChip. With regard to the dataset of GSE139825, it was designed to explore the response to infection with MTB by human extrapulmonary macrophages. Total RNA samples (N = 26) were obtained from alveolar macrophages from TB patients infected with clinical isolates of MTB to compared to alveolar macrophages from control samples. The Illumina HumanHT-12 V4.0 expression beadchip was used to evaluate the genome-wide transcriptional abundance.

Statistical analysis of RNA expression data from GEO database

With regard to GSE133803 dataset, we conducted a differential gene expression (DGE) analysis. The Student's t-test is used to assess the significant differences between MTB-infected cells and uninfected cells. Based on the Pearson correlation analysis, we used the *Corrplot* package in R platform to analyze and visualize the co-expression patterns among these highlighted TB-associated genes in the dataset of GSE133803. For both GSE1440943 and GSE1440944, the ANOVA test was used to compare the statistically significant differences between MTB-infected mice and uninfected mice at 4 distinct time points. Furthermore, for GSE139825, the ANOVA test was applied to assess the significant difference among different groups. The Rscript for this analysis was uploaded into the public github website (https://github.com/mayunlong89/TB/blob/master/Anova_test.R).

Abbreviations

TB: Tuberculosis; MTB: *Mycobacterium tuberculosis*; GWAS: Genome-Wide Association Study; HLA: the Human Leukocyte Antigens; SNP: Single Nucleotide Polymorphism; eQTL: Expression Quantitative Trait Loci; eSNP: Expression-associated SNP; LBF: the Logarithm of the Bayes Factor; MAF: Minor Allele Frequency; GHS: the Gutenberg Heart Study; GTEx: the Genotype-Tissue Expression Project; MAGMA: Multi-marker Analysis of GenoMic Annotation; LD: Linkage Disequilibrium; CGPS: Combined Gene set analysis incorporating Prioritization and Sensitivity; KEGG: the Kyoto Encyclopedia of Genes and Genomes; GO: Gene Ontology; BP: Biological Process; CC: Cellular Component; MF: Molecular Function; FDR: False Discovery Rate; GGI: Gene-Gene Interaction; GEO: the database of Gene Expression Omnibus.

AUTHOR CONTRIBUTIONS

MX, ZX, JL and XP managed the reported papers searches, data collection and analysis. XM and XP wrote the first draft of the manuscript. YM conceived the study and wrote and reviewed the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank all our colleagues who give useful advices for data analysis and manuscript writing. We also thank Dr. Jing Tao from Zhejiang University for excellent editing of this manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

FUNDING

This study was supported by the China Postdoctoral Science Foundation (2018M630667). The funders had no role in the designing and conducting of this study and collection, analysis, and interpretation of data and in writing the manuscript.

REFERENCES

1. Dye C, Williams BG. The population dynamics and control of tuberculosis. Science. 2010; 328:856–61. https://doi.org/10.1126/science.1185449 PMID:20466923

2. van Tong H, Velavan TP, Thye T, Meyer CG. Human genetic factors in tuberculosis: an update. Trop Med Int Health. 2017; 22:1063–71. https://doi.org/10.1111/tmi.12923 PMID:28685916

3. Zumla A, Raviglione M, Hafner R, von Reyn CF. Tuberculosis. N Engl J Med. 2013; 368:745–55. https://doi.org/10.1056/NEJMra1200894 PMID:23425167

4. WHO. WHO Global Tuberculosis Report 2017. World Health Organization. 2017.

5. Campbell IA, Bah-Sow O. Pulmonary tuberculosis: diagnosis and treatment. BMJ. 2006; 332:1194–97. https://doi.org/10.1136/bmj.332.7551.1194 PMID:16709993

6. Comstock GW. Tuberculosis in twins: a re-analysis of the prophit survey. Am Rev Respir Dis. 1978; 117:621–24. https://doi.org/10.1164/arrd.1978.117.4.621 PMID:565607

7. van der Eijk EA, van de Vosse E, Vandenbroucke JP, van Dissel JT. Heredity versus environment in tuberculosis in twins: the 1950s United Kingdom prophit survey simonds and comstock revisited. Am J Respir Crit Care Med. 2007; 176:1281–88. https://doi.org/10.1164/rccm.200703-435OC PMID:17823356

8. Simonds B. Twin research in tuberculosis. Eugen Rev. 1957; 49:25–32. PMID:21260731

9. Salie M, Daya M, Lucas LA, Warren RM, van der Spuy GD, van Helden PD, Hoal EG, Möller M. Association of toll-like receptors with susceptibility to tuberculosis suggests sex-specific effects of TLR8 polymorphisms. Infect Genet Evol. 2015; 34:221–29. https://doi.org/10.1016/j.meegid.2015.07.004 PMID:26160538

10. Bukhari M, Aslam MA, Khan A, Iram Q, Akbar A, Naz AG, Ahmad S, Ahmad MM, Ashfaq UA, Aziz H, Ali M. TLR8 gene polymorphism and association in bacterial load in southern punjab of Pakistan: an association study with pulmonary tuberculosis. Int J Immunogenet. 2015; 42:46–51. https://doi.org/10.1111/iji.12170 PMID:25572425

11. Dittrich N, Berrocal-Almanza LC, Thada S, Goyal S, Slevogt H, Sumanlatha G, Hussain A, Sur S, Burkert S, Oh DY, Valluri V, Schumann RR, Conrad ML. Toll-like receptor 1 variations influence susceptibility and immune response to mycobacterium tuberculosis. Tuberculosis (Edinb). 2015; 95:328–35. https://doi.org/10.1016/j.tube.2015.02.045 PMID:25857934

12. Jafari M, Nasiri MR, Sanaei R, Anoosheh S, Farnia P, Sepanjnia A, Tajik N. The NRAMP1, VDR, TNF-α, ICAM1, TLR2 and TLR4 gene polymorphisms in Iranian patients with pulmonary tuberculosis: a case-control study. Infect Genet Evol. 2016; 39:92–98. https://doi.org/10.1016/j.meegid.2016.01.013 PMID:26774366

13. Torres-Juarez F, Cardenas-Vargas A, Montoya-Rosales A, González-Curiel I, Garcia-Hernandez MH, Enciso-Moreno JA, Hancock RE, Rivas-Santiago B. LL-37 immunomodulatory activity during mycobacterium tuberculosis infection in macrophages. Infect Immun. 2015; 83:4495–503. https://doi.org/10.1128/IAI.00936-15 PMID:26351280

14. Zacharia VM, Manzanillo PS, Nair VR, Marciano DK, Kinch LN, Grishin NV, Cox JS, Shiloh MU. Cor, a novel carbon monoxide resistance gene, is essential for mycobacterium tuberculosis pathogenesis. mBio. 2013; 4:e00721–13. https://doi.org/10.1128/mBio.00721-13 PMID:24255121

15. Thuong NT, Dunstan SJ, Chau TT, Thorsson V, Simmons CP, Quyen NT, Thwaites GE, Thi Ngoc Lan N, Hibberd M, Teo YY, Seielstad M, Aderem A, Farrar JJ, Hawn TR. Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles. PLoS Pathog. 2008; 4:e1000229. https://doi.org/10.1371/journal.ppat.1000229 PMID:19057661

16. Thye T, Vannberg FO, Wong SH, Owusu-Dabo E, Osei I, Gyapong J, Sirugo G, Sisay-Joof F, Enimil A, Chinbuah MA, Floyd S, Warndorff DK, Sichali L, et al, and African TB Genetics Consortium, and Wellcome Trust Case Control Consortium. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat Genet. 2010; 42:739–41. https://doi.org/10.1038/ng.639 PMID:20694014

17. Curtis J, Luo Y, Zenner HL, Cuchet-Lourenço D, Wu C, Lo K, Maes M, Alisaac A, Stebbings E, Liu JZ, Kopanitsa L, Ignatyeva O, Balabanova Y, et al. Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. Nat Genet. 2015; 47:523–27. https://doi.org/10.1038/ng.3248 PMID:25774636

18. Thye T, Owusu-Dabo E, Vannberg FO, van Crevel R, Curtis J, Sahiratmadja E, Balabanova Y, Ehmen C, Muntau B, Ruge G, Sievertsen J, Gyapong J, Nikolayevskyy V, et al. Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat Genet. 2012; 44:257–59. https://doi.org/10.1038/ng.1080 PMID:22306650

19. Oki NO, Motsinger-Reif AA, Antas PR, Levy S, Holland SM, Sterling TR. Novel human genetic variants associated with extrapulmonary tuberculosis: a pilot genome wide association study. BMC Res Notes. 2011; 4:28. https://doi.org/10.1186/1756-0500-4-28 PMID:21281516

20. Chimusa ER, Zaitlen N, Daya M, Möller M, van Helden PD, Mulder NJ, Price AL, Hoal EG. Genome-wide association study of ancestry-specific TB risk in the South African coloured population. Hum Mol Genet. 2014; 23:796–809.

https://doi.org/10.1093/hmg/ddt462 PMID:24057671

21. Mahasirimongkol S, Yanai H, Mushiroda T, Promphittayarat W, Wattanapokayakit S, Phromjai J, Yuliwulandari R, Wichukchinda N, Yowang A, Yamada N, Kantipong P, Takahashi A, Kubo M, et al. Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. J Hum Genet. 2012; 57:363–67. https://doi.org/10.1038/jhg.2012.35 PMID:22551897

22. Png E, Alisjahbana B, Sahiratmadja E, Marzuki S, Nelwan R, Balabanova Y, Nikolayevskyy V, Drobniewski F, Nejentsev S, Adnan I, van de Vosse E, Hibberd ML, van Crevel R, et al. A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. BMC Med Genet. 2012; 13:5. https://doi.org/10.1186/1471-2350-13-5 PMID:22239941

23. Grant AV, Sabri A, Abid A, Abderrahmani Rhorfi I, Benkirane M, Souhi H, Naji Amrani H, Alaoui-Tahiri K, Gharbaoui Y, Lazrak F, Sentissi I, Manessouri M, Belkheiri S, et al. A genome-wide association study of pulmonary tuberculosis in Morocco. Hum Genet. 2016; 135:299–307. https://doi.org/10.1007/s00439-016-1633-2 PMID:26767831

24. Sobota RS, Stein CM, Kodaman N, Scheinfeldt LB, Maro I, Wieland-Alter W, Igo RP Jr, Magohe A, Malone LL, Chervenak K, Hall NB, Modongo C, Zetola N, et al. A locus at 5q33.3 confers resistance to tuberculosis in highly susceptible individuals. Am J Hum Genet. 2016; 98:514–24. https://doi.org/10.1016/j.ajhg.2016.01.015 PMID:26942285

25. Sveinbjornsson G, Gudbjartsson DF, Halldorsson BV, Kristinsson KG, Gottfredsson M, Barrett JC, Gudmundsson LJ, Blondal K, Gylfason A, Gudjonsson SA, Helgadottir HT, Jonasdottir A, Jonasdottir A, et al. HLA class II sequence variants influence tuberculosis risk in populations of european ancestry. Nat Genet. 2016; 48:318–22. https://doi.org/10.1038/ng.3498 PMID:26829749

26. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009; 106:9362–67. https://doi.org/10.1073/pnas.0903103106 PMID:19474294

27. Uren C, Henn BM, Franke A, Wittig M, van Helden PD, Hoal EG, Möller M. A post-GWAS analysis of predicted regulatory variants and tuberculosis susceptibility. PLoS One. 2017; 12:e0174738. https://doi.org/10.1371/journal.pone.0174738 PMID:28384278

28. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Respir Med. 2016; 4:213–24. https://doi.org/10.1016/S2213-2600(16)00048-5 PMID:26907218

29. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, Li H. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am J Hum Genet. 2013; 92:667–80. https://doi.org/10.1016/j.ajhg.2013.03.022 PMID:23643380

30. Ayalew M, Le-Niculescu H, Levey DF, Jain N, Changala B, Patel SD, Winiger E, Breier A, Shekhar A, Amdur R, Koller D, Nurnberger JI, Corvin A, et al. Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. Mol Psychiatry. 2012; 17:887–905. https://doi.org/10.1038/mp.2012.37 PMID:22584867

31. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005; 37:710–17. https://doi.org/10.1038/ng1589 PMID:15965475

32. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, Yang J. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016; 48:481–87. https://doi.org/10.1038/ng.3538 PMID:27019110

33. Ma Y, Li J, Xu Y, Wang Y, Yao Y, Liu Q, Wang M, Zhao X, Fan R, Chen J, Zhang B, Cai Z, Han H, et al. Identification of 34 genes conferring genetic and pharmacological risk for the comorbidity of schizophrenia and smoking behaviors. Aging (Albany NY). 2020; 12:2169–225. https://doi.org/10.18632/aging.102735 PMID:32012119

34. Apt A, Kramnik I. Man and mouse TB: contradictions and solutions. Tuberculosis (Edinb). 2009; 89:195–98. https://doi.org/10.1016/j.tube.2009.02.002 PMID:19345146

35. Möller M, Hoal EG. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. Tuberculosis (Edinb). 2010; 90:71–83.

https://doi.org/10.1016/j.tube.2010.02.002 PMID:20206579

36. Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, Suveges D, Vrousgou O, Whetzel PL, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019; 47:D1005–12. https://doi.org/10.1093/nar/gky1120 PMID:30445434

37. Zheng R, Li Z, He F, Liu H, Chen J, Chen J, Xie X, Zhou J, Chen H, Wu X, Wu J, Chen B, Liu Y, et al. Genome-wide association study identifies two risk loci for tuberculosis in han Chinese. Nat Commun. 2018; 9:4072. https://doi.org/10.1038/s41467-018-06539-w PMID:30287856

38. Miao R, Ge H, Xu L, Sun Z, Li C, Wang R, Ding S, Yang C, Xu F. Genetic variants at 18q11.2 and 8q24 identified by genome-wide association studies were not associated with pulmonary tuberculosis risk in Chinese population. Infect Genet Evol. 2016; 40:214–218. https://doi.org/10.1016/j.meegid.2016.03.005 PMID:26964908

39. Bush WS, Moore JH. Chapter 11: genome-wide association studies. PLoS Comput Biol. 2012; 8:e1002822. https://doi.org/10.1371/journal.pcbi.1002822 PMID:23300413

40. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res. 2017; 45:D896–901. https://doi.org/10.1093/nar/gkw1133 PMID:27899670

41. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–06. https://doi.org/10.1093/nar/gkt1229 PMID:24316577

42. Qi H, Zhang YB, Sun L, Chen C, Xu B, Xu F, Liu JW, Liu JC, Chen C, Jiao WW, Shen C, Xiao J, Li JQ, et al. Discovery of susceptibility loci associated with tuberculosis in han Chinese. Hum Mol Genet. 2017; 26:4752–63. https://doi.org/10.1093/hmg/ddx365 PMID:29036319

43. Schurz H, Kinnear CJ, Gignoux C, Wojcik G, van Helden PD, Tromp G, Henn B, Hoal EG, Möller M. A sex-stratified genome-wide association study of tuberculosis using a multi-ethnic genotyping array. Front Genet. 2019; 9:678. https://doi.org/10.3389/fgene.2018.00678 PMID:30713548

44. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, Hinds DA. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat Commun. 2017; 8:599. https://doi.org/10.1038/s41467-017-00257-5 PMID:28928442

45. Zhu Z, Lee PH, Chaffin MD, Chung W, Loh PR, Lu Q, Christiani DC, Liang L. A genome-wide cross-trait analysis from UK biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat Genet. 2018; 50:857–64. https://doi.org/10.1038/s41588-018-0121-0 PMID:29785011

46. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, Byun J, Dunning A, Pooley KA, et al, and SpiroMeta Consortium. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017; 49:1126–32. https://doi.org/10.1038/ng.3892 PMID:28604730

47. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batini C, Fawcett KA, Song K, Sakornsakolpat P, Li X, Boxall R, Reeve NF, et al, and Understanding Society Scientific Group. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat Genet. 2019; 51:481–93. https://doi.org/10.1038/s41588-018-0321-7 PMID:30804560

48. Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, Obeidat M, Henry AP, Portelli MA, Hall RJ, Billington CK, Rimington TL, Fenech AG, et al, and Understanding Society Scientific Group, and Geisinger-Regeneron DiscovEHR Collaboration. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nat Genet. 2017; 49:416–25. https://doi.org/10.1038/ng.3787 PMID:28166213

49. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK biobank. Nat Genet. 2018; 50:1593–99. https://doi.org/10.1038/s41588-018-0248-z PMID:30349118

50. Yang CP, Li X, Wu Y, Shen Q, Zeng Y, Xiong Q, Wei M, Chen C, Liu J, Huo Y, Li K, Xue G, Yao YG, et al. Comprehensive integrative analyses identify GLT8D1 and CSNK2B as schizophrenia risk genes. Nat Commun. 2018; 9:838.
https://doi.org/10.1038/s41467-018-03247-3
PMID:29483533

51. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, Zeng L, Ntalla I, Lai FY, Hopewell JC, Giannakopoulou O, Jiang T, Hamby SE, et al, and EPIC-CVD Consortium, CARDIoGRAMplusC4D, and UK Biobank CardioMetabolic Consortium CHD working group. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat Genet. 2017; 49:1385–91.
https://doi.org/10.1038/ng.3913
PMID:28714975

52. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, et al, and MAGIC investigators, and GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010; 42:579–89.
https://doi.org/10.1038/ng.609
PMID:20581827

53. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, Webb TR, Zeng L, Dehghan A, et al. A comprehensive 1,000 genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015; 47:1121–30.
https://doi.org/10.1038/ng.3396
PMID:26343387

54. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, Dong R, Guarani V, Vaites LP, et al. The BioPlex network: a systematic exploration of the human interactome. Cell. 2015; 162:425–40.
https://doi.org/10.1016/j.cell.2015.06.043
PMID:26186194

55. Kim EJ, Park MK, Kang GJ, Byun HJ, Kim HJ, Yu L, Kim B, Chae HS, Chin YW, Shim JG, Lee H, Lee CH. YDJC induces epithelial-mesenchymal transition via escaping from interaction with CDC16 through ubiquitination of PP2A. J Oncol. 2019; 2019:3542537.
https://doi.org/10.1155/2019/3542537
PMID:31485224

56. Kim EJ, Park MK, Byun HJ, Kang GJ, Yu L, Kim HJ, Shim JG, Lee H, Lee CH. YdjC chitooligosaccharide deacetylase homolog induces keratin reorganization in lung cancer cells: involvement of interaction between YDJC and CDC16. Oncotarget. 2018; 9:22915–28.
https://doi.org/10.18632/oncotarget.25145
PMID:29796162

57. Sisti J, Boffetta P. What proportion of lung cancer in never-smokers can be attributed to known risk factors? Int J Cancer. 2012; 131:265–75.
https://doi.org/10.1002/ijc.27477 PMID:22322343

58. Wong JY, Zhang H, Hsiung CA, Shiraishi K, Yu K, Matsuo K, Wong MP, Hong YC, Wang J, Seow WJ, Wang Z, Song M, Kim HN, et al. Tuberculosis infection and lung adenocarcinoma: mendelian randomization and pathway analysis of genome-wide association study data from never-smoking Asian women. Genomics. 2020; 112:1223–32.
https://doi.org/10.1016/j.ygeno.2019.07.008
PMID:31306748

59. Jin J, Li Y, Ren J, Man Lam S, Zhang Y, Hou Y, Zhang X, Xu R, Shui G, Ma RZ. Neonatal respiratory failure with retarded perinatal lung maturation in mice caused by reticulocalbin 3 disruption. Am J Respir Cell Mol Biol. 2016; 54:410–23.
https://doi.org/10.1165/rcmb.2015-0036OC
PMID:26252542

60. Jin J, Shi X, Li Y, Zhang Q, Guo Y, Li C, Tan P, Fang Q, Ma Y, Ma RZ. Reticulocalbin 3 deficiency in alveolar epithelium exacerbated bleomycin-induced pulmonary fibrosis. Am J Respir Cell Mol Biol. 2018; 59:320–33.
https://doi.org/10.1165/rcmb.2017-0347OC
PMID:29676583

61. Hou Y, Li Y, Gong F, Jin J, Huang A, Fang Q, Ma RZ. A preliminary study on RCN3 protein expression in non-small cell lung cancer. Clin Lab. 2016; 62:293–300.
https://doi.org/10.7754/clin.lab.2015.150411
PMID:27156316

62. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, et al, and UK10K Consortium. The UK10K project identifies rare variants in health and disease. Nature. 2015; 526:82–90.
https://doi.org/10.1038/nature14962
PMID:26367797

63. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA, and 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65.
https://doi.org/10.1038/nature11632 PMID:23128226

64. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, et al, and Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48:1279–83.

https://doi.org/10.1038/ng.3643
PMID:27548312

65. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, Consonni D, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. 2009; 85:679–91.
https://doi.org/10.1016/j.ajhg.2009.09.012
PMID:19836008

66. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, Eleftheriadis M, Sinning CR, Schnabel RB, et al. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS One. 2010; 5:e10693.
https://doi.org/10.1371/journal.pone.0010693
PMID:20502693

67. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet. 2013; 45:580–85.
https://doi.org/10.1038/ng.2653
PMID:23715323

68. GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–60.
https://doi.org/10.1126/science.1262110
PMID:25954001

69. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. Nat Genet. 2017; 49:1664–70.
https://doi.org/10.1038/ng.3969
PMID:29019975

70. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015; 11:e1004219.
https://doi.org/10.1371/journal.pcbi.1004219
PMID:25885710

71. Akula N, Wendland JR, Choi KH, McMahon FJ. An integrative genomic study implicates the postsynaptic density in the pathogenesis of bipolar disorder. Neuropsychopharmacology. 2016; 41:886–95.
https://doi.org/10.1038/npp.2015.218
PMID:26211730

72. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011; 39:W316–22.
https://doi.org/10.1093/nar/gkr483
PMID:21715386

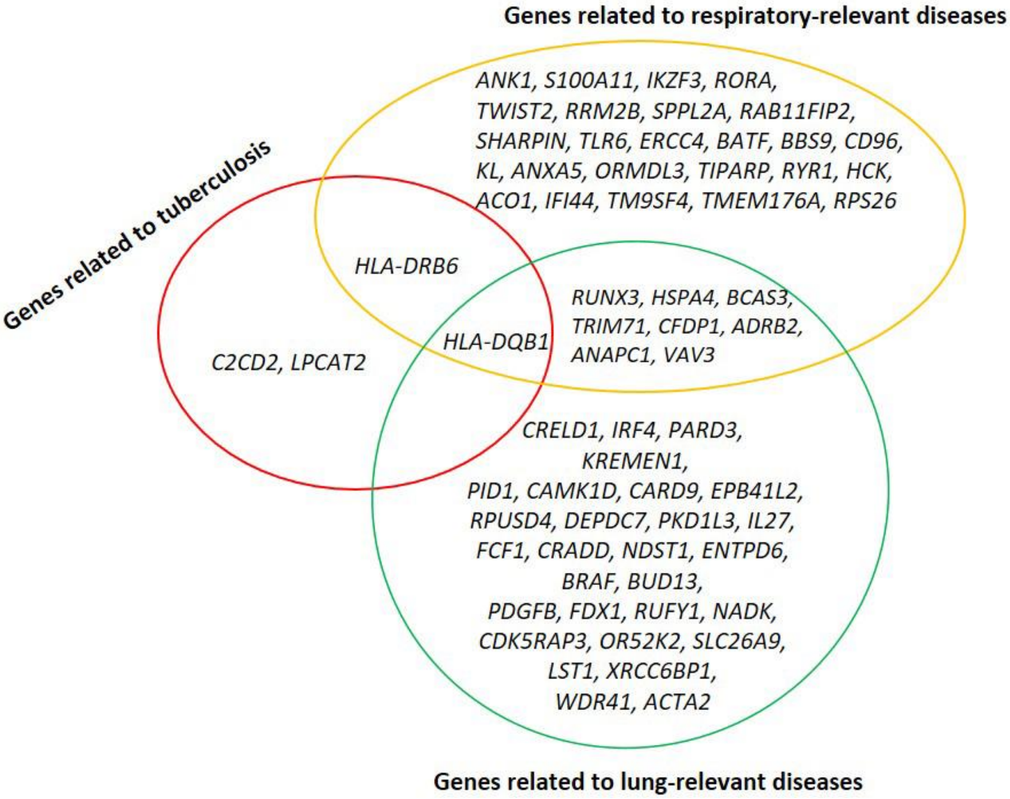
73. Ai C, Kong L. CGPS: a machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. J Genet Genomics. 2018; 45:489–504.
https://doi.org/10.1016/j.jgg.2018.08.002
PMID:30292791

74. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010; 38:W214–20.
https://doi.org/10.1093/nar/gkq537
PMID:20576703

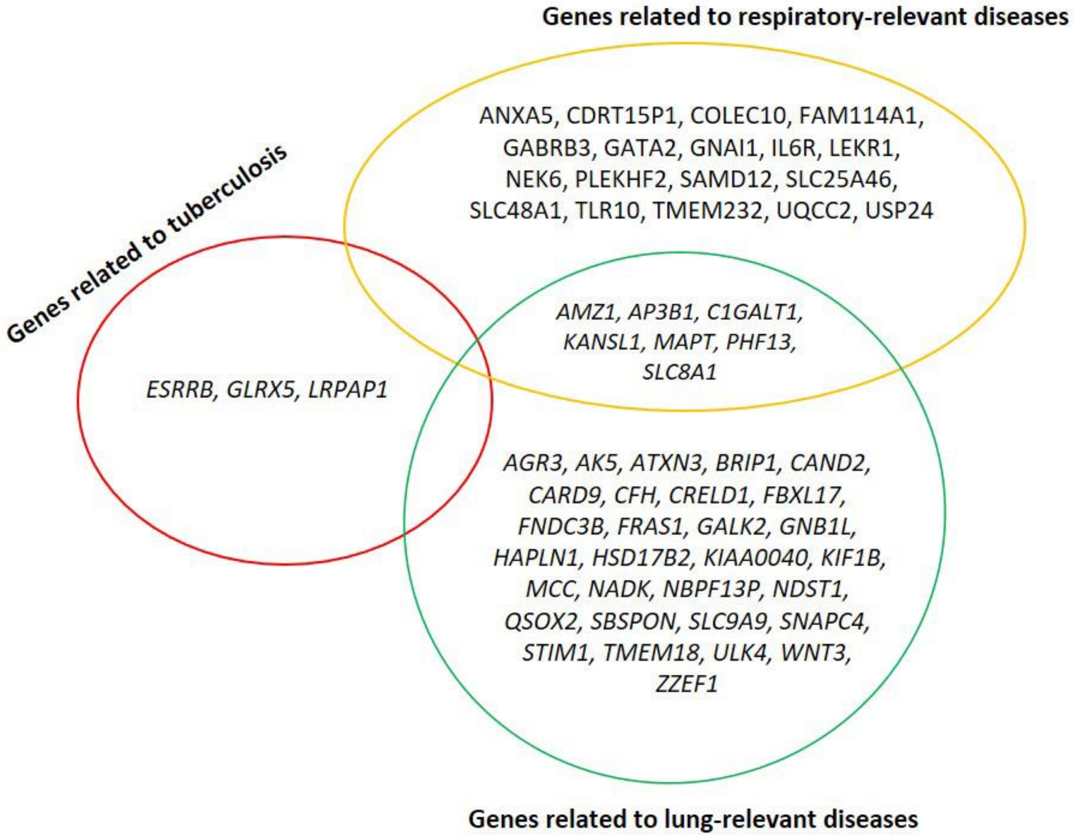
75. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–504.
https://doi.org/10.1101/gr.1239303
PMID:14597658

SUPPLEMENTARY MATERIALS

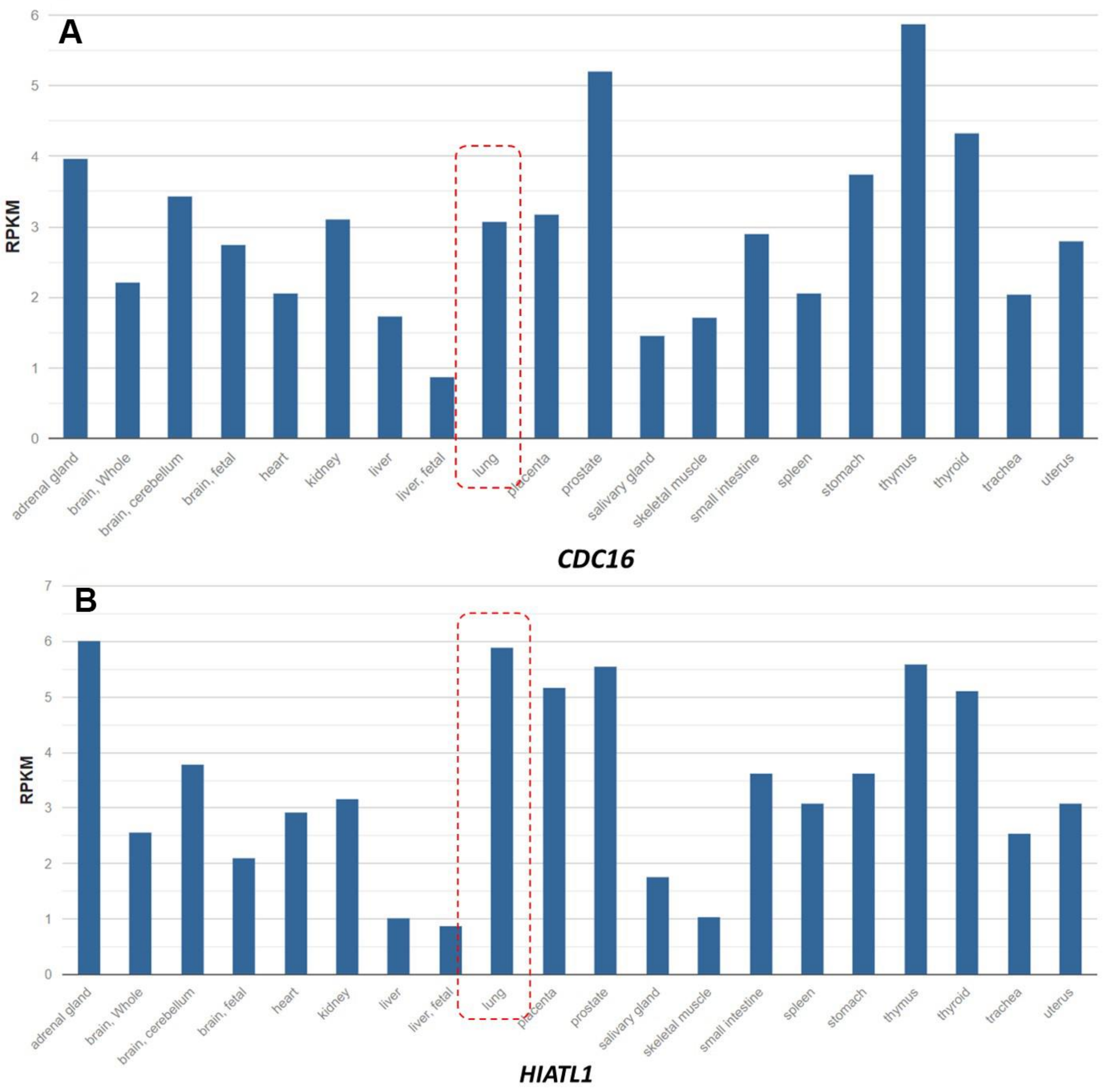
Supplementary Figures


Genes related to respiratory-relevant diseases
Genes related to tuberculosis
Genes related to lung-relevant diseases
ANK1, S100A11, IKZF3, RORA, TWIST2, RRM2B, SPPL2A, RAB11FIP2, SHARPIN, TLR6, ERCC4, BATF, BBS9, CD96, KL, ANXA5, ORMDL3, TIPARP, RYR1, HCK, ACO1, IFI44, TM9SF4, TMEM176A, RPS26
HLA-DRB6
RUNX3, HSPA4, BCAS3, TRIM71, CFDP1, ADRB2, ANAPC1, VAV3
HLA-DQB1
C2CD2, LPCAT2
CRELD1, IRF4, PARD3, KREMEN1, PID1, CAMK1D, CARD9, EPB41L2, RPUSD4, DEPDC7, PKD1L3, IL27, FCF1, CRADD, NDST1, ENTPD6, BRAF, BUD13, PDGFB, FDX1, RUFY1, NADK, CDK5RAP3, OR52K2, SLC26A9, LST1, XRCC6BP1, WDR41, ACTA2

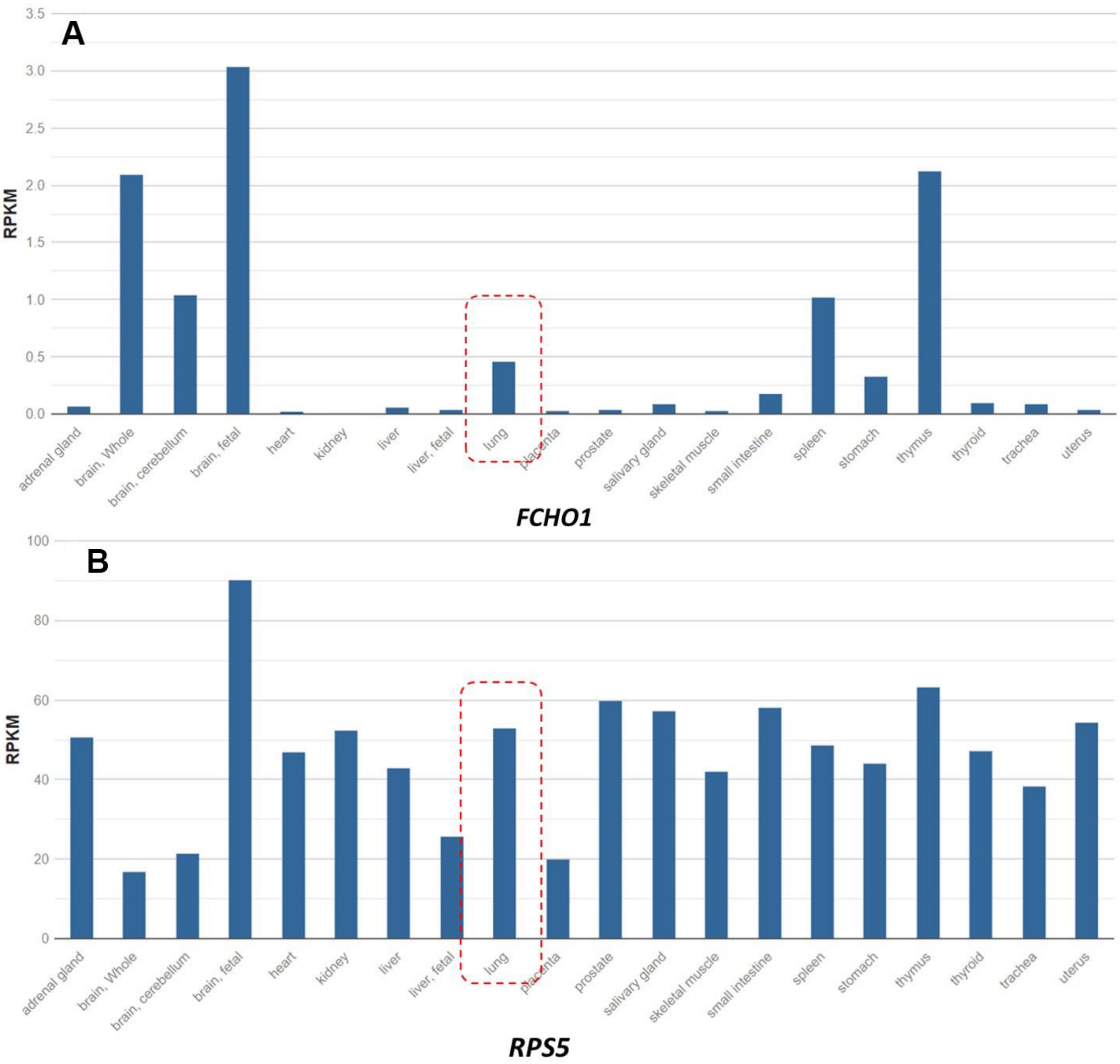
Supplementary Figure 1. Previous studies provide supportive evidence of these Sherlock-identified genes in the discovery stage.


Genes related to respiratory-relevant diseases
ANXA5, CDRT15P1, COLEC10, FAM114A1, GABRB3, GATA2, GNAI1, IL6R, LEKR1, NEK6, PLEKHF2, SAMD12, SLC25A46, SLC48A1, TLR10, TMEM232, UQCC2, USP24
Genes related to tuberculosis
AMZ1, AP3B1, C1GALT1, KANSL1, MAPT, PHF13, SLC8A1
ESRRB, GLRX5, LRPAP1
AGR3, AK5, ATXN3, BRIP1, CAND2, CARD9, CFH, CRELD1, FBXL17, FNDC3B, FRAS1, GALK2, GNB1L, HAPLN1, HSD17B2, KIAA0040, KIF1B, MCC, NADK, NBPF13P, NDST1, QSOX2, SBSPON, SLC9A9, SNAPC4, STIM1, TMEM18, ULK4, WNT3, ZZEF1
Genes related to lung-relevant diseases

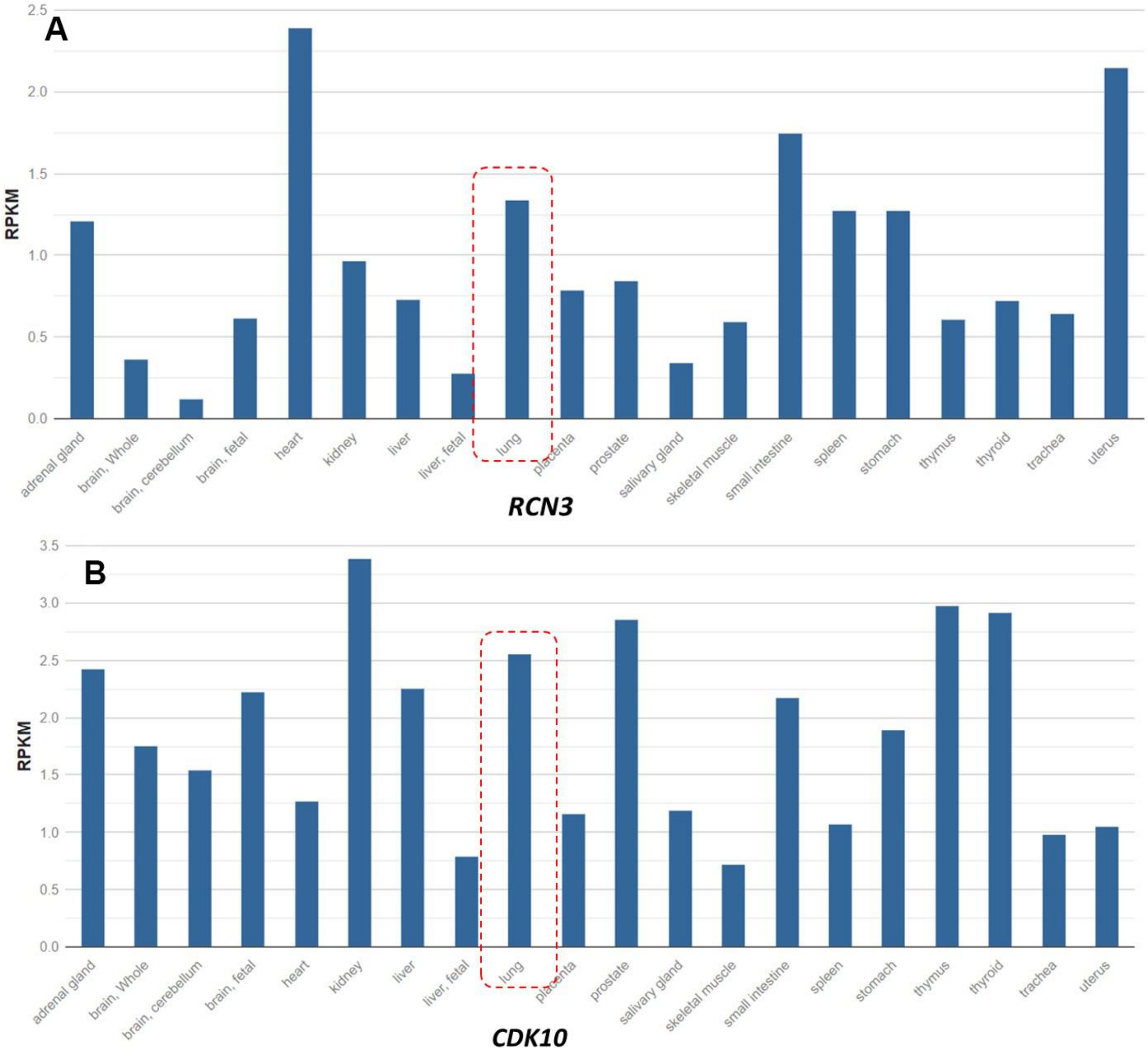
Supplementary Figure 2. Previous studies provide supportive evidence of these Sherlock-identified genes in the replication stage (based on both Dataset #4 and #5).



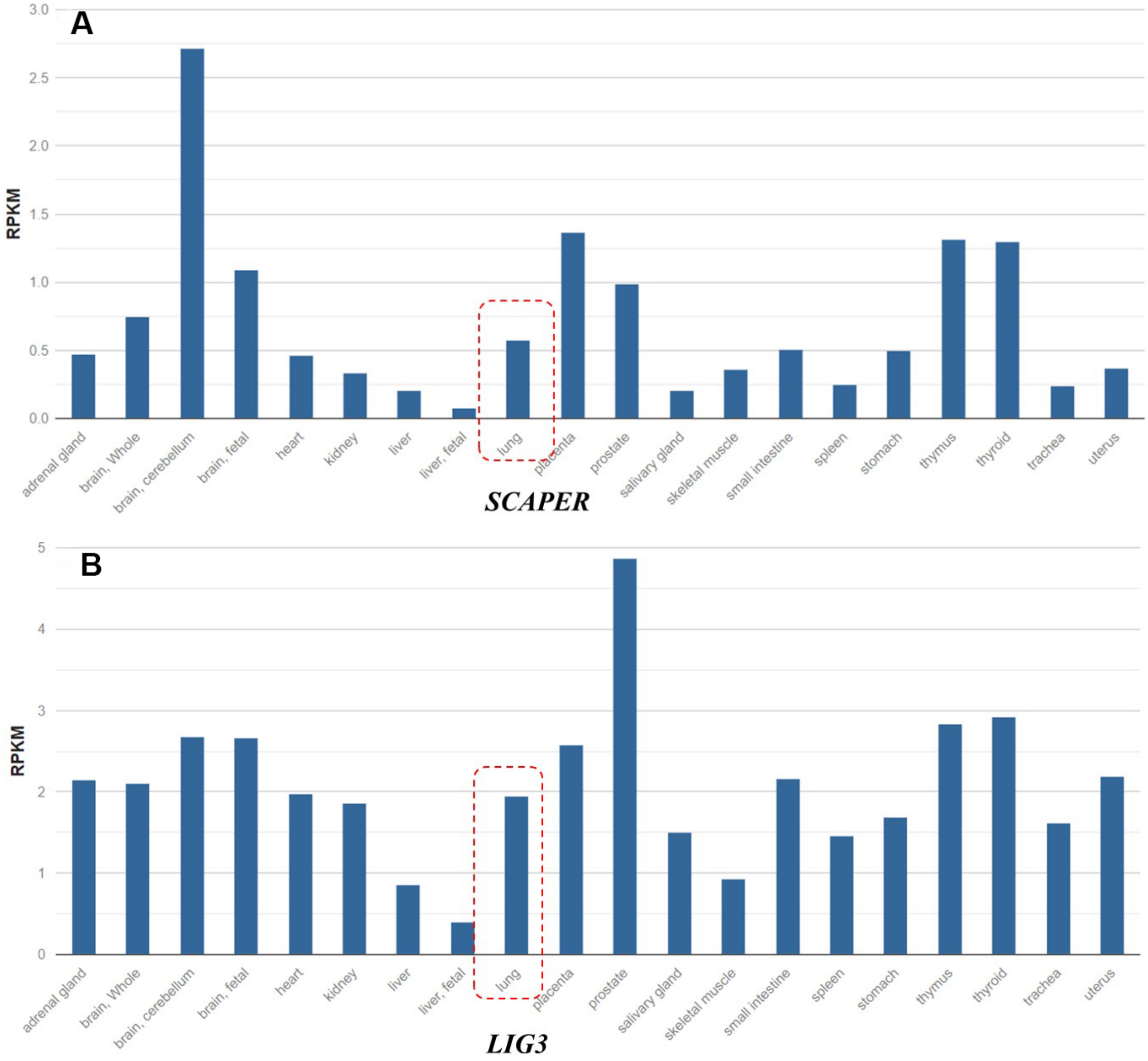
Supplementary Figure 3. Expression abundance of *CDC16* and *HIATL1* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



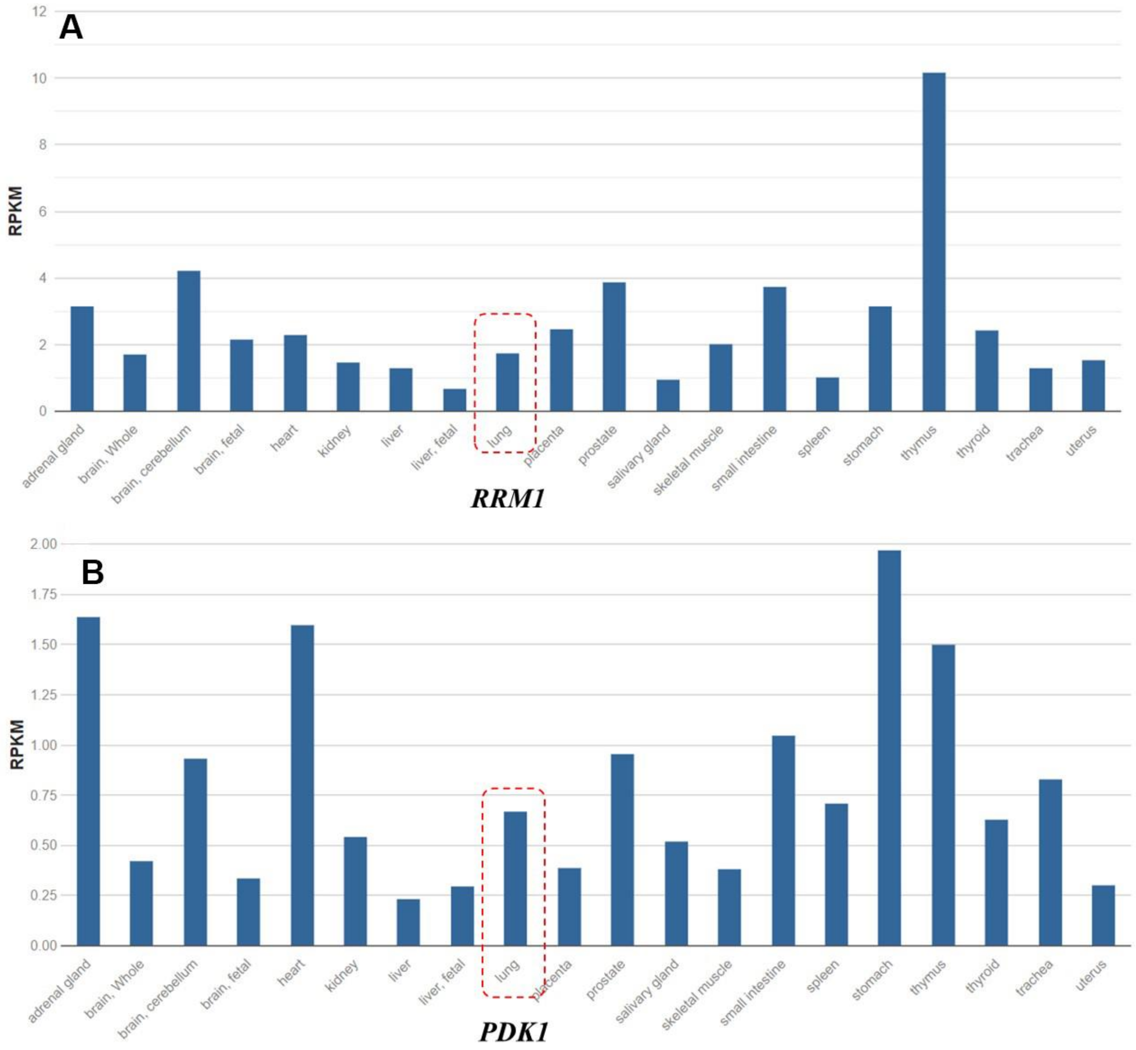
Supplementary Figure 4. Expression abundance of *FCHO1* and *RPS5* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



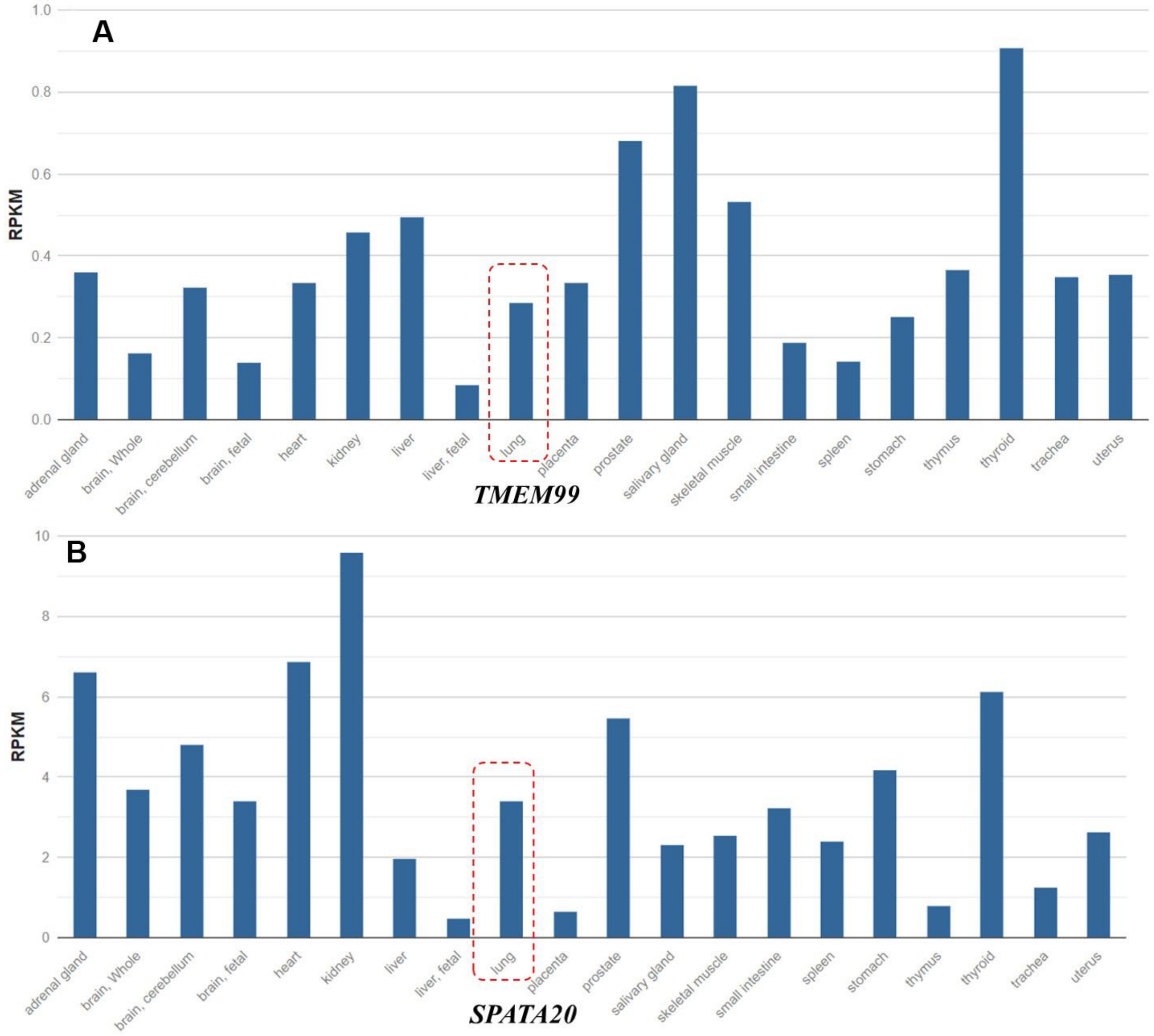
Supplementary Figure 5. Expression abundance of *RCN3* and *CDK10* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



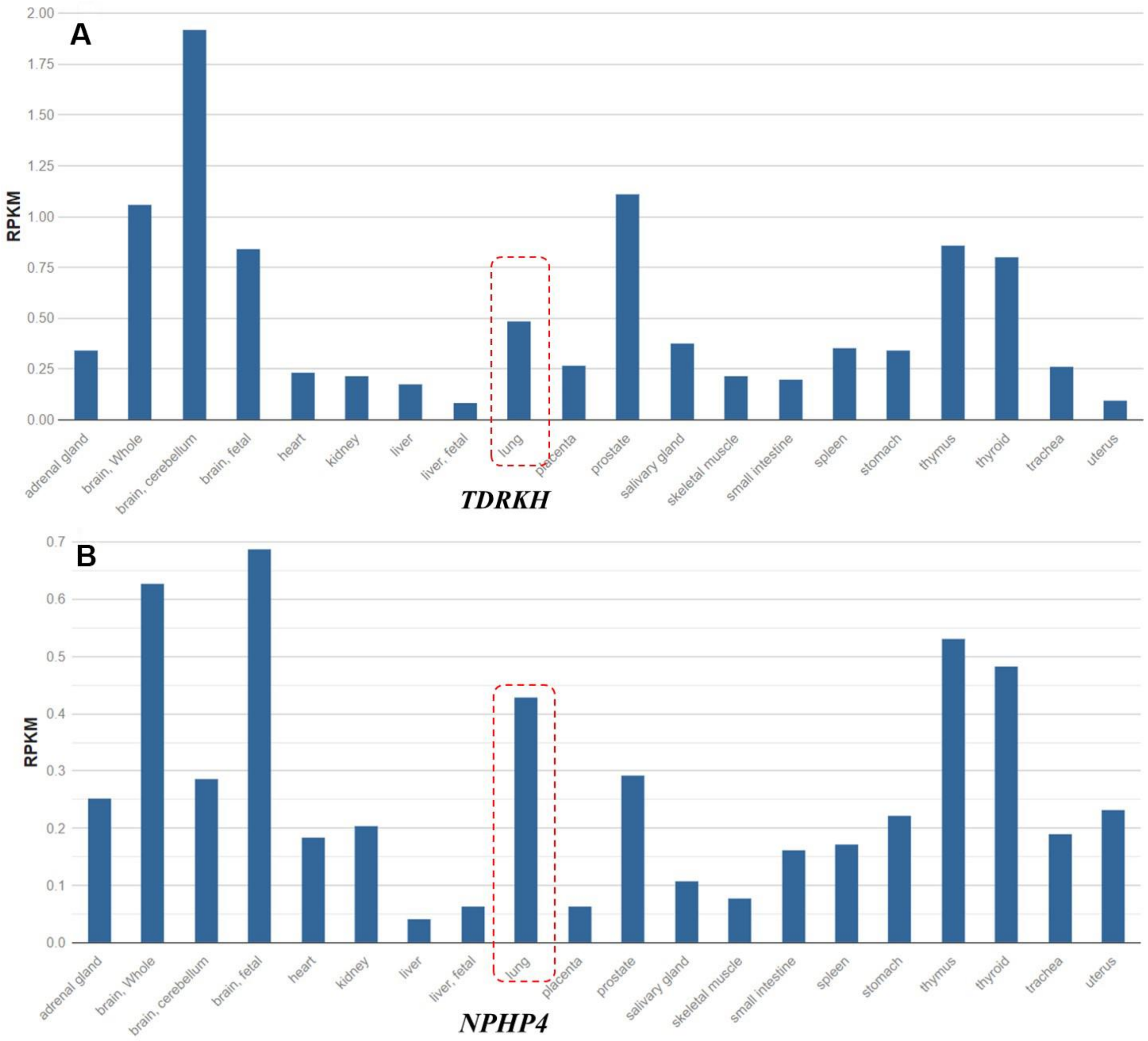
Supplementary Figure 6. Expression abundance of *SCAPER* and *LIG3* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



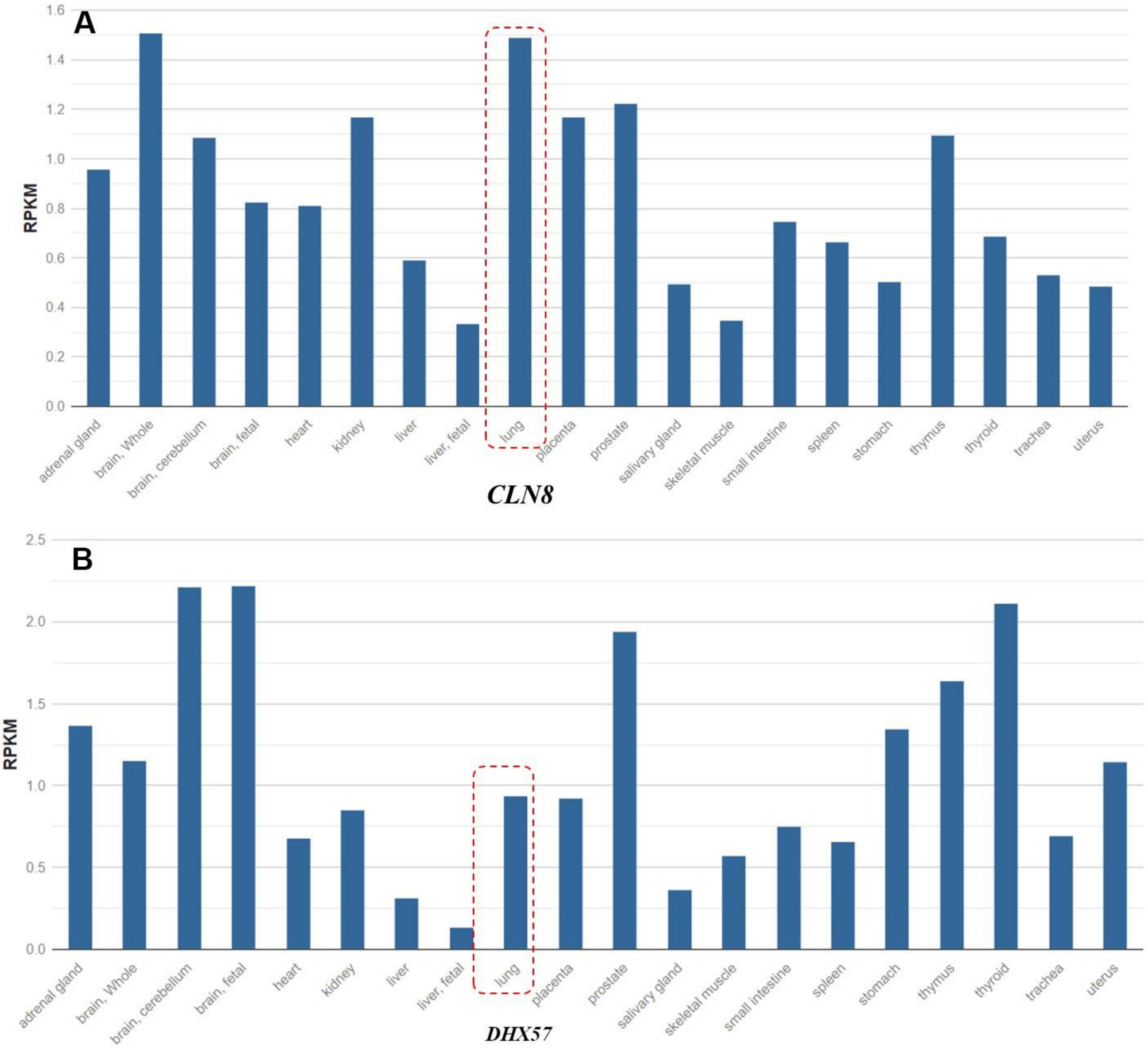
Supplementary Figure 7. Expression abundance of *RRM1* and *PDK1* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



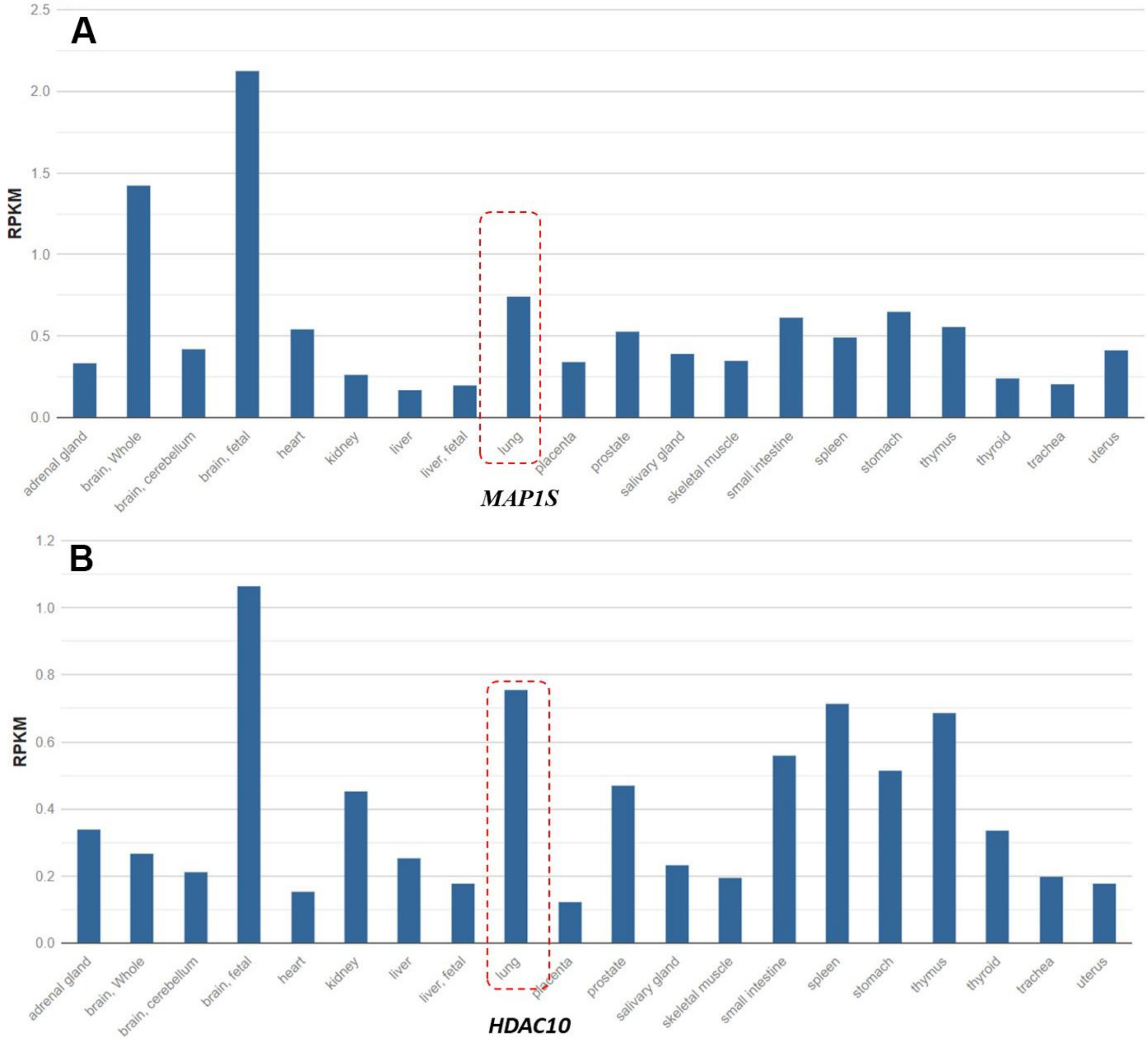
Supplementary Figure 8. Expression abundance of *TMEM99* and *SPATA20* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



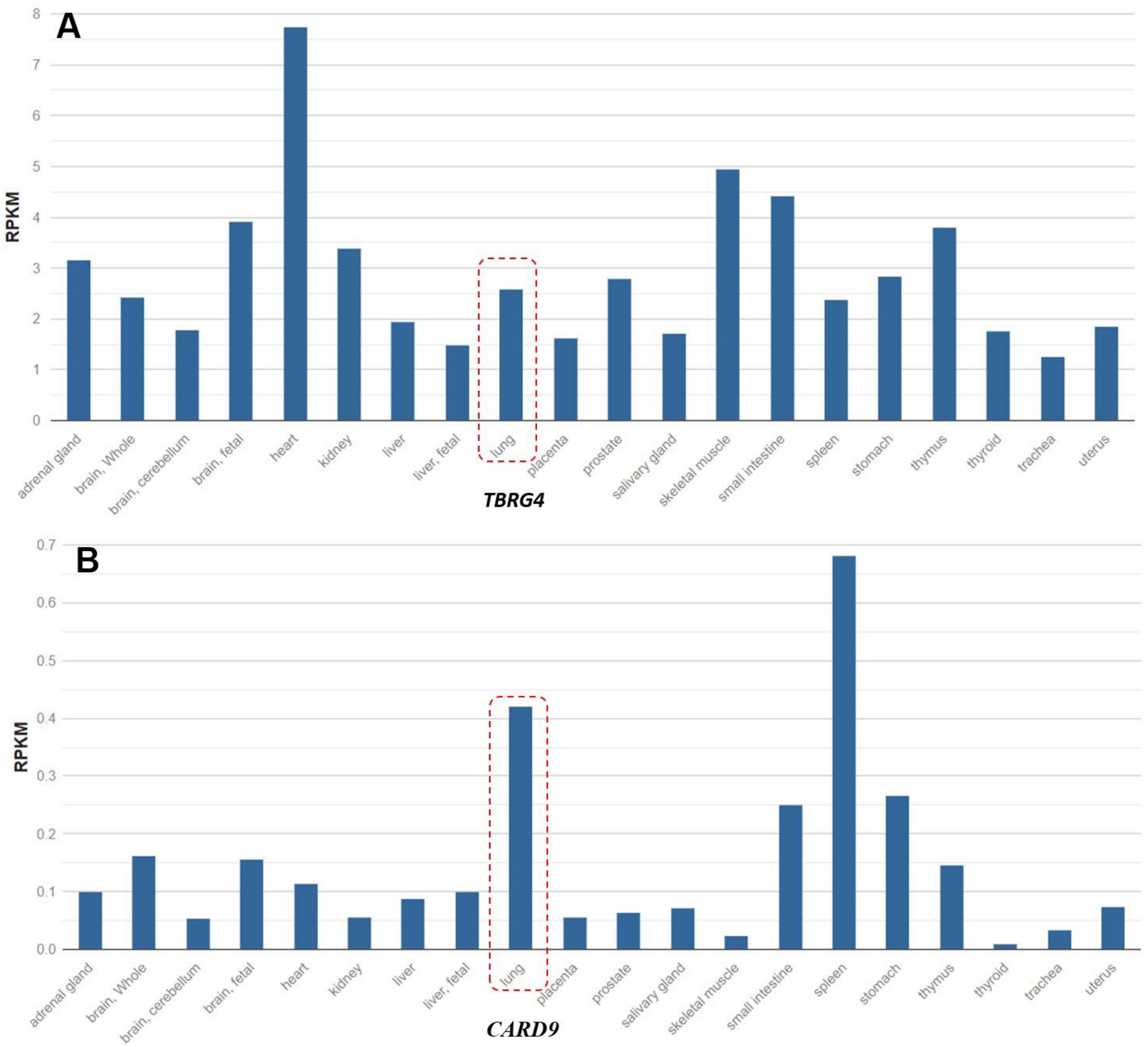
Supplementary Figure 9. Expression abundance of *TDRKH* and *NPHP4* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



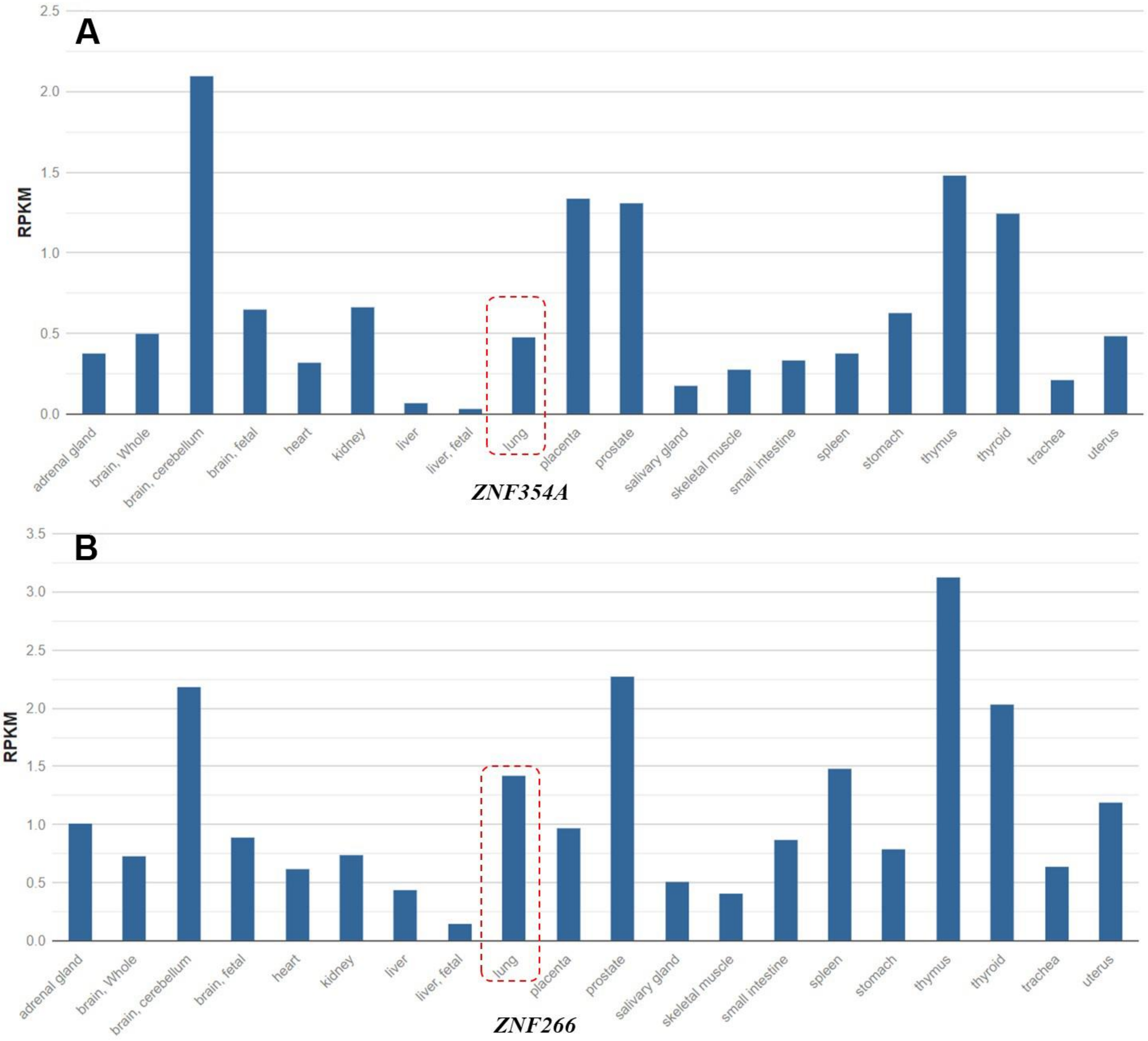
Supplementary Figure 10. Expression abundance of *CLN8* and *DHX57* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



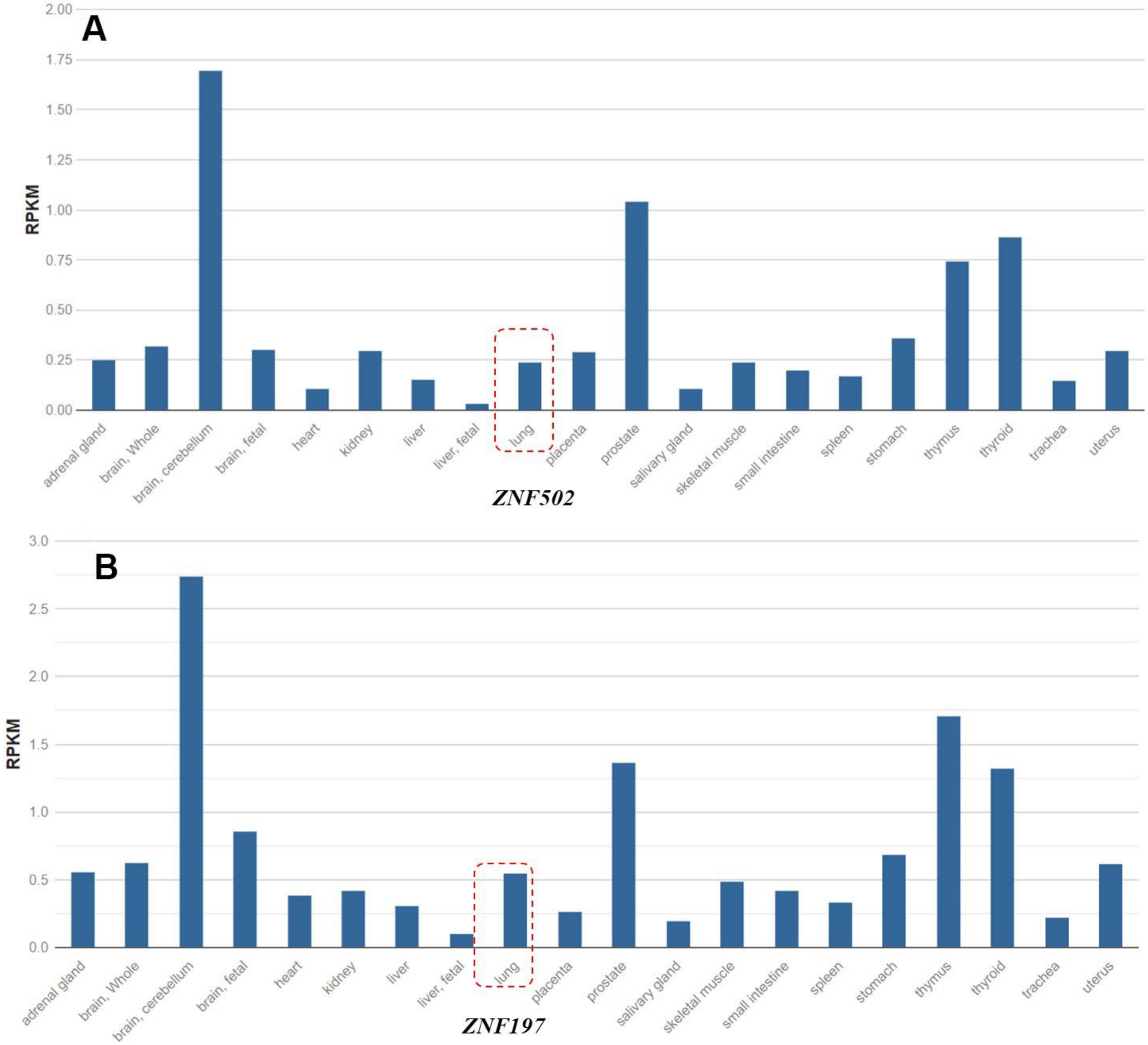
Supplementary Figure 11. Expression abundance of *MAP1S* and *HDAC10* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



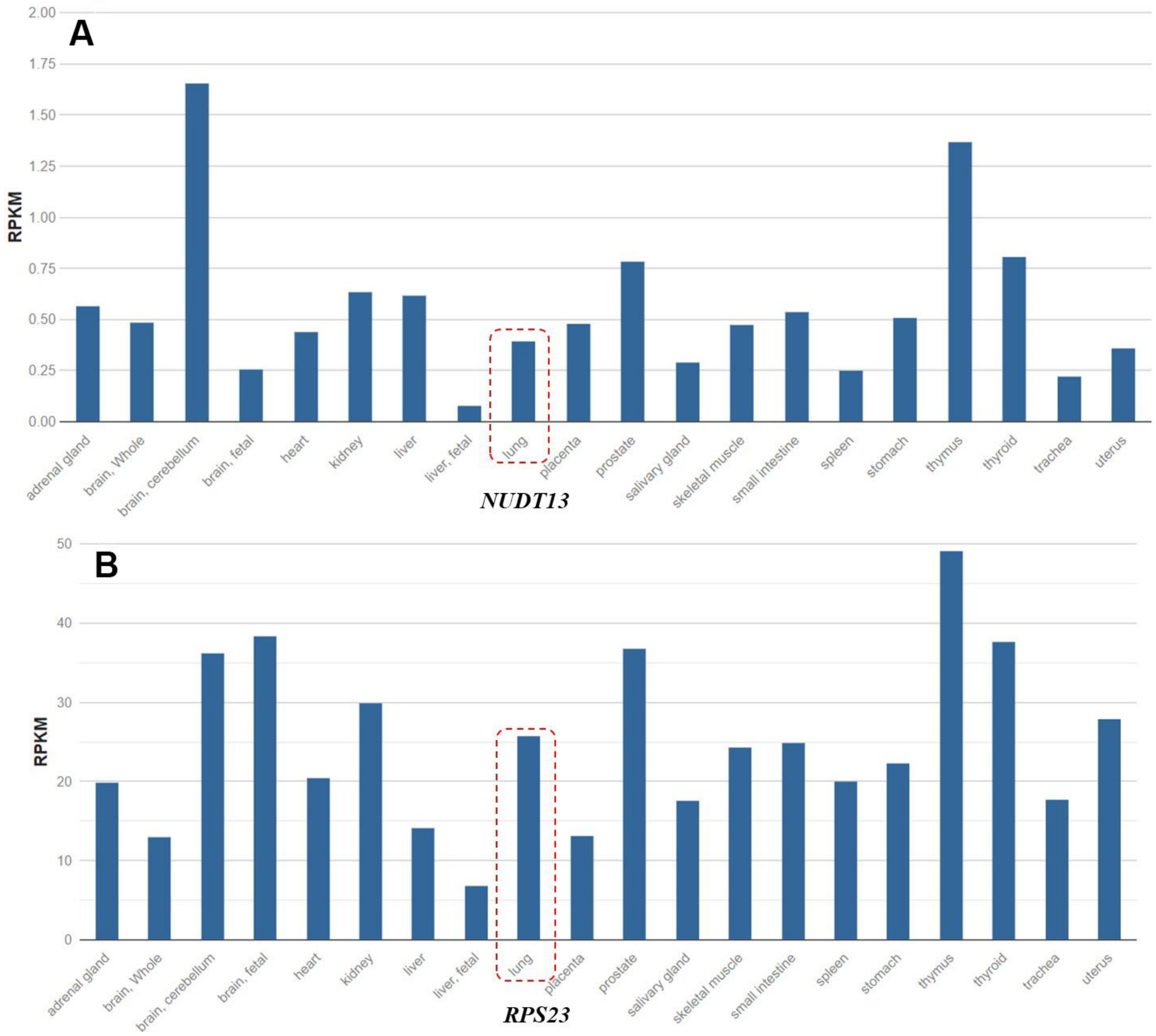
Supplementary Figure 12. Expression abundance of *TBRG4* and *CARD9* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



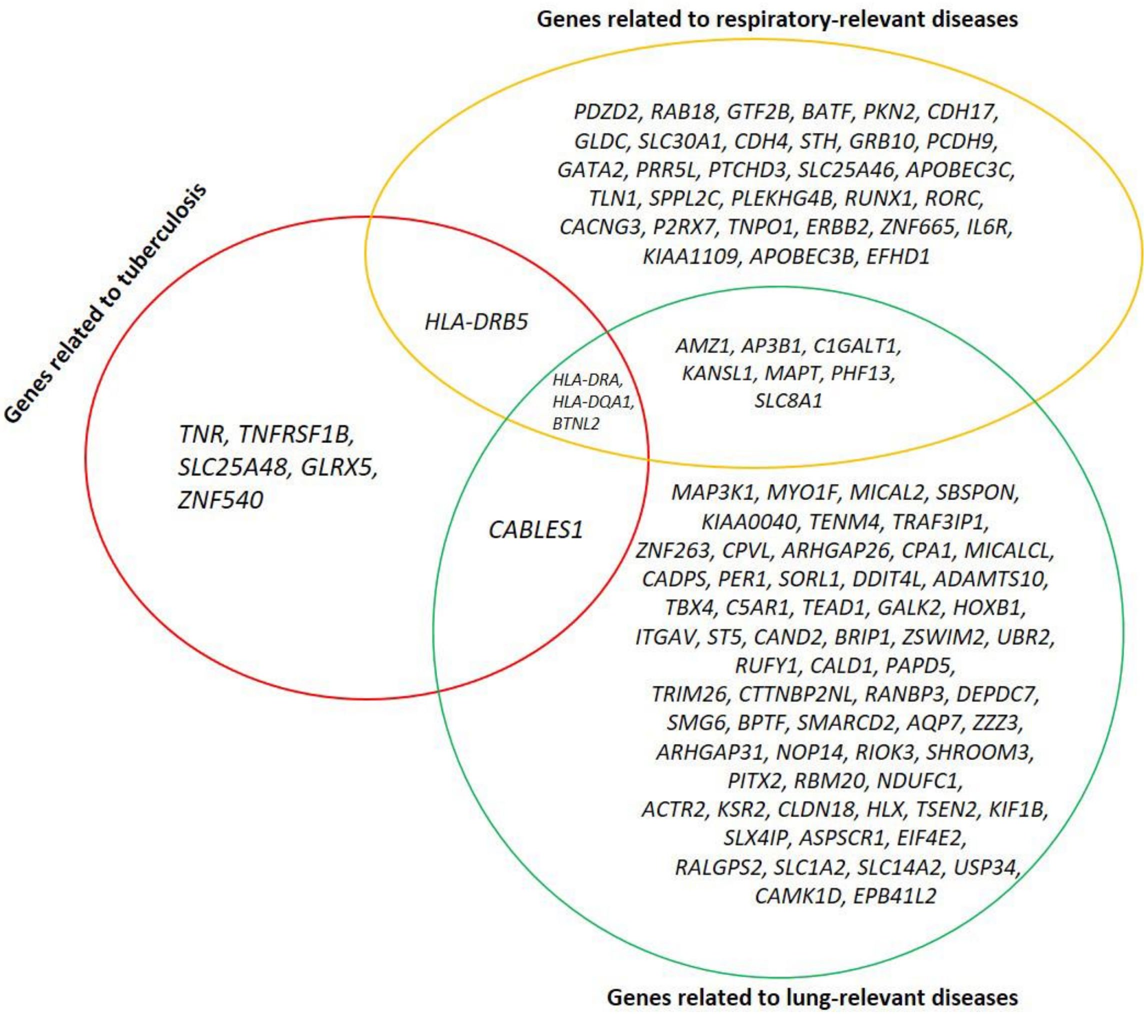
Supplementary Figure 13. Expression abundance of *ZNF354A* and *ZNF266* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



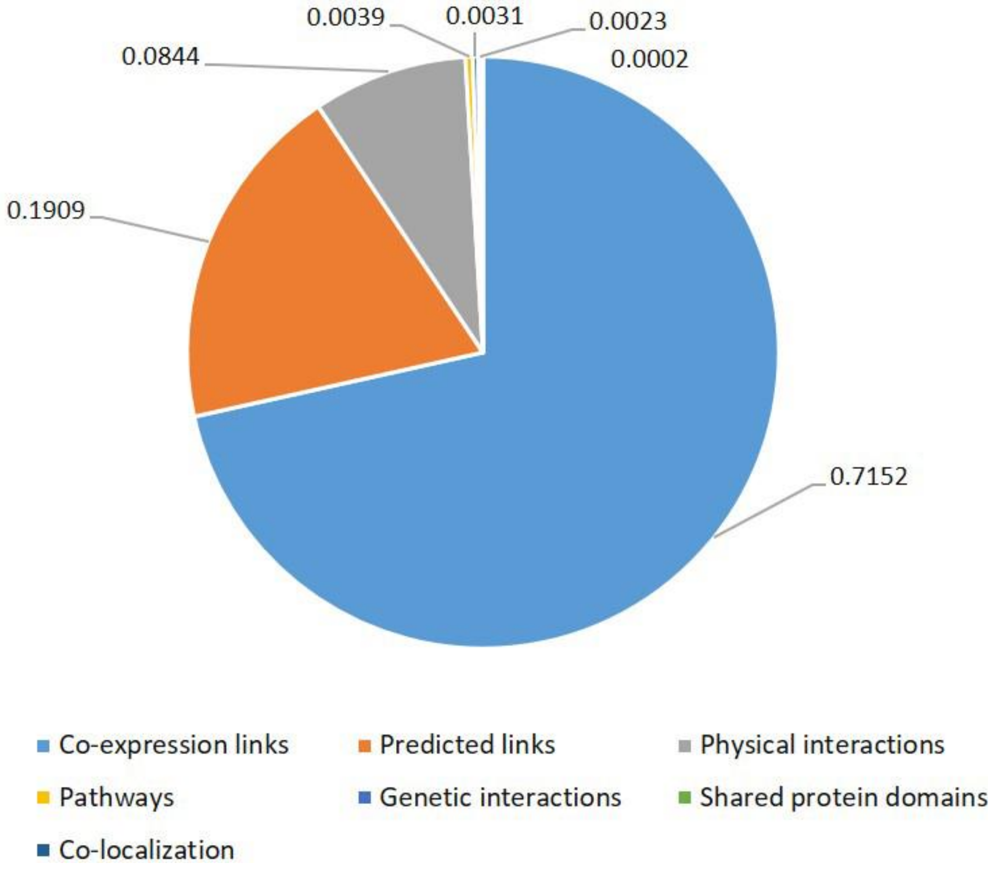
Supplementary Figure 14. Expression abundance of *ZNF502* and *ZNF197* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).



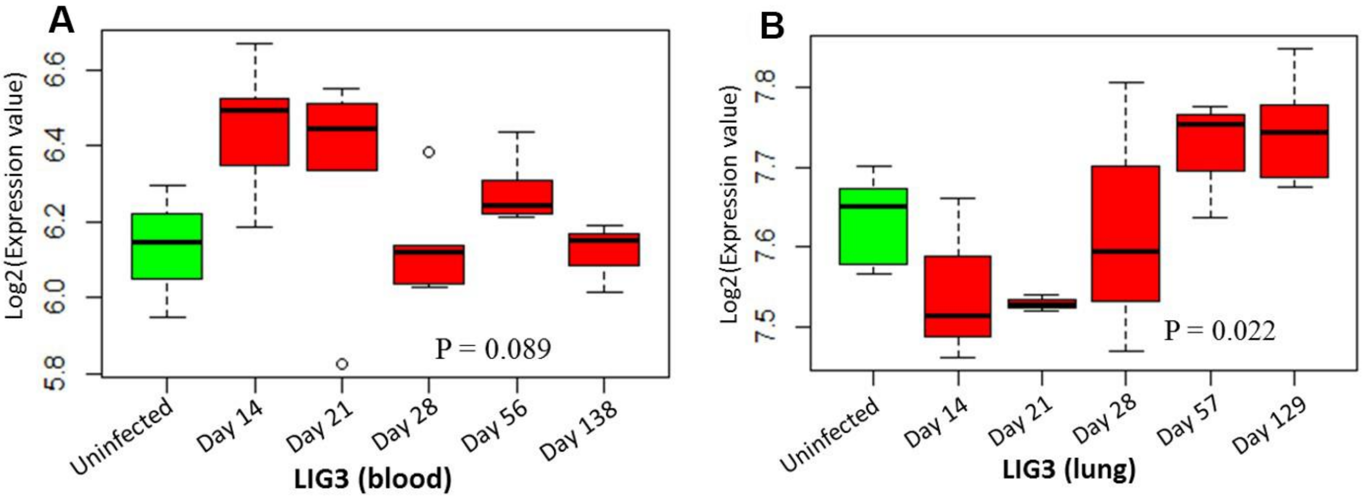
Supplementary Figure 15. Expression abundance of *NUDT13* and *RPS23* based on RNA sequencing from 20 human tissues. The expression data were obtained from BioProject (Accession No. PRJNA280600). Expression values are shown in Reads Per Kilobase per Million mapped reads (RPKM).


Genes related to respiratory-relevant diseases
PDZD2, RAB18, GTF2B, BATF, PKN2, CDH17,
GLDC, SLC30A1, CDH4, STH, GRB10, PCDH9,
GATA2, PRR5L, PTCHD3, SLC25A46, APOBEC3C,
TLN1, SPPL2C, PLEKHG4B, RUNX1, RORC,
CACNG3, P2RX7, TNPO1, ERBB2, ZNF665, IL6R,
KIAA1109, APOBEC3B, EFHD1
Genes related to tuberculosis
HLA-DRB5
AMZ1, AP3B1, C1GALT1,
KANSL1, MAPT, PHF13,
SLC8A1
HLA-DRA,
HLA-DQA1,
BTNL2
TNR, TNFRSF1B,
SLC25A48, GLRX5,
ZNF540
CABLES1
MAP3K1, MYO1F, MICAL2, SBSPON,
KIAA0040, TENM4, TRAF3IP1,
ZNF263, CPVL, ARHGAP26, CPA1, MICALCL,
CADPS, PER1, SORL1, DDIT4L, ADAMTS10,
TBX4, C5AR1, TEAD1, GALK2, HOXB1,
ITGAV, ST5, CAND2, BRIP1, ZSWIM2, UBR2,
RUFY1, CALD1, PAPD5,
TRIM26, CTTNBP2NL, RANBP3, DEPDC7,
SMG6, BPTF, SMARCD2, AQP7, ZZZ3,
ARHGAP31, NOP14, RIOK3, SHROOM3,
PITX2, RBM20, NDUFC1,
ACTR2, KSR2, CLDN18, HLX, TSEN2, KIF1B,
SLX4IP, ASPSCR1, EIF4E2,
RALGPS2, SLC1A2, SLC14A2, USP34,
CAMK1D, EPB41L2
Genes related to lung-relevant diseases

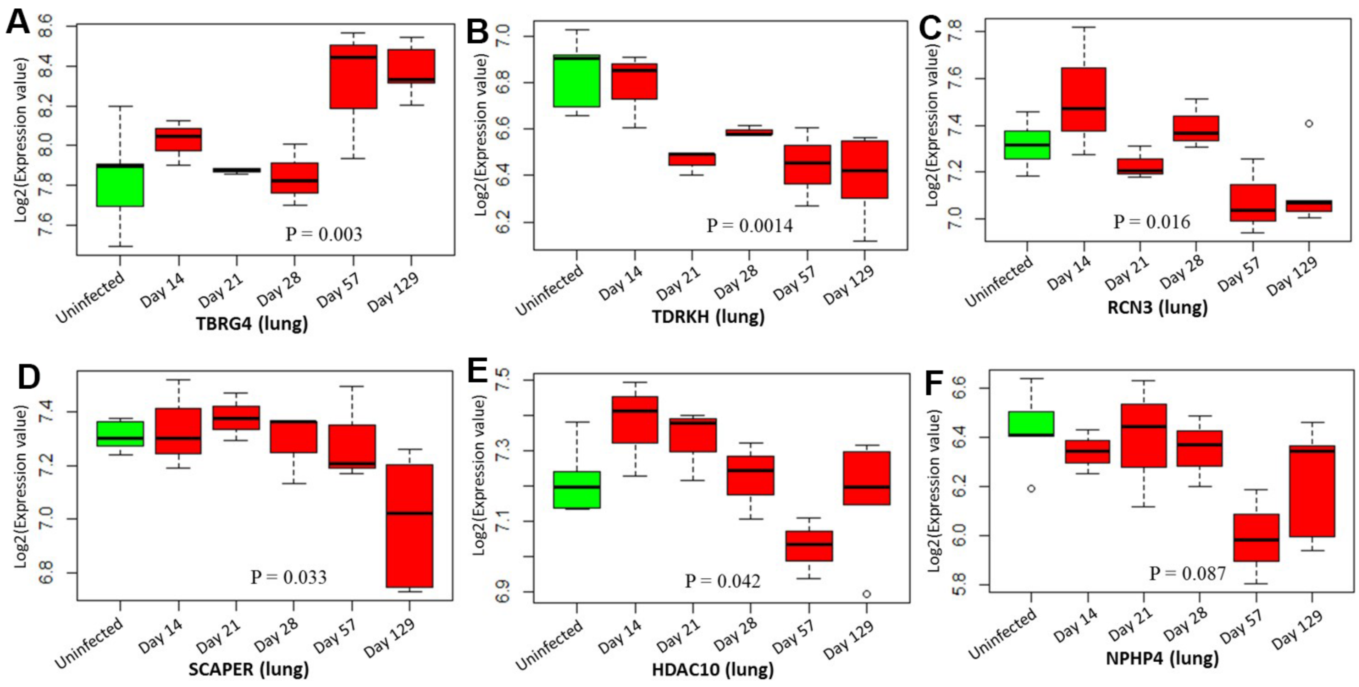
Supplementary Figure 16. Previous studies provides supportive evidence of these MAGMA-identified genes in the replication stage (based on Dataset #1).



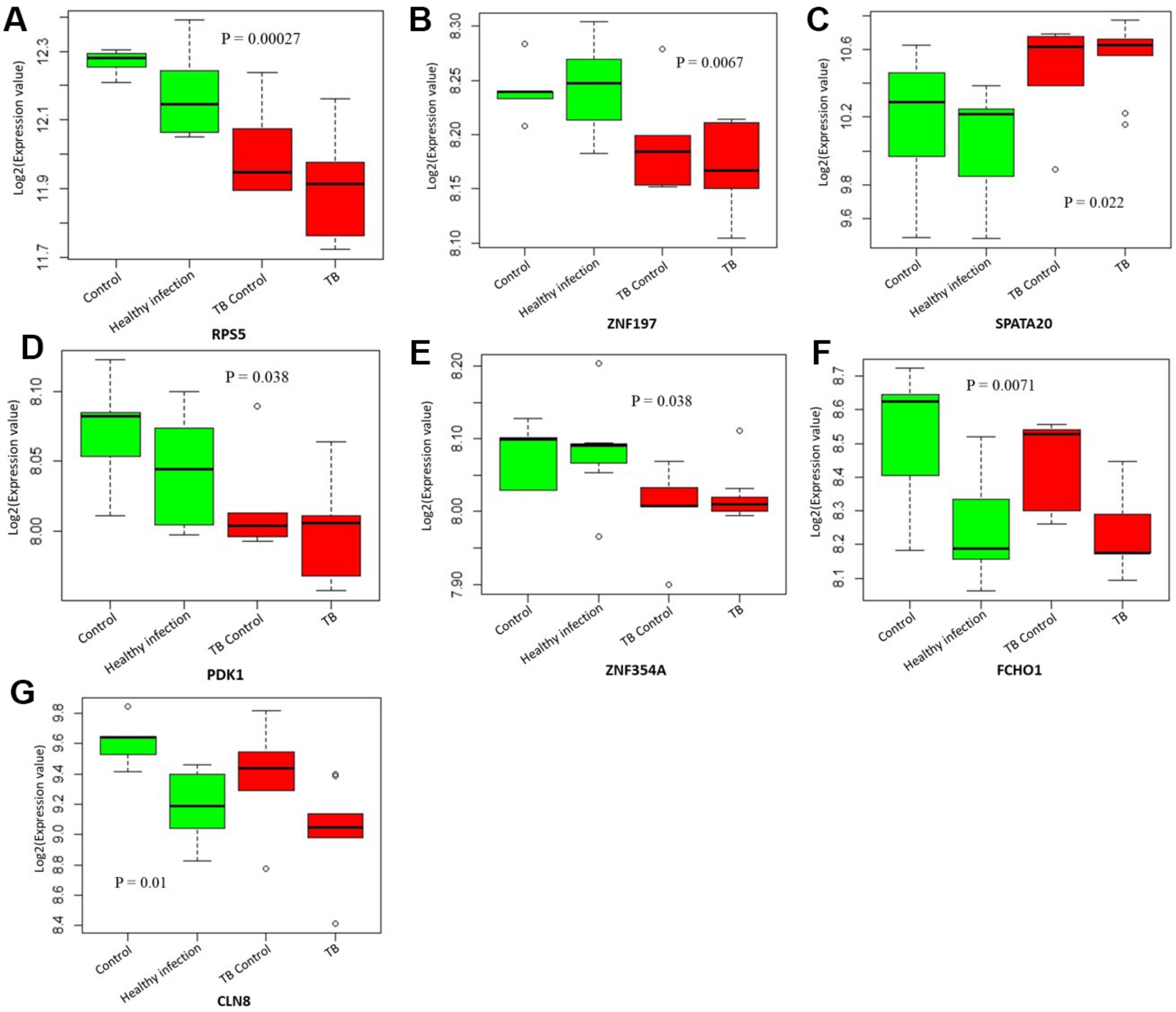
Supplementary Figure 17. The proportion of multiple layers of evidence in constructed GGI network using the GeneMANIA tool.



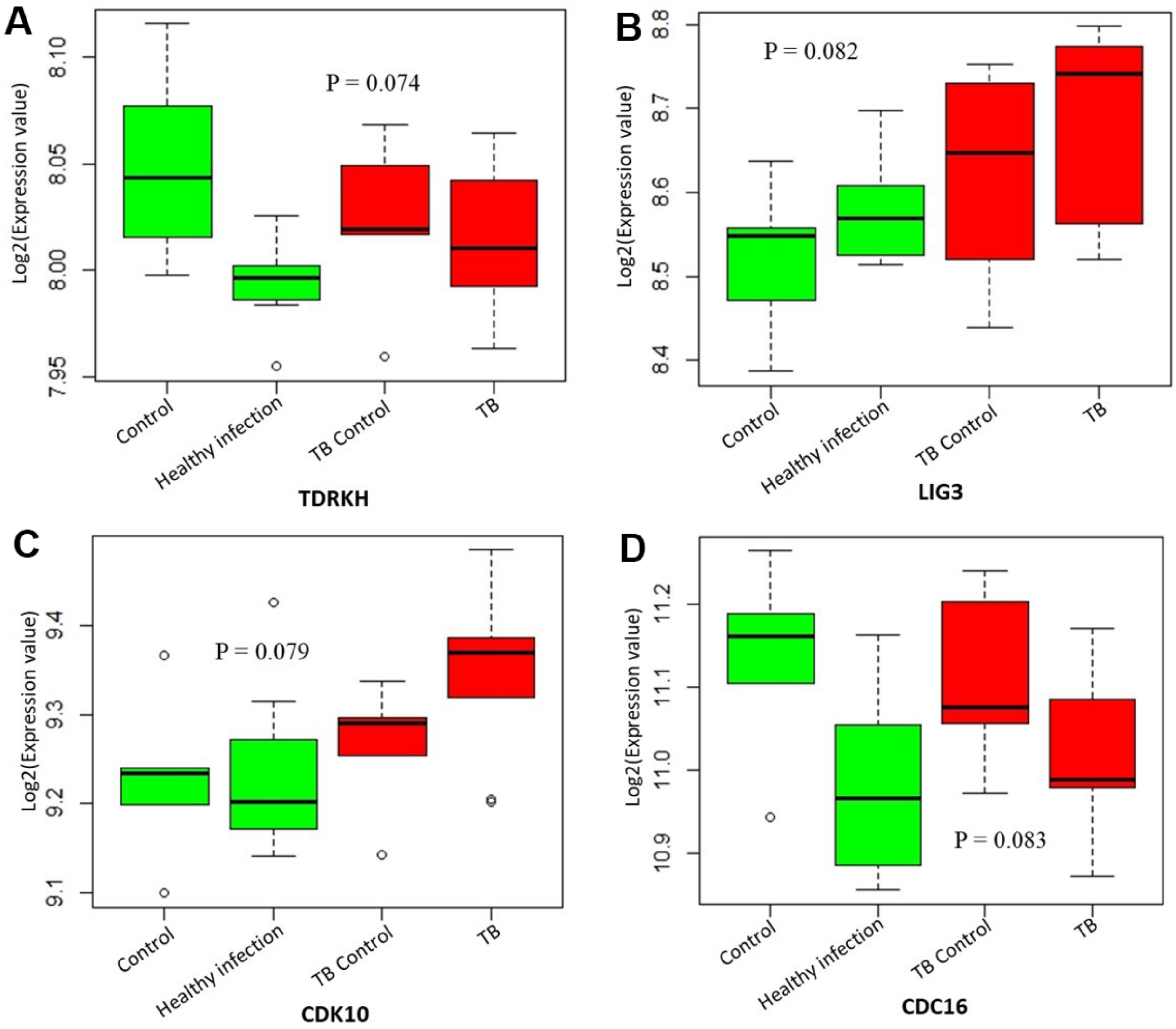
Supplementary Figure 18. Boxplots show the differential expression levels of tuberculosis-genes between uninfected mice and infected mice with 5 distinct time points based on two GSE1440943 (blood) and GSE1440944 (lung) datasets. (**A**) *LIG3* for blood; (**B**) *LIG3* for lung. P values were generated by Anova test.



Supplementary Figure 19. Boxplots show the differential expression levels of tuberculosis-genes between uninfected mice and infected mice with 5 distinct time points based on the GSE1440944 (lung) dataset. (**A**) *TBRG4* for lung; (**B**) *TDRKH* for lung; (**C**) *RCN3* for lung; (**D**) *SCAPER* for lung; (**E**) *HDAC10* for lung; (**F**) *NPHP4* for lung. P values were generated by Anova test.



Supplementary Figure 20. Boxplots show the significantly differential expression levels of tuberculosis-genes in alveolar macrophages with four groups of TB infection, TB control, healthy infection, and healthy control based on the GSE139825 dataset. (**A**) *RPS5*; (**B**) *ZNF197*; (**C**) *SPATA20*; (**D**) *PDK1*; (**E**) *ZNF354A*; (**F**) *FCHO1*; (**G**) *CLN8*. P values were generated by Anova test.



Supplementary Figure 21. Boxplots show the suggestively differential expression levels of tuberculosis-genes in alveolar macrophages with four groups of TB infection, TB control, healthy infection, and healthy control based on the GSE139825 dataset. (**A**) *TDRKH*; (**B**) *LIG3*; (**C**) *CDK10*; (**D**) *CDC16*. P values were generated by Anova test.

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 1 to 6, 14, 18, 19, 20.

Supplementary Table 7. Significant pathways enriched by tuberculosis-associated genes (Gene set #2) identified from Sherlock Bayesian analysis of dataset #4 in the replication stage.

Pathway ID	Database	Input number	Background number	P-Value	FDR
R-HSA-1430728	Reactome	28	2075	4.09E-07	1.19E-04
R-HSA-168256	Reactome	28	2096	4.98E-07	1.33E-04
R-HSA-1643685	Reactome	19	1049	5.30E-07	1.33E-04
R-HSA-392499	Reactome	26	2012	2.37E-06	3.72E-04
hsa01100	KEGG PATHWAY	21	1433	3.61E-06	5.05E-04
R-HSA-162582	Reactome	30	2689	6.93E-06	8.72E-04
R-HSA-597592	Reactome	20	1412	1.01E-05	1.09E-03
R-HSA-2470946	Reactome	3	10	2.53E-05	2.39E-03
R-HSA-68884	Reactome	3	14	5.94E-05	4.87E-03
R-HSA-71387	Reactome	8	288	6.82E-05	5.47E-03
R-HSA-168249	Reactome	15	1043	1.12E-04	6.97E-03
R-HSA-211945	Reactome	5	105	1.51E-04	8.92E-03
R-HSA-199991	Reactome	11	631	1.89E-04	1.10E-02
R-HSA-8953854	Reactome	11	667	3.00E-04	1.66E-02
R-HSA-5653656	Reactome	11	669	3.07E-04	1.68E-02
R-HSA-77075	Reactome	3	27	3.39E-04	1.80E-02
R-HSA-167160	Reactome	3	27	3.39E-04	1.80E-02
R-HSA-72086	Reactome	3	29	4.12E-04	1.90E-02
R-HSA-74160	Reactome	17	1448	4.17E-04	1.90E-02
R-HSA-73857	Reactome	16	1316	4.24E-04	1.90E-02
R-HSA-167172	Reactome	4	73	4.27E-04	1.90E-02
R-HSA-6807505	Reactome	4	74	4.49E-04	1.96E-02
hsa00030	KEGG PATHWAY	3	30	4.51E-04	1.96E-02
R-HSA-397014	Reactome	6	209	4.70E-04	1.97E-02
R-HSA-382551	Reactome	11	720	5.61E-04	2.19E-02
R-HSA-446203	Reactome	7	304	5.81E-04	2.22E-02
R-HSA-5649702	Reactome	2	7	7.19E-04	2.42E-02
R-HSA-167287	Reactome	3	36	7.41E-04	2.45E-02
R-HSA-167290	Reactome	3	36	7.41E-04	2.45E-02
R-HSA-8983711	Reactome	2	9	1.09E-03	3.27E-02
R-HSA-15869	Reactome	4	95	1.10E-03	3.27E-02
R-HSA-5362517	Reactome	3	43	1.20E-03	3.55E-02
R-HSA-5685939	Reactome	2	10	1.31E-03	3.73E-02
R-HSA-212436	Reactome	14	1193	1.33E-03	3.79E-02
R-HSA-5668914	Reactome	4	103	1.47E-03	4.06E-02
R-HSA-2468052	Reactome	2	11	1.54E-03	4.24E-02
R-HSA-167152	Reactome	3	48	1.62E-03	4.37E-02
R-HSA-6785807	Reactome	4	108	1.73E-03	4.64E-02
R-HSA-6798695	Reactome	8	478	1.80E-03	4.72E-02
R-HSA-5696398	Reactome	4	111	1.91E-03	4.90E-02

Note: Proportion of risk genes: these identified risk genes (Input number) accounted for the proportion of all genes in each pathway (Background number) enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Supplementary Table 8. Significant GO-terms enriched by tuberculosis-associated genes (Gene set #2) identified from Sherlock Bayesian analysis of dataset #4 in the replication stage.

GO-terms ID	Input number	Background number	P-Value	FDR
GO:0043229	33	1863	4.63E-11	8.74E-08
GO:0005488	31	2247	5.75E-08	4.34E-05
GO:0097159	18	845	1.07E-07	5.78E-05
GO:0032555	11	292	1.69E-07	7.08E-05
GO:0009987	34	2852	3.69E-07	1.16E-04
GO:0005737	24	1641	7.34E-07	1.63E-04
GO:0043227	27	2030	8.76E-07	1.74E-04
GO:0005622	28	2228	1.61E-06	2.89E-04
GO:0043168	12	454	1.76E-06	2.89E-04
GO:0019222	18	1070	2.92E-06	4.24E-04
GO:0097708	9	261	4.54E-06	5.91E-04
GO:0005829	12	531	8.30E-06	9.79E-04
GO:0003824	18	1162	8.82E-06	9.79E-04
GO:0032553	9	292	1.08E-05	1.11E-03
GO:0031982	13	667	1.62E-05	1.57E-03
GO:0043226	24	2086	3.77E-05	3.39E-03
GO:0005515	21	1688	3.96E-05	3.39E-03
GO:0010468	15	1012	8.07E-05	6.08E-03
GO:0005654	10	474	8.21E-05	6.08E-03
GO:1901265	9	388	9.28E-05	6.61E-03
GO:0003723	6	155	9.91E-05	6.80E-03
GO:0008152	26	2527	1.09E-04	6.97E-03
GO:0000794	3	18	1.15E-04	6.97E-03
GO:1901363	13	813	1.16E-04	6.97E-03
GO:0030054	6	177	1.99E-04	1.12E-02
GO:0036094	9	467	3.55E-04	1.83E-02
GO:0051173	9	469	3.66E-04	1.83E-02
GO:1903708	3	28	3.74E-04	1.83E-02
GO:0005634	15	1182	4.14E-04	1.90E-02
GO:0006266	2	5	4.22E-04	1.90E-02
GO:0044237	21	2027	4.64E-04	1.97E-02
GO:0046914	6	215	5.43E-04	2.19E-02
GO:1990904	4	79	5.68E-04	2.19E-02
GO:0060089	6	217	5.69E-04	2.19E-02
GO:0043902	4	80	5.94E-04	2.22E-02
GO:0036211	12	857	6.76E-04	2.42E-02
GO:0048037	4	84	7.08E-04	2.42E-02
GO:0016818	6	227	7.16E-04	2.42E-02
GO:0110165	26	2864	7.18E-04	2.42E-02
GO:0042629	2	7	7.19E-04	2.42E-02
GO:0001649	3	38	8.59E-04	2.79E-02
GO:0070942	2	8	8.95E-04	2.84E-02
GO:0031331	4	92	9.80E-04	3.06E-02
GO:0016486	2	9	1.09E-03	3.27E-02
GO:0009262	2	9	1.09E-03	3.27E-02
GO:2000108	2	10	1.31E-03	3.73E-02
GO:0010976	3	48	1.62E-03	4.37E-02

GO:0019320	2	12	1.79E-03	4.72E-02
GO:0016787	9	593	1.83E-03	4.74E-02
GO:0000166	7	377	1.95E-03	4.95E-02

Note: Proportion of risk genes: these identified risk genes (Input number) accounted for the proportion of all genes in each pathway (Background number) enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Supplementary Table 9. Significant KEGG and NHGRI GWAS catalog disease enriched by tuberculosis-associated genes (Gene set #2) identified from Sherlock Bayesian analysis of dataset #4 in the replication stage.

Disease terms	Database	Input number	Background number	P-Value	FDR
QT interval	NHGRI GWAS Catalog	6	37	4.28E-08	4.34E-05
Obesity-related traits	NHGRI GWAS Catalog	16	691	1.89E-07	7.11E-05
Ulcerative colitis	NHGRI GWAS Catalog	6	138	5.35E-05	4.48E-03
Congenital disorders of metabolism	KEGG DISEASE	12	695	1.06E-04	6.97E-03
Hematological and biochemical traits	NHGRI GWAS Catalog	3	31	4.93E-04	2.05E-02
Skin and soft tissue diseases	KEGG DISEASE	4	103	1.47E-03	4.06E-02
Skin diseases	KEGG DISEASE	4	103	1.47E-03	4.06E-02

Note: Proportion of risk genes: these identified risk genes (Input number) accounted for the proportion of all genes in each pathway (Background number) enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Supplementary Table 10. Significant pathways enriched by tuberculosis-associated genes (Gene set #3) identified from Sherlock Bayesian analysis of dataset #5 in the replication stage.

Pathway ID	Database	Input number	Background number	P-Value	FDR
R-HSA-1430728	Reactome	35	2075	2.43E-08	3.16E-05
R-HSA-392499	Reactome	33	2012	1.18E-07	5.19E-05
R-HSA-74160	Reactome	26	1448	5.42E-07	1.24E-04
R-HSA-212436	Reactome	22	1193	2.73E-06	4.24E-04
R-HSA-597592	Reactome	23	1412	1.17E-05	1.34E-03
R-HSA-73857	Reactome	22	1316	1.23E-05	1.37E-03
R-HSA-72649	Reactome	5	58	3.22E-05	3.21E-03
R-HSA-72702	Reactome	5	58	3.22E-05	3.21E-03
R-HSA-72662	Reactome	5	59	3.48E-05	3.32E-03
hsa05168	KEGG PATHWAY	12	492	4.15E-05	3.76E-03
R-HSA-72695	Reactome	4	51	2.88E-04	1.67E-02
R-HSA-72766	Reactome	8	291	3.67E-04	2.07E-02
R-HSA-9006934	Reactome	10	458	4.22E-04	2.28E-02
R-HSA-156827	Reactome	5	111	5.74E-04	2.79E-02
R-HSA-72706	Reactome	5	112	5.97E-04	2.87E-02
R-HSA-5653656	Reactome	12	669	6.51E-04	3.09E-02
R-HSA-499943	Reactome	3	28	7.46E-04	3.34E-02
R-HSA-72737	Reactome	5	119	7.76E-04	3.39E-02
R-HSA-72613	Reactome	5	119	7.76E-04	3.39E-02
R-HSA-1614517	Reactome	2	6	8.98E-04	3.55E-02
R-HSA-196807	Reactome	3	31	9.81E-04	3.76E-02
R-HSA-168273	Reactome	5	131	1.17E-03	3.98E-02
hsa05133	KEGG PATHWAY	4	76	1.19E-03	3.98E-02
R-HSA-382551	Reactome	12	720	1.21E-03	3.98E-02
R-HSA-199991	Reactome	11	631	1.35E-03	4.22E-02
hsa00983	KEGG PATHWAY	4	79	1.37E-03	4.22E-02
R-HSA-1643685	Reactome	15	1049	1.41E-03	4.22E-02
R-HSA-168255	Reactome	5	141	1.61E-03	4.56E-02
R-HSA-159763	Reactome	2	9	1.74E-03	4.78E-02

Note: Proportion of risk genes: these identified risk genes (Input number) accounted for the proportion of all genes in each pathway (Background number) enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Supplementary Table 11. Significant GO-terms enriched by tuberculosis-associated genes (Gene set #3) identified from Sherlock Bayesian analysis of dataset #5 in the replication stage.

GO-terms ID	Input number	Background number	P-Value	FDR
GO:0005515	32	1688	7.25E-09	1.41E-05
GO:0110165	42	2864	4.32E-08	3.36E-05
GO:0019222	23	1070	1.23E-07	5.19E-05
GO:0043227	33	2030	1.44E-07	5.19E-05
GO:0005488	35	2247	1.60E-07	5.19E-05
GO:0005622	34	2228	3.86E-07	1.07E-04
GO:0000166	13	377	5.26E-07	1.24E-04
GO:0043229	30	1863	6.65E-07	1.36E-04
GO:0043231	27	1606	1.12E-06	2.07E-04
GO:0043167	20	962	1.36E-06	2.31E-04
GO:0043233	16	725	7.63E-06	9.91E-04
GO:0030659	6	75	7.69E-06	9.91E-04
GO:1901363	17	813	7.90E-06	9.91E-04
GO:0046872	15	657	9.96E-06	1.17E-03
GO:0005509	7	137	2.15E-05	2.26E-03
GO:0005576	15	736	3.58E-05	3.32E-03
GO:0003723	7	155	4.56E-05	3.95E-03
GO:0005737	23	1641	1.11E-04	8.80E-03
GO:0010468	17	1012	1.13E-04	8.80E-03
GO:0032991	13	639	1.20E-04	8.80E-03
GO:1990904	5	79	1.28E-04	9.07E-03
GO:0005856	9	320	1.35E-04	9.23E-03
GO:0005739	8	260	1.76E-04	1.14E-02
GO:0031982	13	667	1.81E-04	1.14E-02
GO:0031090	11	493	1.84E-04	1.14E-02
GO:0008152	30	2527	1.91E-04	1.14E-02
GO:0032553	8	292	3.75E-04	2.07E-02
GO:0031967	6	165	4.89E-04	2.54E-02
GO:1901265	9	388	5.34E-04	2.70E-02
GO:0005654	10	474	5.47E-04	2.70E-02
GO:0009295	2	5	6.76E-04	3.10E-02
GO:0005635	4	65	6.85E-04	3.10E-02
GO:0031975	6	183	8.26E-04	3.47E-02
GO:0016787	11	593	8.30E-04	3.47E-02
GO:0044237	24	2027	8.66E-04	3.51E-02
GO:0016020	19	1443	9.02E-04	3.55E-02
GO:0043228	11	606	9.85E-04	3.76E-02
GO:0071704	28	2548	1.01E-03	3.78E-02
GO:0032549	4	74	1.09E-03	3.95E-02
GO:0016229	2	7	1.15E-03	3.96E-02
GO:0070129	2	7	1.15E-03	3.96E-02
GO:0005310	2	7	1.15E-03	3.96E-02
GO:0097367	8	351	1.20E-03	3.98E-02
GO:0033036	9	440	1.26E-03	4.10E-02
GO:0005886	12	726	1.29E-03	4.12E-02
GO:0065003	8	358	1.36E-03	4.22E-02
GO:0034707	2	8	1.43E-03	4.22E-02

GO:0003777	2	8	1.43E-03	4.22E-02
GO:0043025	4	82	1.56E-03	4.49E-02
GO:0005215	6	210	1.63E-03	4.58E-02
GO:0035091	3	38	1.70E-03	4.68E-02

Note: Proportion of risk genes: these identified risk genes (Input number) accounted for the proportion of all genes in each pathway (Background number) enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Supplementary Table 12. Significant KEGG and NHGRI GWAS catalog disease enriched by tuberculosis-associated genes (Gene set #3) identified from Sherlock Bayesian analysis of dataset #5 in the replication stage.

Disease terms	Database	Input number	Background number	P-Value	FDR
Parkinson's disease	NHGRI GWAS Catalog	6	56	1.60E-06	2.59E-04
Hematological and biochemical traits	NHGRI GWAS Catalog	4	31	4.83E-05	4.09E-03
Hematologic diseases	KEGG DISEASE	7	181	1.16E-04	8.80E-03
Congenital disorders of metabolism	KEGG DISEASE	13	695	2.66E-04	1.57E-02
Mean platelet volume	NHGRI GWAS Catalog	4	55	3.77E-04	2.07E-02
Metabolite levels	NHGRI GWAS Catalog	5	107	4.89E-04	2.54E-02
Obesity-related traits	NHGRI GWAS Catalog	12	691	8.56E-04	3.51E-02
Cardiovascular diseases	KEGG DISEASE	8	342	1.02E-03	3.79E-02
Serum total protein level	NHGRI GWAS Catalog	2	8	1.43E-03	4.22E-02
Triglycerides	NHGRI GWAS Catalog	4	81	1.49E-03	4.37E-02
QT interval	NHGRI GWAS Catalog	3	37	1.58E-03	4.52E-02
Bone mineral density	NHGRI GWAS Catalog	4	85	1.77E-03	4.81E-02
Nervous system diseases	KEGG DISEASE	13	859	1.78E-03	4.81E-02

Note: Proportion of risk genes: these identified risk genes (Input number) accounted for the proportion of all genes in each pathway (Background number) enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Supplementary Table 13. 2 common diseases enriched by tuberculosis-associated genes across 3 gene sets identified from Sherlock analysis.

Disease terms	Database	Gene set #1		Gene set #2		Gene set #3	
		Proportion of risk genes	FDR	Proportion of risk genes	Corrected P-Value	Proportion of risk genes	Corrected P-Value
Congenital disorders of metabolism	KEGG DISEASE	5.61%	4.96E-12	1.73%	6.97E-03	1.87%	1.57E-02
Obesity-related traits	NHGRI GWAS Catalog	5.07%	1.23E-09	2.32%	7.11E-05	1.74%	3.51E-02

Note: Proportion of risk genes: these identified risk genes (Input number) accounted for the proportion of all genes in each pathway (Background number) enriched by these genes. FDR values were calculated by using the method of Benjamini-Hochberg false discovery rate (FDR) correction.

Supplementary Table 15. 21 pathways enriched by using MAGMA gene-set analysis based on the KEGG pathway resource.

Pathway Name	Gene Number	Beta	MAGMA-based P values	KOBAS-based P values (genes from the discovery stage)
Pyruvate metabolism	38	0.35	4.88E-03	5.70E-02
Acute myeloid leukemia	53	0.28	5.67E-03	1.03E-02
Toxoplasmosis	121	0.20	6.47E-03	6.68E-04
Type II diabetes mellitus	44	0.30	9.75E-03	3.13E-03
Neurotrophin signaling pathway	121	0.17	9.91E-03	8.91E-04
RIG-I-like receptor signaling pathway	62	0.27	1.09E-02	5.87E-02
B cell receptor signaling pathway	70	0.20	1.55E-02	4.15E-03
Adipocytokine signaling pathway	64	0.22	1.69E-02	5.68E-02
Natural killer cell mediated cytotoxicity	128	0.16	1.81E-02	5.60E-05
VEGF signaling pathway	73	0.20	2.11E-02	1.32E-03
Insulin signaling pathway	127	0.14	2.46E-02	1.93E-03
Toll-like receptor signaling pathway	95	0.18	2.55E-02	2.26E-03
Jak-STAT signaling pathway	142	0.14	2.69E-02	4.68E-03
Drug metabolism - cytochrome P450	71	0.23	2.70E-02	4.98E-05
mTOR signaling pathway	45	0.21	3.08E-02	8.13E-04
Prostate cancer	85	0.16	3.42E-02	3.44E-02
Pancreatic cancer	66	0.16	4.16E-02	6.88E-02
Hepatitis C	128	0.12	4.62E-02	1.87E-04
Metabolism of xenobiotics by cytochrome P450	71	0.20	4.74E-02	4.90E-04

Supplementary Table 16. The proportion of multiple layers of evidence in constructed GGI network using the GeneMANIA tool.

ID	Evidence of interactions	Proportions
1	Co-expression links	71.52%
2	Predicted links	19.09%
3	Physical interactions	8.44%
4	Pathways	0.39%
5	Genetic interactions	0.31%
6	Shared protein domains	0.23%
7	Co-localization	0.02%

Supplementary Table 17. Differential expression analysis of 26 candidate genes between infected cells and uninfected cells based on the GSE133803 data.

Gene	GSM3927531 (infected cells)	GSM3927532 (infected cells)	GSM3927533 (infected cells)	GSM3927534 (uninfected cells)	GSM3927535 (uninfected cells)	GSM3927536 (uninfected cells)	P values (t-test)
RPS23	9.91	9.69	9.47	12.03	12.05	12.02	5.45E-05
RPS5	11.52	11.14	11.08	13.08	13.16	13.02	2.11E-04
CLN8	6.74	6.85	6.77	7.78	7.78	7.76	8.18E-06
SPATA20	10.15	9.77	9.63	10.97	11.02	10.95	1.98E-03
CDC16	11.36	11.11	11.05	12.33	12.45	12.22	5.63E-04
TMEM99	9.40	9.42	9.09	10.31	10.37	10.11	2.00E-03
LIG3	7.90	7.65	7.67	8.51	8.49	8.35	1.86E-03
RRM1	10.49	10.24	10.00	11.13	11.16	11.12	3.24E-03
SCAPER	8.10	8.04	7.96	8.44	8.41	8.45	8.84E-04
ZNF266	6.78	6.76	6.74	6.77	7.17	7.24	0.110
RCN3	11.07	10.79	10.91	11.37	11.51	11.31	8.73E-03
CARD9	7.13	7.39	7.09	7.55	7.44	7.43	5.48E-02
TBRG4	8.66	8.24	8.44	8.74	8.77	8.63	0.111
ZNF502	7.62	7.67	7.30	7.84	7.75	7.68	0.143
ZNF197	6.96	7.05	6.91	7.06	7.25	7.18	5.40E-02
NUDT13	7.06	6.77	6.96	7.03	7.27	6.96	0.281
HDAC10	7.14	6.78	6.88	7.05	7.06	6.98	0.424
TDRKH	6.91	6.69	6.77	6.64	6.77	6.89	0.833
PDK1	7.24	7.13	7.04	7.13	7.26	6.95	0.825
CDK10	6.77	6.74	6.59	6.68	6.59	6.68	0.494
DHX57	6.70	6.81	6.69	6.64	6.80	6.59	0.475
NPHP4	8.00	7.78	7.86	7.87	7.82	7.62	0.318
ZNF354A	7.29	7.26	7.49	7.24	7.34	7.09	0.293
FCHO1	6.77	6.90	6.65	6.42	6.75	6.67	0.267
MAP1S	10.71	10.32	10.30	9.81	9.88	9.71	1.01E-02
HIATL1	10.41	10.42	10.37	8.24	8.31	8.24	1.83E-07

Note: The P values were calculated by using the Student's t test.