



Gene Expression Profiling Based on High-Throughput Sequencing Technology and its Application in Disease Diagnosis

Dai Yuanbo *

*Wuhan University, Wuhan, China.

Corresponding Email: [*daiyuanbo2024@outlook.com](mailto:daiyuanbo2024@outlook.com)

Received: 11 September 2025 **Accepted:** 28 November 2025 **Published:** 13 January 2025

Abstract: *Pulmonary tuberculosis tends to co-occur with other diseases, leading to the emergence of drug-resistant strains of *Mycobacterium tuberculosis*, which complicates the identification and treatment of pulmonary tuberculosis. The limitations of previous methods have become increasingly evident. Therefore, based on high-throughput sequencing technology, this study aims to screen for diagnostic marker genes by comparing the differences in PBMC expression profiles between pulmonary tuberculosis patients and healthy individuals. During the research process, PBMCs were first isolated from samples, total RNA was extracted, and cDNA libraries were prepared for bridge PCR amplification. Subsequently, deep sequencing was performed using an Illumina Genome Analyzer IIx sequencer to determine the read counts of target genes, calibrate sequence tags for high-throughput sequencing data, screen for differentially expressed genes of PTB and HC with higher expression levels, and make diagnoses based on the condition.*

Keywords: *High-Throughput Sequencing Technology, Gene Expression, Disease Diagnosis, Sample Analysis, Tuberculosis Detection.*

1. INTRODUCTION

Digital Gene Expression Profiling (DGEP) refers to the construction of unbiased cDNA libraries from cells or tissues in a specific state, followed by large-scale cDNA sequencing, collection of cDNA sequence fragments, and qualitative and quantitative analysis of mRNA population composition to depict the types and abundances of gene expression in that specific



cell or tissue under a particular condition. This technology enables comprehensive, economical, and rapid detection of gene expression profiles in a specific tissue under a given condition. Through bioinformatics searches, queries, comparisons, and analyses, one can obtain information such as transcription levels, differential gene expression patterns between samples, protein function, and interrelationships(Qian N. 2024). The basic principle of Digital Gene Expression Profiling (DGEP) technology is to use double-strand enzyme digestion on cDNAs derived from reverse-transcribed mRNAs, resulting in one corresponding label for each mRNA, which then undergoes high-throughput sequencing and analysis. After bioinformatics normalization, the number of different labels across various samples is compared to identify differentially expressed labels. The differentially expressed labels obtained through DGEP undergo image recognition, base identification, and data volume output statistics processing, leading to experimental saturation analysis, determination of gene expression levels, detection of differentially expressed genes, clustering analysis of differential gene expression patterns, Gene Ontology (GO) functional enrichment analysis, pathway significance enrichment analysis, protein interaction network analysis, and new transcript prediction(Jafary F. 2024). Finally, through comparison between samples, the differential expression results of genes are obtained. In species with reference genomes and some species without reference genomes but with a single-gene sequence set database, DGEP can be directly applied. In species without both reference genomes and a single-gene sequence set database, RNA-seq technology must first be used to construct a single-gene sequence set database for that species, serving as the reference data for comparing DGEP labels (Chunhui W. 2024).

High-throughput sequencing technology is based on the principle of SBS (Synthetic-Base Sequencing), where DNA single strands are used as templates to replicate and generate complementary strands, and the colors emitted by dNTPs labeled with different fluorescent tags are used to determine different bases. This technology, which is based on next-generation high-throughput sequencing, directly sequences labels, making it more suitable for detecting low-abundance gene transcription and subtle gene changes(Chen Y. 2024). Therefore, it has been widely applied in basic scientific research, medical research, and drug development. High-throughput sequencing technology reversibly protects newly added dNTP termini with protective groups, ensuring that only one base is added per reaction and accurately removing the protective group after the base reading is complete, thus enabling successful subsequent reactions. To increase fluorescence intensity for easier acquisition by imaging systems, bridge amplification of the target fragment is necessary before sequencing(Rachmatulloh B. 2024). Currently, the main technologies for studying gene expression profiles include microarrays and high-throughput sequencing. Compared to microarrays, DGEP, a technology based on next-generation high-throughput sequencing that directly sequences labels, is more suitable



for detecting low-abundance gene transcription and subtle gene changes. Therefore, it has been widely applied in basic scientific research, medical research, and drug development(Zou X. 2024).

Based on the current research on gene expression profiling and high-throughput sequencing technology, this paper chooses digital gene expression profiling to analyze the differences in PBMC expression profiles of pulmonary tuberculosis, in order to find specific markers for the diagnosis of pulmonary tuberculosis(Liu Q. 2024).

1. Preparation of Mrna Sequencing Samples And Library Establishment

1.1 Collection and Separation of Blood Specimens

In the early morning, heparin anticoagulation was utilized to collect and screen the blood of qualified volunteers, with 20 milliliters being drawn at the outpatient clinic of the Lung Department of Shenzhen Third People's Hospital. The specific screening criteria are detailed as follows:

PTB (Pulmonary Tuberculosis): 10 cases, comprising 5 males and 5 females, aged 22.1 ± 5.13 years. According to the "Clinical Diagnosis and Treatment Guidelines for Tuberculosis" issued by the Chinese Ministry of Health, all patients underwent specific TB IFN- γ ELISPOT testing, with a spot count >30 and a positive sputum smear. They were also confirmed to have no concurrent chronic diseases or autoimmune diseases.

LTBI (Latent Tuberculosis Infection): 10 cases, including 5 males and 5 females, aged 25.6 ± 6.13 years. The local induration diameter, measured 48 hours after the tuberculin test, was ≥ 15 mm for all patients. Additionally, their IFN- γ ELISPOT results were >30 , and comprehensive examinations revealed no tuberculosis symptoms or other diseases.

HC (Healthy Control): 10 cases, with 5 males and 5 females, aged 23.7 ± 5.10 years. The local induration diameter, measured 48 hours after the tuberculin test, showed virtually no induration, and their IFN- γ ELISPOT results were all 0.

PBMCs (Peripheral Blood Mononuclear Cells), which encompass lymphocytes and monocytes, constitute a vital component of the immune system, playing a crucial role in combating infections and adapting to invasions. The method employed for isolating PBMCs in this experiment was Ficoll-Hypaque (dextran-polyethyleneimine) density gradient centrifugation. The fundamental steps involved mixing 20 milliliters of whole blood with an equal volume of PBS, adding this mixture to an equal volume of Ficoll lymphocyte separation solution, and centrifuging at 2500 revolutions per minute (rpm) at room temperature for 30 minutes. The supernatant was divided into three layers, with the middle layer being collected. The lymphocyte separation solution was washed off using 10 times its



volume of PBS, a process repeated twice. The cells were then counted, and if deemed qualified, RPMI 1640 complete medium-2 suspension was added for further sorting.

1.2 Sequencing

The selected samples were dispatched to BGI for sequencing using the Illumina Genome Analyzer IIX, with the sequencing data being utilized for further analysis. From RNA extraction to the conversion of all mRNA into cDNA and the establishment of the library, the entire process took approximately two days. The preparation procedure is illustrated in Figure 1. Once the cDNA library met the necessary requirements, it was deemed ready for sequencing (Li W. 2024).

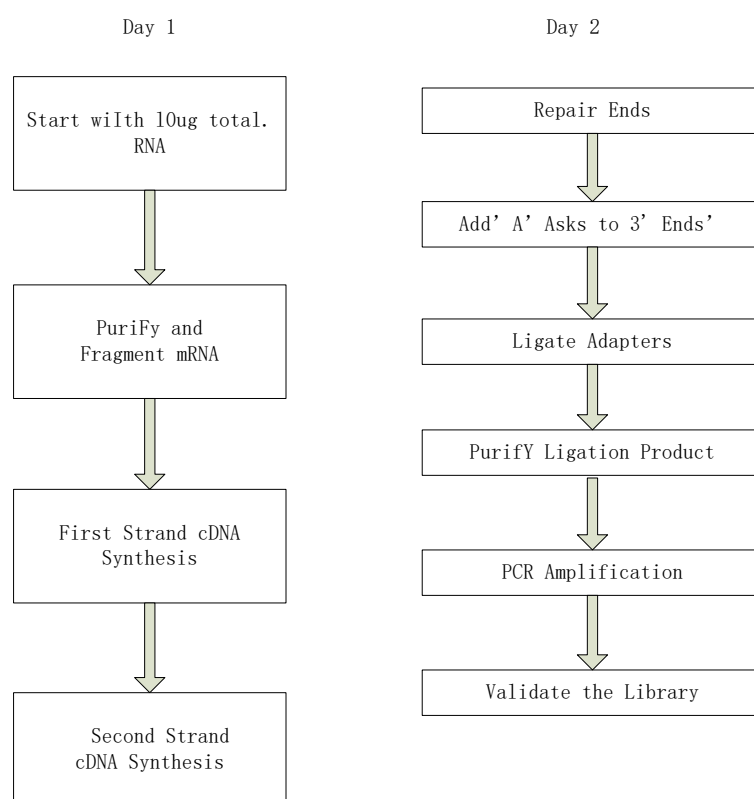


Figure 1 Workflow of mRNA sample preparation

2. Sequence Tag (Tag) Analysis of High-Throughput Sequencing Data

The tags obtained after sequencing (raw tags) are raw sequence data (raw data) with a 3 adapter, containing various impurities and a small number of low-quality tags. After removing a series of tags containing N (Tags Containing N) and Tag CopyNum<2, high-quality tags (Clean_tag) are obtained. Clean_tag is aligned to the human reference genome hg18, and tags with fewer than 2 mismatches and unique alignments are selected,



along with their corresponding genes. For a gene, there may be multiple different expression sequence tags.

For the PTB_PBMC samples, the original label generated after sequencing is 3839500, corresponding to 293077 types of labels. After filtering out labels with copy numbers less than 2 and those containing N, 3651827 high-quality labels (clean_tag) remain. Comparing these high-quality labels with the human reference genome, only tags uniquely positioned to the reference sequence and with a mismatch count less than 2 are retained. Finally, 2430312 tags of 36390 types are obtained, with 12270 genes. The human reference genome used is hg18 (NCBI Build 36.1), with 30456 genes and 275988 labels. After processing the LTBI and HC samples, the final gene counts are 12001 and 12244, respectively. Other analytical data are shown in Table 1.

In order to measure the expression level of each gene scientifically and accurately, it is necessary to standardize the expression of each gene. The standardized result is expressed by TPM (transcript per million clean tags), $TPM = \frac{\text{the number of high-quality sequencing tags compared to the gene}}{\text{the total number of high-quality sequencing tags in the sample}} \times 10^6$. Take the logarithm of the expression of genes in \log_{10} , and count the number of genes in each regional range. The expression of the three samples is roughly in the positive and skewed distribution, and the majority of the gene expression is between $0\log_{10}$ and $2\log_{10}$. Less than 10% of the genes have an expression greater than $2\log_{10}$.

Table 1 List of Label Distribution for Three Types of Sample Sequence

	PTB_PBMC		LTBI_PBMC		HC_PBMC	
	Label type	Total number of tags	Label type	Total number of tags	Label type	Total number of tags
raw data	293077	3839500	281894	3633000	306819	3636500
Tags containing N	3299	5095	3109	4854	3081	4747
Labels with a copy number less than 2	182578	182578	178369	178369	202853	202853
High quality labels	107200	3651827	100416	3449777	100885	3428900
Tags compared with	36390	2430312	33066	2078418	35009	2127374



reference genes						
number of genes	12270		12001		12244	

Note: Original data: is the data obtained after sequencing by Illumina Genome Analyzer IIX sequencer and analyzed by CASAVA1.8.2 software; the label containing N indicates that the index tag contains unknown bases; the label comparing with the reference gene refers to the comparison with the human reference gene hg18, only those with a mismatch number less than 2 and uniquely comparing with the reference gene are retained, while those with more mismatches and multiple positioning points on the reference gene are filtered out. Gene quantity: a gene has multiple different labels (Bao. 2024).

3. Screening of Differentially Expressed Genes Between PTB And HC

PTBPBMC and HCPBMC gene expression levels are compared in Figure 1. The log₂ value of the expression ratio of PTB_pbMC/HC_pbMC was obtained after standardizing the gene expression of PTB and HC samples (0.01 for genes with zero expression in a single sample), and $|\log_2 \text{Ratio}| > 1$, Genes with FDR < 0.001 are considered differentially expressed genes. PTBvsHC initially screened out 3097 genes, among which 1601 genes show PTB_PBMC > HC_PBMC (upregulated), and 1496 genes show PTB_PBMC < HC_PBMC (downregulated) (hereupregulated and downregulated refer to the PTB_PBMC samples).

Given that the number of differentially expressed genes identified in the initial screening was too large to focus on, the screening criteria were further increased by a factor of 4. The screening criteria were expanded to four times, resulting in the identification of 74 upregulated genes and 269 downregulated genes. Among the 74 upregulated genes, 48 genes are only expressed in PTB samples with low expression levels; the remaining 26 genes have an expression level greater than $2\log_{10}$ when controlling for PTB_PBMC. In the 269 downregulated genes, 118 genes are only expressed in HC, except for the LOC100286895 gene which has an expression level greater than $2\log_{10}$ in HC, while the expression levels of the other genes are relatively low. Due to their wide distribution in expression levels (Figure 3), 17 genes with expression levels greater than 300 were selected as potential target genes. The specific gene expression profiles are shown in Table 2. $\log_{10}(\text{PTB_PBMC} < \text{TPM} >)$

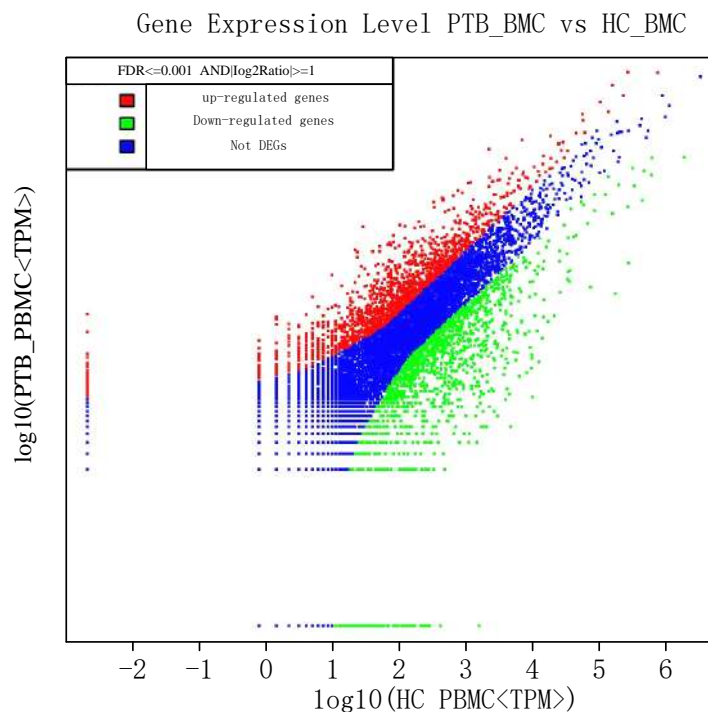


Figure 1 Comparison of gene expression levels between pulmonary tuberculosis PBMCs and normal control PBMCs

Note: X and Y axes: The expression levels of samples are all taken as log10: $FDR \leq 0.001$: refers to the P-value for controlling false positive rate ≤ 0.001 ; log2Ratio: Ratio refers to the ratio of standardized expression levels; Red part: indicates upregulated gene expression, $\log_2\text{Ratio} \geq 1$, i.e., $PTB_PBMC \geq HC_PBMC$, number of genes 1601; Green part: indicates downregulated gene expression, $\log_2\text{Ratio} \leq -1$, i.e., $PTB_PBMC \leq HC_PBMC$, number of genes 1469. Blue part: NotDEGs refers to genes with no significant difference, i.e., $|\log_2\text{Ratio}| \leq 1$.

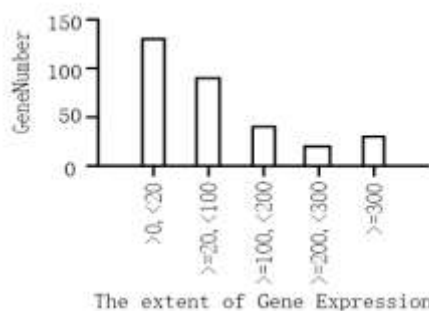


Figure 2 Distribution of expression levels of 269 down regulated genes HC-PBMC



Table 2: Significant Differences in Genes between PTB_PBMC and HC-PBMC

	Gene	PTB_PB MC	HC_PB MC	log2 Rario (PTB_PBMC/HC -PBMC)	P-Value	FDR
PTB> HC	LILRB2	289.99	16.04	4.35415914	0	0
	OSM	276.85	6.71	5.19535479	2.62E-10	2.62E-09
	NFIL3	223.45	13.71	4.15948725	0	0
	CFP	167.59	9.92	4.15686461	2.71E-13	2.44E-12
	CSF2RB	153.9	5.54	4.12354794	3.29E-09	1.93E-08
	FCGRIA	147.05	8.46	4.36254895	4.35E-16	3.31E-11
	LOC100293 412	144.86	5.82	4.98452326	1.69E-09	1.02E-08
	IFIT3	127.33	5.83	4.65966526	1.69E-10	1.02E-08
PTB< HC	LOC100286 895	0.01	103.82	-13.34179677	7.47E-113	2.19E-11
	CXCR4	110.63	355858	-5.007487104	0	0
	TWIST2	50.39	25592.08	-5.684828979	0	0
	TYROBP	58.84	1451.49	-4.623858195	0	0
	CD97	28.48	612.15	-4.42586416	0	0
	SASH3	17.37	530.78	-4.444935854	0	0
	PLEKHA2	15.66	475.95	-4.10493696	0	0
	RPS17	8.76	460.21	-4.857009515	0	0
	TP53BP2	19.17	424.63	-5.599131621	0	
	ILIA	3.655	409.46	-4.41680028	0	0
	GZMA	10.41	366.3	-6.68500465	0	0
	C1orf56	9.58	360.47	-5.124285223	0	0
	sLC29a1	7.39	358.42	-5.23370973	0	0
	PRMT8	12.6	350.26	-5.599932968	0	0
	EVI2B	18.89	337.43	-4.796930602	0	0
	CD48	6.57	322.84	-4.158893639	6.00E-307	2.11E-30
C3orf57	17.53	306.22	-5.618782159	7.14E-271	2.10E-	



						30
	SNX9	19.17	350.26	-4.126670611	1.15E-244	2.48E-26

Note: PTB>HC: genes with PTB expression greater than 2log10 and log2Ratio greater than 4 times were selected, among which OSM, NFIL3 and LILRB2 were expected to be potential candidate genes with significant differences;

PTB<HC: Select genes with HC expression levels greater than 300 and log2 Ratio greater than 4 times, among which CXCR4, TWIST2, and TYROBP are highly significantly different and potentially serve as candidate genes; these three genes have expression levels exceeding 3log10 in HC; LOC100286895 gene has an expression level greater than 2log10 in HC but does not express in LTBI patients(Ding H. 2024).

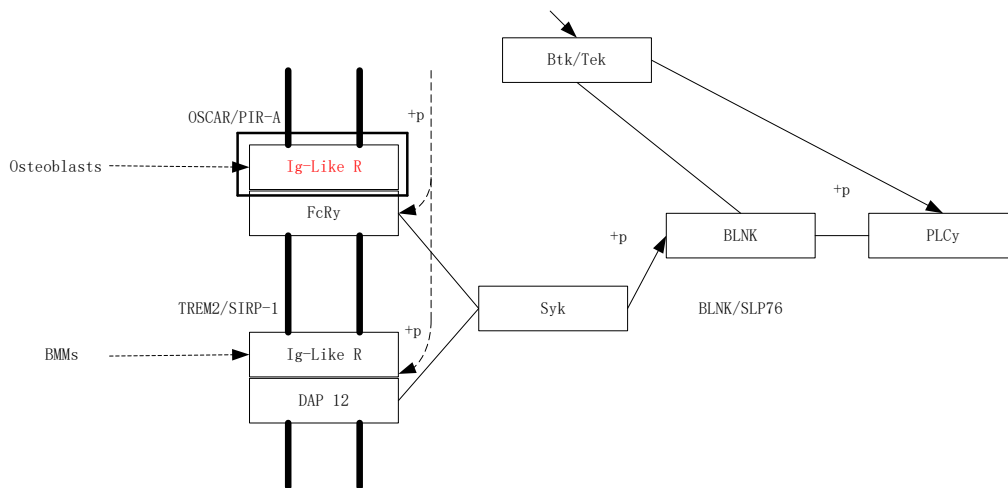
4. PTBvsHC Differential Gene GO Functional Enrichment Analysis

According to GO functional enrichment analysis, the functions of each differentially expressed gene are analyzed, followed by the identification of the differentially expressed genes within each functional entry, including both upregulated and downregulated genes. The methodology involves conducting P-value correction and selecting GO classification entries (terms) with a P-value less than 0.05. The results of the GO functional enrichment analysis for the differentially expressed genes in PTBvsHC are presented in Figure 3, encompassing cell composition analysis, molecular function analysis, and biological process analysis.

Cell composition analysis: Among the 3097 differential genes, 2572 genes can be annotated for GO cell composition analysis, revealing 41 entries that satisfy the criterion of P<0.05. The number of genes in these entries ranges from 22 to 2456, with 14 GO entries containing more than 500 genes, as depicted in Figure 4. Notably, there is no significant difference in the number of upregulated and downregulated genes (P>0.05).

Molecular function analysis: Of the 3097 differentially expressed genes, 2490 can be annotated for GO molecular function analysis, yielding 8 entries that meet the criterion of P<0.05. The number of genes in these entries varies between 13 and 2150, as illustrated in Figure 5. Similarly, there is no significant difference in the count of upregulated and downregulated genes (P>0.05).

Biological process analysis: Out of the 3097 differentially expressed genes, 2396 can be annotated for GO biological processes, revealing 81 entries that fulfill the criterion of P<0.05. The number of genes in these entries ranges from 21 to 2122, with 15 GO entries containing more than 300 genes, as shown in Figure 6. Again, there is no significant difference in the number of upregulated and downregulated genes (P>0.05).



Note: The position of LILRB2 is indicated by the Ig-like R in the black box
 Figure 3: The differentiation metabolic pathway of hsa 04380 osteoclasts where LILRB2 is located

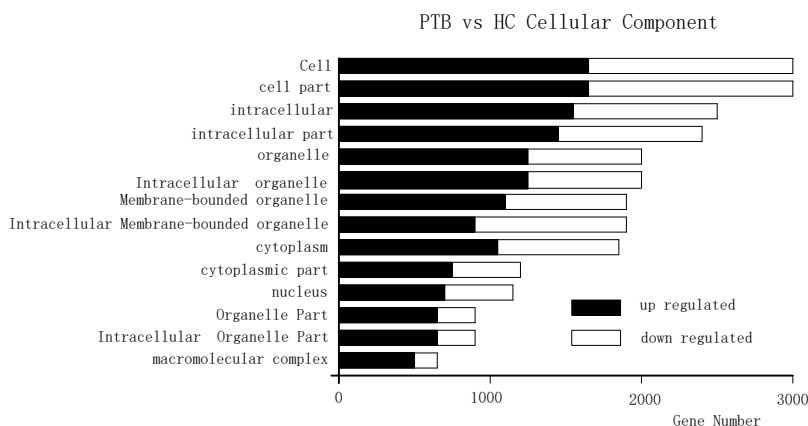


Figure 4 Composition analysis of PTB vs HC GO cells

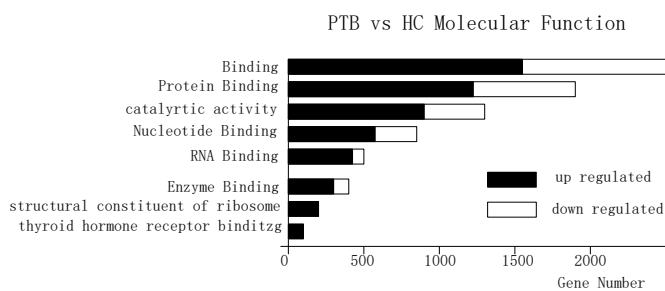


Figure 5 Functional analysis of PTB vs HC GO molecules

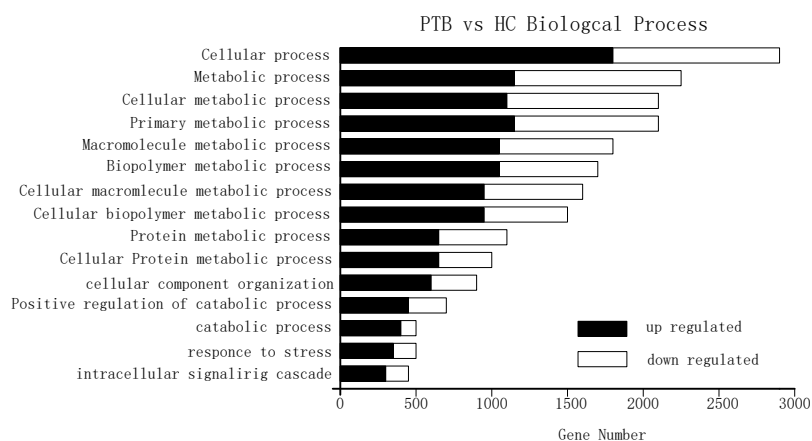


Figure 6: Biological Process Analysis of PTB vs HC GO

Note: There is no significant difference in the number of genes upregulated or downregulated for each GO entry in Figure 4, Figure 6, and Figure 6.

Discussion of the Results of Differentially Expressed Genes Screened by PTBVsHC:

Screening criteria of $FDR \leq 0.001$ and $|\log_2 \text{Ratio}| \geq 1$ identified 3097 differentially expressed genes, with 1601 upregulated and 1469 downregulated genes. Although the number of upregulated genes is higher, it can be seen from Figure 1 that the fold changes in downregulated genes are significantly higher than those in upregulated genes. When changing the screening criterion for $|\log_2 \text{Ratio}|$ to set $|\log_2 \text{Ratio}| \geq 4$, it was found that there were 74 upregulated genes (48 of which are only expressed in PTB) and 269 downregulated genes (118 of which are only expressed in HC), consistent with Figure 1. This indicates that after the body is infected with *Mycobacterium tuberculosis*, the expression levels of many genes decrease significantly or even become non-existent, leading to this phenomenon. The reason for this may be that the presence of *Mycobacterium tuberculosis* disrupts the existing immune defense system, affecting normal physiological metabolism. Whether increasing the expression levels of these genes through certain methods would enhance the body's immunity and benefit the treatment of pulmonary tuberculosis remains to be further studied. For genes that are significantly upregulated, it is possible that the body has activated these genes to protect itself against the effects of *Mycobacterium tuberculosis*, or it may be that *Mycobacterium tuberculosis* has induced the high expression of certain genes, causing further harm to the body.



2. CONCLUSION

This study aimed to analyze the gene expression levels of tuberculosis (TB) and latent tuberculosis infection (LTBI) by constructing three types of libraries: PTB library, LTBI library, and HC (healthy control) library. Initially, differential genes were screened at the gene level. Further screening was conducted based on expression levels and fold changes to identify genes with highly significant differences. GO functional enrichment analysis and KEGG metabolic pathway analysis of these differential genes were then performed to reveal how these genes function and what roles they play. The specific results are as follows:

Representative differentially expressed genes in PTBvsHC: upregulated genes include LILRB2, OSM, NFIL3, etc.; downregulated genes include CXCR4, TWIST2, TYROBP, etc. Among these, CXCR4, TWIST2, and TYROBP showed highly significant differences in PTBvsHC, suggesting that their expression may be essential for maintaining normal physiological functions. Given that their expression levels in LTBI are higher than in HC, their high expression likely plays a protective role against infection. A substantial decrease in LTBI expression, or blockade/inhibition of their expression pathways, may easily lead to the onset of TB. Currently, CXCR4 has been studied in the context of TB, but TWIST2 and TYROBP have not been explored in this field, highlighting their research value. Due to the high expression of IL8 in PTB, exceeding $4\log_{10}$, and its high detection sensitivity, coupled with its role as one of the main mediators of inflammatory responses, IL8 has potential as a biomarker for diagnosing TB.

3. REFERENCE

1. A,Qian N. Paving towards the sustainable development goals: Analyzing the nexus of financial technology,business-centric-tourism,and green growth [J]. Journal of Environmental Management,2024,371123153-123153.
2. Jafary F. Exploring the impact of financial development on energy intensity in a global study: Do levels of innovation and technology matter? [J]. Heliyon,2024,10 (21): e39331-e39331.
3. Chunhui W . Research on the High-Quality Development of Agricultural Supply Chain Financial Ecosystem Empowered by Science and Technology [J]. Proceedings of Business and Economic Studies, 2024, 7 (5): 98-105.
4. Chen Y . How does the financial technology innovation regulatory pilot influence financial regulation [J]. Finance Research Letters, 2024, 69 (PB): 106255-106255.
5. Rachmatulloh B . The Development of Financial Technology, Financial Literacy, and Financial Management Behavior in Generation Z (A Case Study of Economics Students



- at UIN Maulana Malik Ibrahim Malang) [J]. *Asian Journal of Economics, Business and Accounting*, 2024, 24 (10): 64-83.
6. Zou X . The green development effect of science and technology financial policy in China [J]. *Frontiers in Environmental Science*, 2024, 12 1463679-1463679.
 7. Liu Q . The interaction between industry-talent integration and two-phase green innovation in pharmaceutical manufacturing companies: Moderating effects of corporate financing constraints and executive short-term compensation incentives [J]. *Journal of Environmental Management*, 2024, 372 123199-123199.
 8. Li W . Environmental, social, and governance performance, financing constraints, and corporate investment efficiency: Empirical evidence from China [J]. *Heliyon*, 2024, 10 (22): e40401-e40401.
 9. Bao The impact of environmental, social, and governance (ESG) rating disparities on corporate risk: The mediating role of financing constraints [J]. *Journal of Environmental Management*, 2024, 371 123113-123113.
 10. Ding H. Environmental, Social and Corporate Governance (ESG) and Total Factor Productivity: The Mediating Role of Financing Constraints and R&D Investment [J]. *Sustainability*, 2024, 16 (21): 9500-9500.