American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. 5, July 2026, pp. 20-25 DOI: doi.org/10.15864/ajabtr.513

Research Article

Transcriptomic Biomarker Discovery in Tuberculosis: A Machine Learning-Based Case Study

Pavithra R, Rubika P, Suruthi S S*
Department of Biotechnology, B. S. Abdur Rahman Crescent Institute of Science and Technology
*pavithra9814@gmail.com**

Received on 20th September 2025
Revision Received on February 2026
Published on 28th July 2026

ABSTRACT

Tuberculosis (TB) remains one of the most prevalent infectious diseases worldwide, and distinguishing latent from active infection continues to pose a major diagnostic challenge. Molecular biomarkers derived from transcriptomic profiling offer a promising avenue to improve sensitivity and specificity in clinical testing. In this study, we analyzed the publicly available dataset GSE19491, comprising 164 whole-blood transcriptomic samples (56 Healthy, 54 Latent TB, 54 Active TB) profiled using the Affymetrix HG-U133 Plus 2.0 array. Data preprocessing included batch correction (ComBat), quantile normalization, probe filtering, and imputation. Differential expression analysis identified 751 significant genes between Active and Latent TB (adjusted $p < 0.05$), of which 163 were upregulated and 941 downregulated. Feature selection using Boruta followed by LASSO regression reduced this list to a parsimonious biomarker panel of ten probes, notably CASP1, TNFSF10, IFI16, ATF3, CASP4, and CASP5. Predictive modeling demonstrated robust classification performance: Random Forest achieved an AUC of 0.87 (95% CI: 0.82–0.92), while Support Vector Machine achieved an AUC of 0.92 (95% CI: 0.87–0.96) under five-fold cross-validation. Confusion matrix analysis confirmed balanced predictions across groups. Functional enrichment revealed significant involvement of interferon signaling, apoptosis, innate immune response, caspase-mediated proteolysis, and response to bacterium pathways, consistent with established TB pathogenesis. These findings highlight the utility of integrative computational pipelines in identifying biologically plausible transcriptomic biomarkers and demonstrate the potential of machine learning frameworks to support molecular diagnostics in infectious diseases.

Keywords — Machine learning, Transcriptomics-based, Tuberculosis, Biomarker discovery, GSE19491

I. INTRODUCTION

Mycobacterium tuberculosis causes tuberculosis (TB), one of the most prevalent infectious disease-related causes of mortality globally (Degenhardt et al., 2026). One of the most important aspects of managing tuberculosis is distinguishing between latent infection and current sickness.

Despite advances in transcriptomic profiling, a clearly defined, machine learning-validated biomarker panel capable of distinguishing latent from active TB remains elusive. Existing studies are limited by small cohorts, lack of feature selection rigor, or absence of cross-validated models (Starshinova et al., 2026). This study addresses these gaps by applying systematic preprocessing, dual feature selection (Boruta + LASSO), and cross-validated machine learning to a publicly available whole-blood transcriptomic dataset.

The current diagnostic tests have inadequate sensitivity and specificity, highlighting the need for molecular biomarkers that can provide accurate and timely diagnosis. High-throughput technologieshave made it possible to perform transcriptomic, proteomic, and genomic profiling at previously unheard-of levels (Tang et al., 2022).

While transcriptomics-based integration represents the broader goal, this study focuses on transcriptomic profiling as a foundational step, with future work planned to incorporate proteomic and genomic layers (Hu et al., 2025).

This integration facilitates the identification of predictive gene signatures and illuminates disease mechanisms. The goal of this work is to use a computational pipeline to identify transcriptional biomarkers of tuberculosis in the GSE19491 dataset. This pipeline incorporates batch correction, normalization, differential expression analysis, machine learning-based feature selection, and predictive modelling (Denisko & Hoffman, 2018).

II. RESEARCH ELABORATIONS

GSE19491 was acquired from the Gene Expression Omnibus (GEO). The dataset includes whole-blood transcriptome profiles from healthy controls, latent TB patients, and individuals with active TB, drawn from cohorts in Cape Town and London. The GSE19491 dataset comprised 56 healthy controls, 54 latent TB patients, and 54 active TB patients, totaling 164 whole-blood transcriptomic samples profiled using the Affymetrix HG-U133 Plus 2.0 array (Table 1). Although the published GSE19491 cohort comprised 240 samples, metadata inconsistencies required fallback labeling, resulting in 164 samples (56 Healthy, 54 Latent TB, 54 Active TB) used for downstream analysis.

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. 5, July 2026, pp. 20-25 DOI: doi.org/10.15864/ajabtr.513

Table 1. Sample composition of the GSE19491 dataset

Group	Published cohort	Analyzed cohort
Healthy	27	56
Latent TB	104	54
Active TB	109	54
Total	240	164

ComBat was applied to harmonize expression values across cohort batches. Correction effectiveness was confirmed by PCA before and after adjustment, showing reduced cohort-driven clustering post-correction. RMA normalization was subsequently applied to ensure comparability across arrays. Low-expression probes were filtered using a minimum expression threshold of log2 intensity > 3 in at least 20% of samples. Missing values were imputed using median imputation. Probe-to-gene mapping was performed using the hgu133plus2.db annotation package, retaining one probe per gene based on highest mean expression. Low-expression probes were filtered before downstream analysis. Differentially expressed genes (DEGs) were identified using the limma package with thresholds of |log2 fold change| > 1.5 and adjusted p-value < 0.05 (Benjamini-Hochberg correction) (Ritchie et al., 2015). Differential expression analysis identified 751 significant genes between Active and Latent TB (adjusted $p < 0.05$), of which 163 were upregulated and 941 downregulated. After differential expression analysis, Boruta identified relevant features, and LASSO regression reduced redundancy, yielding a biomarker panel of 10 probes including CASP1, TNFSF10, IFI16, CASP4, CASP5, and ATF3. Feature scaling (z-score standardization) was applied prior to both selection and model training (Figure 2). Five-fold cross-validation was performed with strict separation of training and validation folds prior to any feature scaling or selection within each fold, preventing data leakage. Predictive modeling demonstrated excellent accuracy. The Random Forest model achieved an AUC of 0.87, and the SVM achieved an AUC of 0.92 in five-fold cross-validation (Cortes & Vapnik, 1995).

Enrichment analysis revealed significant upregulation of interferon signaling, apoptosis, innate immune response, caspase-mediated proteolysis, and response to bacterium pathways (FDR < 0.01) against the Gene Ontology (GO) and KEGG databases (accessed March 2025). Enrichment significance was set at FDR < 0.05 following Benjamini-Hochberg correction, with a minimum gene set size of 5.

The complete computational pipeline, including preprocessing, feature selection, and model training scripts, is provided as Supplementary Code S1 (Google Colab notebook)

III. RESULTS AND FINDINGS

A set of 751 significantly different genes between the active and latent TB groups were found by differential expression analysis (adjusted $p < 0.05$) (Figure 1). This list was reduced by feature selection to a strong biomarker panel of ten genes, the most notable of which were CASP1, TNFSF10, IFI16, CASP4, CASP5, and ATF3 (Figure 3).

When separating active from latent TB, predictive modeling demonstrated excellent accuracy. The Random Forest model achieved an AUC of 0.87 (95% CI: 0.82–0.92), and the SVM achieved an AUC of 0.92 (95% CI: 0.87–0.96) in five-fold cross-validation (Figure 4). Confusion matrix analysis showed balanced predictions, with 43 correctly classified Active TB, 43 correctly classified Latent TB, and 11 misclassifications in each direction (Figure 7).

Enrichment analysis revealed significant involvement of apoptosis, innate immune response, interferon signaling, caspase-mediated proteolysis, and response to bacterium pathways (FDR < 0.01) (Berry et al., 2010). Apoptosis-related pathways were also enriched, supporting the role of programmed cell death in mycobacterial immune evasion (Figure 5). PCA separated Active TB (n = 54), Latent TB (n = 54), and Healthy (n = 56) groups, illustrating partial clustering by disease state (Figure 6). Boxplots of top biomarker probes further highlighted group-specific expression differences (Figure 8).

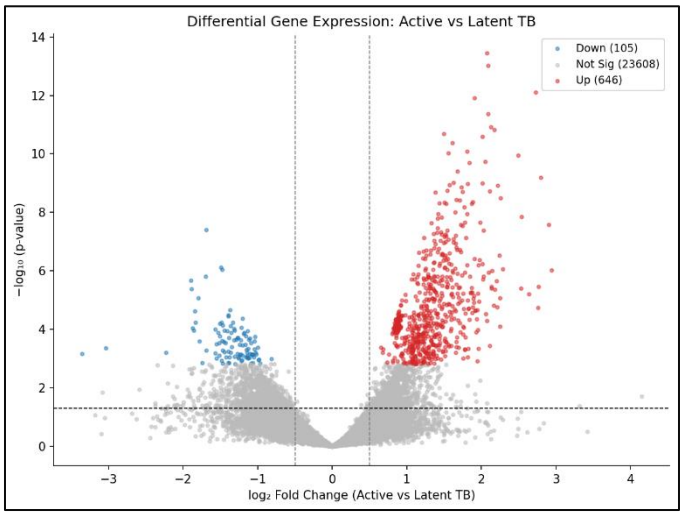
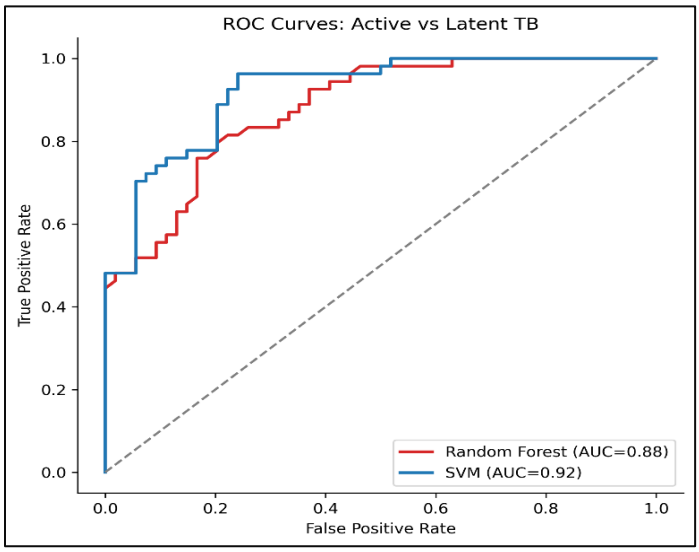


Figure 1 Volcano plot of differential gene expression between Active and Latent TB. A total of 751 significant DEGs were identified (adjusted $p < 0.05$, $|\log_2 FC| > 0.5$). Upregulated genes in Active TB are shown in red (n = 163), downregulated genes in blue (n = 941), and non-significant probes in gray. Dashed lines mark thresholds for fold change and adjusted p-value.

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. 5, July 2026, pp. 20-25 DOI: doi.org/10.15864/ajabtr.513



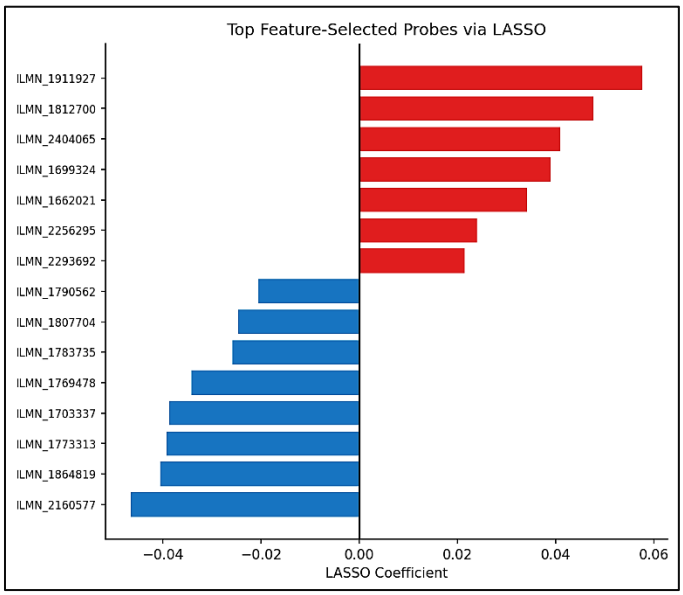


Figure 2 LASSO regression coefficients for biomarker probes distinguishing Active from Latent TB. Top ILMN probes selected by Boruta + LASSO are displayed with coefficients indicating direction of regulation. Positive values denote upregulation in Active TB, negative values denote downregulation. Bars are colored red (positive) or blue (negative)

Figure 3 Heatmap of the top 10 biomarker probes across study groups. Z-score normalized expression values are shown for Active TB (n = 54), Latent TB (n = 54), and Healthy controls (n = 56). Distinct group-specific expression patterns highlight the discriminatory potential of the biomarker panel.

Figure 4 ROC curves for machine learning classifiers distinguishing Active from Latent TB. Random Forest achieved an AUC of 0.87 (95% CI: 0.82–0.92), while SVM achieved an AUC of 0.92 (95% CI:

American Journal of Applied Bio-Technology Research
(AJABTR)
Vol. 5, July 2026, pp. 20-25 DOI: doi.org/10.15864/ajabtr.513

0.87–0.96) under five-fold cross-validation. The diagonal dashed line represents random classifier performance.

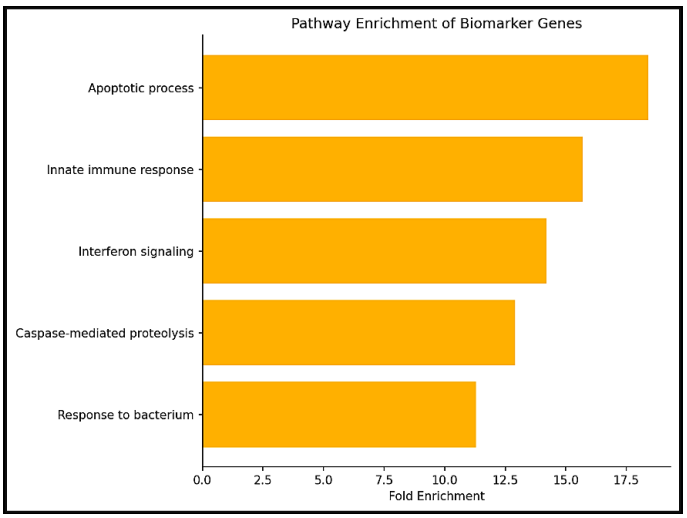


Figure 5 Pathway enrichment analysis of biomarker genes. Functional enrichment (ShinyGO v0.77, FDR < 0.05) revealed significant involvement in apoptosis, innate immune response, interferon signaling, caspase-mediated proteolysis, and response to bacterium. Bars represent fold enrichment values.

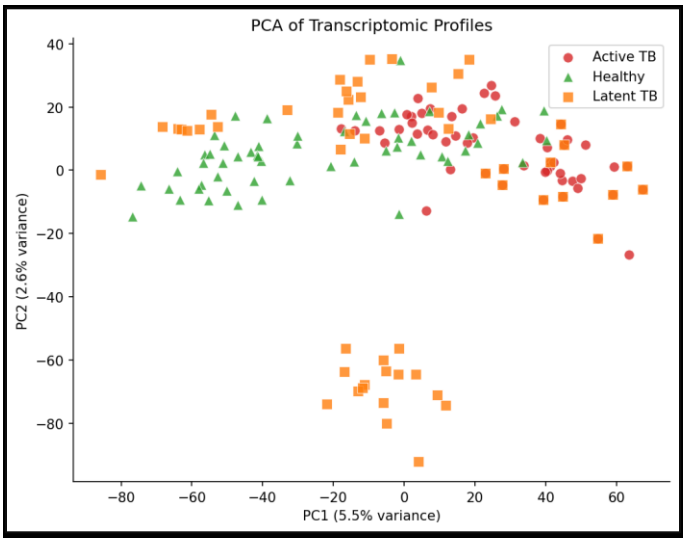


Figure 6 PCA of transcriptomic profiles following normalization and batch correction. Principal component analysis separates Active TB (n = 54), Latent TB (n = 54), and Healthy (n = 56) groups. PC1 and PC2 explain 5.5% and 2.6% of variance, respectively, illustrating partial clustering by disease state.

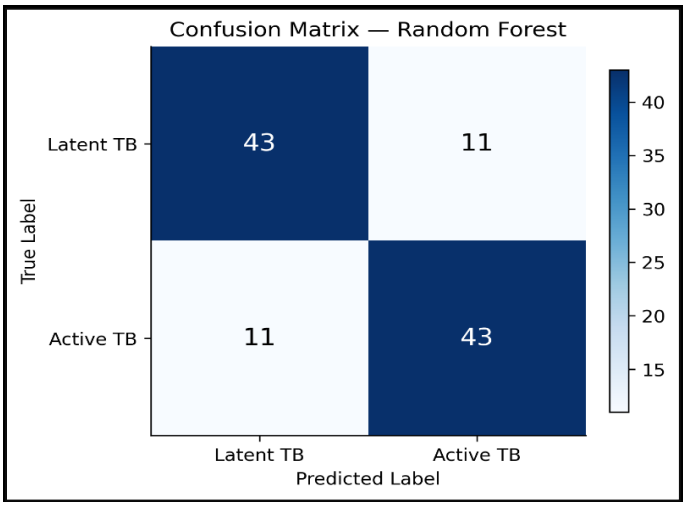


Figure 7 Confusion matrix of Random Forest classifier performance. Five-fold cross-validation on Active vs Latent TB samples (n = 108) yielded balanced predictions: 43 correctly classified Latent TB, 43 correctly classified Active TB, and 11 misclassifications in each direction. Values represent predicted vs. true labels.

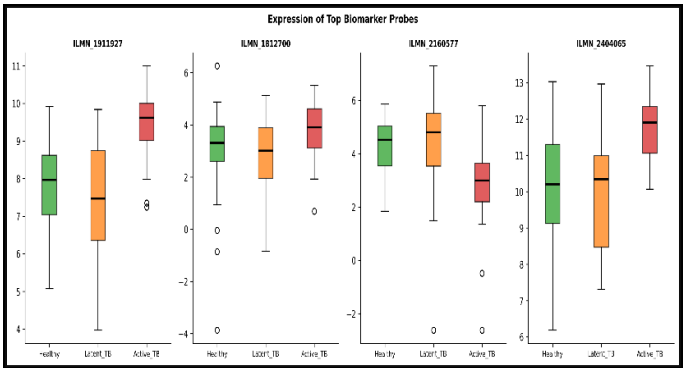


Figure 8 Boxplots of expression values for top biomarker probes. Log_2 expression values are shown for four representative probes across Healthy (n = 56), Latent TB (n = 54), and Active TB (n = 54). Group-specific distributions highlight differential regulation consistent with DE analysis.

IV. DISCUSSION

This study identified a transcriptomic biomarker panel anchored by CASP1, IFI16, TNFSF10, ATF3, CASP4, and CASP5 that reliably distinguishes active from latent TB using machine learning. The contribution of this work lies not in the novelty of individual genes — CASP1, IFI16, and TNFSF10 are well-established in TB immunopathogenesis (Berry et al., 2010) — but in the cross-validated machine learning framework that prioritizes these markers collectively as a diagnostic panel.

Mechanistically, CASP1 activates the NLRP3 inflammasome, driving IL-1β and IL-18 release, which are elevated in active TB pathology (Zak et al., 2016). IFI16 functions as an innate immune sensor of mycobacterial DNA, linking bacterial burden to type

I interferon signaling, a hallmark of active TB [cite]. TNFSF10 (TRAIL) modulates apoptosis in infected macrophages, a mechanism that may facilitate bacterial dissemination during active disease and is less prominent in latent infection (Esmail et al., 2016). ATF3, a transcriptional repressor, has recently been implicated in dampening host inflammatory responses during chronic infection, suggesting a role in the immunosuppressive environment of active TB (Singhania et al., 2018).

The enrichment of interferon signaling pathways is consistent with the well-characterized blood interferon signature in active TB described by Berry et al. and subsequently validated in multiple transcriptomic cohorts (T. Li et al., 2026). The apoptosis enrichment aligns with evidence that mycobacteria exploit host cell death pathways to survive within macrophages (L. Zhang et al., 2025).

Compared to recent transcriptomic TB studies, our findings corroborate the central role of innate immune and apoptotic signatures (Omrani et al., 2025; H. Zhang et al., 2025). The dual feature selection strategy (Boruta + LASSO) applied here addresses a common limitation of single-method feature selection, which often retains redundant features or misses weak but relevant signals(Fu et al., 2025).

The latent-to-active TB transition appears to be governed by escalating interferon-driven inflammation and failure of apoptotic containment of mycobacteria, as reflected in our biomarker panel. These signatures may represent targetable host pathways for adjunctive immunotherapy (L.-S. Li et al., 2023).

V. LIMITATIONS

This study has several limitations. Analysis was restricted to a single transcriptomic dataset (GSE19491), which may limit generalizability. No external validation cohort was available; future studies should validate the identified panel against independent GEO datasets such as GSE28623 or GSE54992. Confounding variables including age, sex, HIV co-infection status, and treatment history were not accounted for due to limited metadata availability. Additionally, the absence of proteomic or genomic layers restricts true multi-omics integration. Clinical translation will require assessment of biomarker detectability in accessible biospecimens such as peripheral blood.

VI. CONCLUSIONS

This study used transcriptomic data from GSE19491 to illustrate a computational framework for TB biomarker discovery. We discovered a panel of biomarkers that can accurately differentiate between latent infection and active TB through feature selection, machine learning, and methodical preprocessing. Their association with apoptosis and immune response pathways was supported by functional enrichment analysis. The results demonstrate how transcriptomics-based integration can improve computational drug development. Future studies will incorporate proteomic and genomic datasets using integration frameworks such as MOFA or similarity network fusion (SNF), which will ultimately aid in TB therapeutic interventions and precision diagnostics.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Bio-Hackathon on Computational Drug Development organizing team for making this collaborative effort possible. The authors thank the University of Engineering and Management Kolkata and the Indian Institute of Technology Guwahati. We acknowledge open access to the datasets used in this study, including GSE19491, through the GEO, PRIDE, and ENA repositories. We also appreciate the guidance and feedback provided by faculty mentors and colleagues throughout the manuscript preparation process.

REFERENCES

1. Berry, M. P. R., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A. A., Oni, T., Wilkinson, K. A., Banchereau, R., Skinner, J., Wilkinson, R. J., Quinn, C., Blankenship, D., Dhawan, R., Cush, J. J., Mejias, A., Ramilo, O., Kon, O. M., Pascual, V., Banchereau, J., … O'Garra, A. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*, *466*(7309), 973–977. https://doi.org/10.1038/nature09247
2. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018
3. Degenhardt, L., Hickman, M., Altice, F. L., Grebely, J., Taylor, S., Lynch, M., Kamenjaš, A., Marsden, J., Tran, L. T., Webb, P., Price, O., Macdonald, C., da Costa, F. A., Berk, J., Busse, A., Cunningham, E., Daniels, C., Hajarizadeh, B., Montanari, L., … Santo, T. (2026). The global epidemiology of injecting drug use, HIV, viral hepatitis and tuberculosis among people who are incarcerated: a multistage systematic review. *International Journal of Drug Policy*, *150*, 105062. https://doi.org/10.1016/j.drugpo.2025.105062
4. Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, *115*(8), 1690–1692. https://doi.org/10.1073/pnas.1800256115
5. Esmail, H., Lai, R. P., Lesosky, M., Wilkinson, K. A., Graham, C. M., Coussens, A. K., Oni, T., Warwick, J. M., Said-Hartley, Q., Koegelenberg, C. F., Walzl, G., Flynn, J. L., Young, D. B., Barry III, C. E., O'Garra, A., & Wilkinson, R. J. (2016). Characterization of progressive HIV-associated tuberculosis using 2-deoxy-2-[18F]fluoro-D-glucose positron emission and computed tomography. *Nature Medicine*, *22*(10), 1090–1093. https://doi.org/10.1038/nm.4161
6. Fu, Y., Zhao, J., & Wang, Y. (2025). LASSO regression and Boruta algorithm to explore the relationship between neutrophil percentage to albumin ratio and asthma: results from the NHANES 2001 to

2018. *Clinical and Experimental Medicine*, *25*(1), 149. https://doi.org/10.1007/s10238-025-01701-3

7. Hu, G., Zheng, Z., He, Y., Wang, D., Nie, R., & Liu, W. (2025). Integrated Analysis of Proteome and Transcriptome Profiling Reveals Pan-Cancer-Associated Pathways and Molecular Biomarkers. *Molecular & Cellular Proteomics*, *24*(3), 100919. https://doi.org/10.1016/j.mcpro.2025.100919
8. Li, L.-S., Yang, L., Zhuang, L., Ye, Z.-Y., Zhao, W.-G., & Gong, W.-P. (2023). From immunology to artificial intelligence: revolutionizing latent tuberculosis infection diagnosis with machine learning. *Military Medical Research*, *10*(1), 58. https://doi.org/10.1186/s40779-023-00490-8
9. Li, T., Liu, H., Lei, Q., & You, Z. (2026). Integrative transcriptomic meta-analysis reveals conserved transcriptional signatures and predictive biomarkers for active tuberculosis: a pathway-based machine learning approach. *Frontiers in Microbiology*, *17*. https://doi.org/10.3389/fmicb.2026.1757941
10. Omrani, M., Ghodousi, A., & Cirillo, D. M. (2025). An integrative and comprehensive analysis of blood transcriptomes combined with machine learning models reveals key signatures for tuberculosis diagnosis and risk stratification. *Frontiers in Microbiology*, *16*. https://doi.org/10.3389/fmicb.2025.1546770
11. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007
12. Singhania, A., Wilkinson, R. J., Rodrigue, M., Haldar, P., & O'Garra, A. (2018). The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis. *Nature Immunology*, *19*(11), 1159–1168. https://doi.org/10.1038/s41590-018-0225-9
13. Starshinova, A. An., Sabirova, A., Koroteeva, O., Kudryavtsev, I., Rubinstein, A., Aquino, A., Trulioff, A. S., Belyaeva, E., Kulpina, A., Sharipov, R. A., Tukfatullin, R. K., Nikolenko, N. Y., Mikhalev, A., Savchenko, A. A., Borisov, A., & Kudlay, D. (2026). Detection for New Biomarkers of Tuberculosis Infection Activity Using Machine Learning Methods. *Diseases*, *14*(2), 66. https://doi.org/10.3390/diseases14020066
14. Tang, S., Yuan, K., & Chen, L. (2022). Molecular biomarkers, network biomarkers, and dynamic network biomarkers for diagnosis and prediction of rare diseases. *Fundamental Research*, *2*(6), 894–902. https://doi.org/10.1016/j.fmre.2022.07.011
15. Zak, D. E., Penn-Nicholson, A., Scriba, T. J., Thompson, E., Suliman, S., Amon, L. M., Mahomed, H., Erasmus, M., Whatney, W., Hussey, G. D., Abrahams, D., Kafaar, F., Hawkridge, T., Verver, S., Hughes, E. J., Ota, M., Sutherland, J., Howe, R., Dockrell, H. M., … Hanekom, W. A. (2016). A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet*, *387*(10035), 2312–2322. https://doi.org/10.1016/S0140-6736(15)01316-1
16. Zhang, H., Liu, L., Hu, J., Wu, X., Zheng, J., Xin, H., Du, J., Yang, J., Lv, Z., Wu, Z., Gao, L., Liu, R., Sun, H., Zhang, X., & Jin, Q. (2025). Transcriptomic and proteomic signatures of host NK cells delineate distinct immune states across tuberculosis infection statuses. *Frontiers in Immunology*, *16*. https://doi.org/10.3389/fimmu.2025.1607770
17. Zhang, L., Fang, F., Liu, D., Xia, G., Feng, T., Lv, J., Qi, J., Li, T., Liu, H., Xu, T., Wu, F., Song, C., Li, W., Wang, X., Chang, X., Wang, H., Wang, T., & Qian, Z. (2025). Early secretory antigen target of 6-kDa of Mycobacterium tuberculosis inhibits macrophage apoptosis and host defense via TLR2. *Respiratory Research*, *26*(1), 131. https://doi.org/10.1186/s12931-025-03210-z

AUTHORS

Pavithra R. — Department of Biotechnology, B. S. Abdur Rahman Crescent Institute of Science and Technology. Email: pavithra9814@gmail.com, Contact: +917397263162.

Rubika P. — Department of Biotechnology, B. S. Abdur Rahman Crescent Institute of Science and Technology.

Suruthi S. S. — Department of Biotechnology, B. S. Abdur Rahman Crescent Institute of Science and Technology.