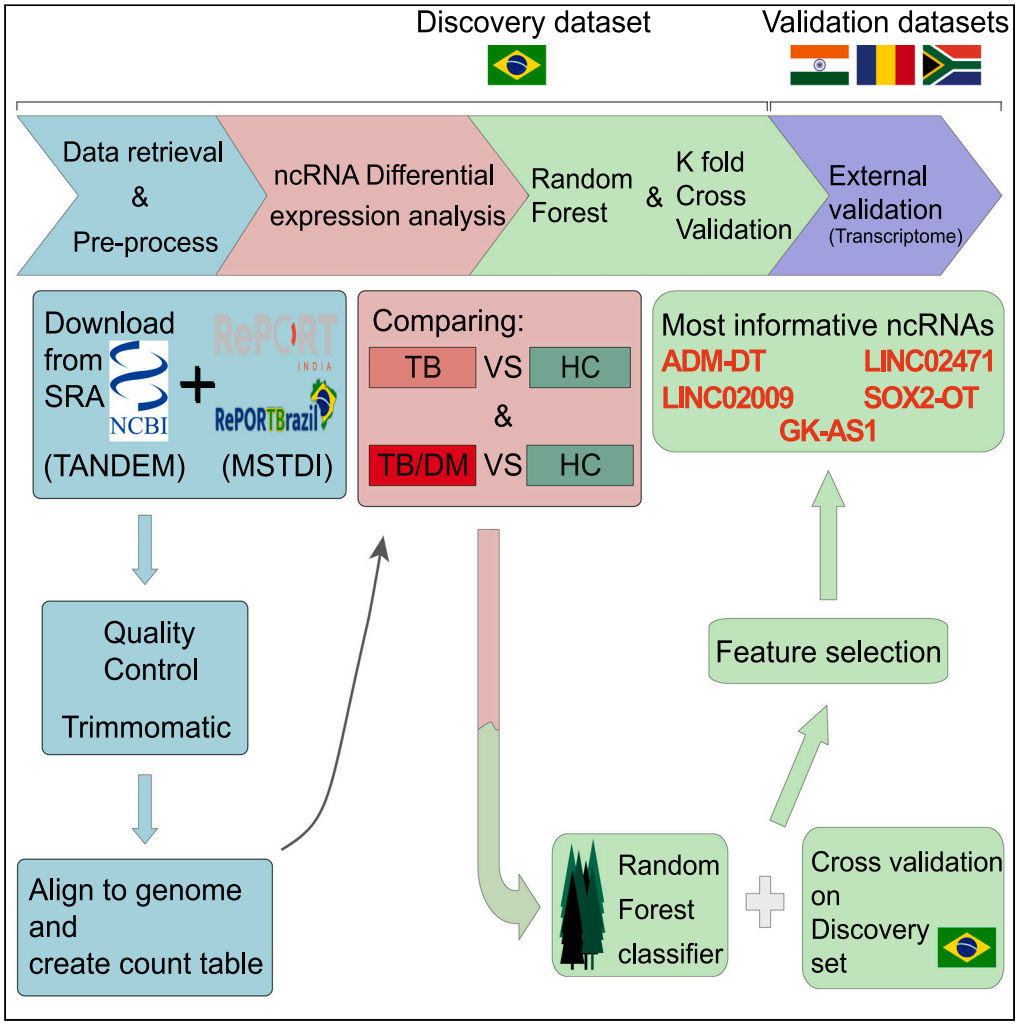
iScience

CellPress
OPEN ACCESS

Article

The sound of silent RNA in tuberculosis and the lncRNA role on infection



Eduardo Fukutani Rocha, Caian Leal Vinhaes, Mariana Araújo-Pereira, ..., Artur Trancoso Lopo de Queiroz, RePORT Brazil, RePORT India Consortia

arturlopo@gmail.com

Highlights

A distinct lncRNA signature characterizes TB regardless of DM status

lncRNA affects a range of biological pathways associated with TB pathophysiology TB

The study of lncRNA may provide new insights regarding their impacts on TB infection

Rocha et al., iScience 27, 108662
January 19, 2024
https://doi.org/10.1016/j.isci.2023.108662

iScience

CellPress
OPEN ACCESS

Article

The sound of silent RNA in tuberculosis and the lncRNA role on infection

Eduardo Fukutani Rocha,[1,3,17] Caian Leal Vinhaes,[2,3,4,17] Mariana Araújo-Pereira,[2,3,4,13] Tiago Feitosa Mota,[1,3] Akshay N. Gupte,[5] Nathella Pavan Kumar,[16] Maria Belen Arriaga,[2,3] Timothy R. Sterling,[7] Subash Babu,[6] Sanjay Gaikwad,[8] Rajesh Karyakarte,[9] Vidya Mave,[10,11] Vandana Kulkarni,[10,11] Mandar Paradkar,[10,11] Vijay Viswanathan,[12] Hardy Kornfeld,[14,15] Amita Gupta,[10] Bruno Bezerril Andrade,[2,3,4,13,18] Artur Trancoso Lopo de Queiroz,[1,2,3,18,19,*] and RePORT Brazil, RePORT India Consortia

SUMMARY

Tuberculosis (TB) is one of the leading causes of death worldwide, and Diabetes Mellitus is one of the major comorbidities (TB/DM) associated with the disease. A total of 103 differentially expressed ncRNAs have been identified in the TB and TB/DM comparisons. A machine learning algorithm was employed to identify the most informative lncRNAs: ADM-DT, LINC02009, LINC02471, SOX2-OT, and GK-AS1. These lncRNAs presented substantial accuracy in classifying TB from HC (AUCs >0.85) and TB/DM from HC (AUCs >0.90) in the other three countries. Genes with significant correlations with the five lncRNAs enriched common pathways in Brazil and India for both TB and TB/DM. This suggests that lncRNAs play an important role in the regulation of genes related to the TB immune response.

INTRODUCTION

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* (Mtb) and is one of the leading causes of death worldwide due to a single pathogen.[1] It is estimated that ¼ of the global population have been infected by the Mtb[2] and about 10% of the infected people develop the active form of TB during their lifetime.[3] The clinical presentation of active TB varies depending on the site of infection and the host inflammatory response. An aggravating condition to TB is the comorbidity with diabetes mellitus (DM), a potentially devastating medical condition with an alarming increase in its prevalence since the beginning of this century.[4] DM is characterized as a metabolic disease with pathologically high blood glucose level due to insulin action failure.[5]

The main TB aggravating factor for DM is the immunological dysfunction caused by the hyperglycemia, as it impairs both the innate and adaptive immune responses toward infections, increasing the host susceptibility to develop active TB.[6,7] The comorbidity of TB and DM (TB/DM) also worsens the treatment for TB, often leading to prolonged sputum culture conversion, and unfavorable anti-TB treatment (ATT) outcomes, such as death, treatment failure, and TB recurrence after treatment.[8] Furthermore, TB/DM is associated with an altered transcriptome and perturbations in biological pathways,[9] which may also contribute to the dysregulation of non-coding RNAs (ncRNAs).

Around 60% of the transcriptional output in human cells is represented by ncRNAs[10] and its largest type is the long non-coding (lncRNA). These are composed of 200 or more nucleotides, and despite their functions being little understood, they have major importance in regulating a large range of biological processes.[11] At genetic level, lncRNAs participate in gene expression regulation by controlling access or

[1]Centro de Integração de Dados e Conhecimentos para Saúde, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil
[2]Laboratório de Inflamação e Biomarcadores, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil
[3]Multinational Organization Network Sponsoring Translational and Epidemiological Research (MONSTER) Initiative, Salvador, Brazil
[4]Escola Bahiana de Medicina e Saúde Pública (EBMSP), Salvador 40290-150, Brazil
[5]Boston University School of Public Health, Boston, MA USA
[6]National Institutes of Health- NIRT - International Center for Excellence in Research, Chennai, India
[7]Division of Infectious Diseases, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN USA
[8]Department of Pulmonary Medicine, Byramjee-Jeejeebhoy Government Medical College and Sassoon General Hospitals, Pune, India
[9]Department of Microbiology, Byramjee-Jeejeebhoy Government Medical College and Sassoon General Hospitals, Pune, India
[10]Byramjee-Jeejeebhoy Government Medical College-Johns Hopkins University Clinical Research Site, Pune, India
[11]Johns Hopkins Center for Infectious Diseases in India, Pune, India
[12]Prof. M. Viswanathan Diabetes Research Centre, Chennai, India
[13]Faculdade de Tecnologia e Ciências, Instituto de Pesquisa Clínica e Translacional, Salvador, Brazil
[14]Department of Medicine, University of Massachusetts Medical School, Worcester, MA USA
[15]UMass Chan Medical School, Worcester, MA USA
[16]ICMR-National Institute for Research in Tuberculosis, Chennai, India
[17]These authors contributed equally to this work and share the first authorship.
[18]These authors contributed equally to this work and share the last authorship.
[19]Lead contact
*Correspondence: arturlopo@gmail.com
https://doi.org/10.1016/j.isci.2023.108662

dismissal of regulatory proteins from chromatin.[12] They can also regulate other ncRNAs[13] and microRNAs activities, by acting as microRNAs sponges as well as affecting the mRNAs translation.[14] Additionally, it is also known that lncRNAs can modify the mRNA expression by regulating the pre-mRNA splicing, editing and even stabilizing mRNAs.[15–17]

In the TB scenario, differentially expressed lncRNAs are important molecules with the role of regulating immune response pathways against Mtb. Such regulation is done on essential molecules and biological processes, such as TGF-β, IFN-γ, T- and B- cells differentiation and adaptive immune responses.[18–20] While for DM, differentially expressed lncRNAs have been mainly associated with insulin secretion by the pancreatic beta cells and insulin resistance.[21] Despite being less explored, lncRNAs emerge as potential biomarkers to evaluate the dynamics of TB infection and prognosis, as some lncRNA signatures have been previously proposed.[22,23] However, more studies are required to further validate the previously proposed TB biomarkers, including other populations.

Our group has recently described the patterns of coding gene expression in response toward TB and TB/DM in four different populations (Brazil, India, Romania, and South Africa).[24] The previous results depicted highly different patterns of gene expression, suggesting influence of population-specific differences on TB and TB/DM gene expression. In the present study, we used a robust bioinformatic approach to propose a lncRNA based biomarker for TB, which is consistently expressed through different regions and maintains its accuracy even with the TB/DM comorbidity. This biomarker was identified in RNA-seq data from patients from Brazil, enrolled by the Report Brazil[25] and had its accuracy validated in data from patients enrolled from India,[26,27] Romania, and South Africa.[28]

RESULTS

Identifying lncRNA that characterize tuberculosis and tuberculosis/diabetes mellitus

We used previously published and public data to identify differentially expressed ncRNAs and evaluate their expression. A detailed population description can be found in our previous work. In Brazil, our discovery set, a total of 189 DEGs between TB and HC groups were identified, from which 120 were upregulated and 69 were downregulated (Figure S1). Regarding the TB/DM vs. HC comparison, a total of 1128 DEGs were identified, from which 182 DEGs were upregulated and 946 were downregulated (Figure S1). Following, the lncRNAs and microRNAs were filtered from the DEGs on each comparison, i.e., TB vs. HC and TB/DM vs. HC. A total of 25 differentially expressed ncRNAs (DEncRNAs) were identified in the TB comparison, being 15 upregulated and 10 downregulated. The TB/DM comparison identified 95 DEncRNAs (28 upregulated and 67 downregulated). A summary of all identified DEGs and DEncRNAs, as well as the statistical values are available in supplemental material S1. The overall study procedure and downstream analyses are resumed in a flowchart (Figure 1).

Discriming tuberculosis and tuberculosis/diabetes mellitus using differentially expressed non-coding RNAs signature from machine learning application

After identifying 103 DEncRNAs, we applied the Random Forest (RF) machine learning algorithm to their expression data, aiming to detect the five best classifying DEncRNAs to characterize TB and TB/DM. The lncRNAs *ADM-DT*, *LINC02009*, *LINC02471*, *SOX2-OT* and *GK-AS1* were the top five features in terms of variable importance according to the RF model (Table S1). The five most informative lncRNAs' fold change was shown in Figure S2. To evaluate these genes' expression in each clinical group in Brazil, the discovery set, we employed a heatmap, displaying the z-scores of VST normalized expression data (Figure 2A). Our analysis revealed a total of two major clusters: the first was predominantly composed of patients with TB (44.4%) and TB/DM (55.5%), while the second was comprised by a mixture of HC (36.8%), TB (34.2%) and TB/DM (28.9%) (Figure 2A). Furthermore, the k-fold cross validation applied to the discovery set appointed an accuracy of 1, 95% C.I. [0.95, 1], a no-information rate of 0.42, sensitivity of 1, specificity of 1, positive and negative predictive values of 1. To check the accuracy of the identified DEncRNA signature in each different region dataset (India, Romania, and South Africa), we used Receiver Operating Characteristic (ROC) curves (Figures 2B and 2C and 2D). When classifying Indian samples, the lncRNA signature achieved an AUC of 0.86, 95% C.I. [0.78, 0.93], when classifying TB and HC samples and an AUC of 0.90, 95% C.I. [0.84, 0.97], when classifying TB/DM and HC samples (Figure 2B). A similar classifying performance was observed in the Romanian dataset, as the biomarker achieved AUCs of 0.96, 95% C.I. [0.90, 1.00], and 0.94, 95% C.I. [0.85, 1.00], when classifying TB and TB/DM from HC samples, respectively (Figure 2C). Lastly, the lncRNA signature was tested with South African samples, achieving AUCs of 0.90, 95% C.I. [0.79, 1.00], and 0.94, 95% C.I. [0.85, 1.00], when classifying TB and TB/DM from HC samples, respectively (Figure 2D). This finding has shown that the DEncRNA signature, herein identified using RF, could discriminate TB/DM with an accuracy higher than 90% in the clinical sites included in our study. Compared to the previously published TB signatures, this lncRNA signature had an overall similar performance, but could provide insights regarding potentially important lncRNAs in TB (Figure S3). Further information regarding the model is available at Table S1.

Impact of lncRNA in the overall gene expression

To assess how the signature affects the overall gene expression in each condition and region, we performed a spearman correlation analysis between the expressions of the five selected ncRNAs and all mRNA genes. The correlated genes and their respective rho and p values are available in supplemental material S2. Brazil was the site with the highest number of strong correlations, $|rho| > 0.7$, with 2292 in TB and 1214 in TB/DM. Additionally, most of them were positive correlations (85% and 89%, respectively) (Figure 3). The Indian region revealed 1484 interactions, 807 in TB and 677 in TB/DM, with 92% and 81% positive correlation, respectively (Figure 3).

As for the Romanian region, 317 and 425 correlations were identified in TB and TB/DM groups, respectively. Despite the decreased number of strong correlations, compared to Brazil and India, positive correlations were also predominant, with 74% in the TB group and 57% in the

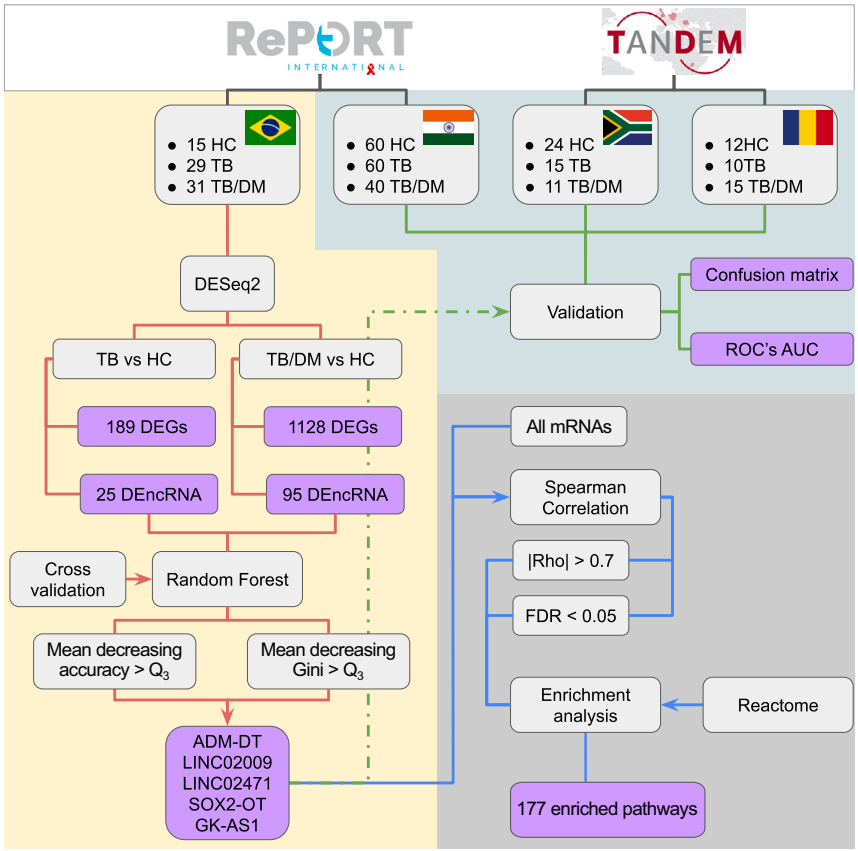


Figure 1. Flowchart of study's procedures from data acquisition through differentially expressed genes detection, feature selection and validation until enrichment analysis of genes correlated with the top five most informative DEncRNA

Red lines represent the analysis with the training dataset, Brazilian samples; Green lines represent the validation of ncRNA signature found in the random forest model; Blue lines represent the correlation and enrichment analysis. Purple boxes are illustrating the results.

TB/DM group. Lastly, the least number of correlations were observed in South Africa, being only four for the TB group (three positive and one negative) and 130 for TB/DM (78 positives and 52 negatives) (Figure 3). The result suggests a considerable impact of the ncRNA in the overall gene expression, marked mainly by positive correlations.

Impact of the correlated mRNA genes in the biological pathways

Next, to evaluate the impact of the above-mentioned correlated genes on biological pathways, we performed an enrichment analysis (Figure 4). Thus, all strongly correlated genes' Entrez IDs and their respective fold changes were used as input, grouped by the region in which the genes were correlated. A total of 177 pathways were enriched by the genes which were strongly correlated with our lncRNA TB signature. Further information about all enriched pathways is available at supplemental material S3. The top 10 gene ratio pathways for each region were identified and if the pathway was also enriched in the other regions, its respective gene ratios were retrieved. The genes which comprise these pathways had their respective correlated lncRNA assessed, to check each lncRNAs impact on the enrichment. Within the five most informative lncRNAs, three were correlated with the majority of the correlated genes comprising these pathways: *LINC02471*, *ADM-DT*, and *GK-AS1* (Figure S4). Both Neutrophil degranulation and Signaling by Interleukins were among the top 10 gene ratios in all TB infected groups (TB and TB/DM) in Brazil and India regions (Figure 4). The pathways which were among the top 10 gene ratios for at least one group, but also enriched by the other ones were Interleukin (IL)−4 and IL−13 signaling, Regulated Necrosis, Signaling by CSF3 (G−CSF), Inactivation of CSF3 (G−CSF) signaling, Diseases associated with the TLR signaling cascade, Diseases of Immune System, IRAK4 deficiency (TLR2/4) and MyD88 deficiency (TLR2/4) (Figure 4). The pathways commonly enriched by the correlated genes identified in Brazil TB, Brazil TB/DM and India TB were Programmed Cell Death, Toll−like Receptor Cascades, Interferon gamma signaling, Antigen processing−Cross presentation and ER−Phagosome pathway (Figure 4). The Interferon alpha/beta signaling was enriched only by the Brazil TB, Brazil TB/DM, and India TB/DM correlated genes. Interleukin−3, Interleukin−5, and GM−CSF signaling was enriched by the Brazil TB, India TB and India TB/DM correlated genes. The FCGR activation pathway was enriched by the Brazil TB/DM, India TB and India TB/DM correlated genes. Lastly, the Class I MHC mediated antigen processing & presentation pathway was exclusively enriched by the Brazil region (Figure 4).

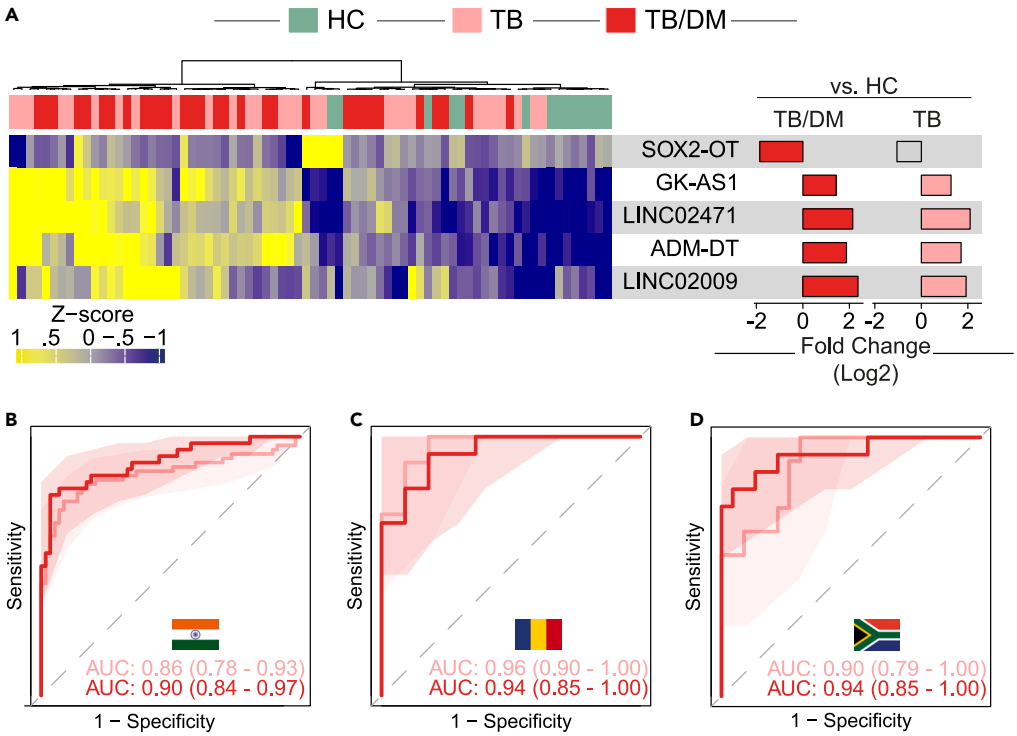


Figure 2. Random forest model and validation

(A) Heatmap displaying the Brazil region Z-scaled VST normalized expression data of the 5 classifying lncRNAs. The barplot alongside the heatmap is displaying the 5 lncRNAs log2 fold change when compared to the HC group. The bars are collored in red (TB/DM) and pink (TB) for statistically significant fold changes, while gray bars represent non significant fold changes.

(B–D) ROC curves displaying the lncRNA biomarker overall classifying performance when classifying samples from each validation dataset. The AUC with confidence interval values are depicted in each ROC curve. (B) India region dataset.

(C) Romania region dataset.

(D) South Africa region dataset.

DISCUSSION

This study aimed to evaluate the role of a TB signature composed of lncRNAs in TB and TB/DM, improving the molecular knowledge and providing further insights regarding the pathophysiology. Such insights could contribute, in the future, toward the development of new TB host directed therapy, regardless of DM comorbidity. We use the data from samples collected by the RePORT-Brazil in Salvador as the main discovery dataset, due to this data being paired end and presenting only one batch. Datasets from other regions were maintained as test datasets, as they were single-end RNA-seq data (Romania and South Africa) and were sequenced in more than one batch, demanding the application of a batch effect correction algorithm (India). This approach of using machine learning algorithms to identify biomarkers has been used before with HTLV-1,[29] mosquitoes with dengue, Zika, Chikungunya, and Yellow Fever[30,31] and to identify a predictive model in cardiovascular diseases.[32] The model composed of five lncRNAs (*ADM-DT*, *LINC02009*, *LINC02471*, *SOX2-OT*, and *GK-AS1*) achieved AUCs >0.85 when discriminating patients with TB from HC and >0.9 with TB/DM from HC, even in samples from other populations, exhibiting an outstandingly consistent accuracy. When compared with the previously proposed TB signatures, this model had similar accuracy, but provides unique insights regarding important lncRNAs in TB.

The identification of a concise transcriptomic signature to characterize TB/DM interaction has been the focus of several groups.[33,34] Recently, by applying a similar methodology, we identified a signature for TB and TB/DM composed of four mRNA genes using samples from the same cohort. Despite the signature's accuracy, it was noted that the expression of these four genes had a high degree of variability across the study regions, suggesting a strong influence of population-specific expression pattern.[24] Here we identified a more consistent pattern of expression, as three lncRNAs (ADM-DT, GK-AS1, and LINC02471) had similar fold changes for TB and TB/DM in all regions and LINC2009-fold changes were similar in Brazil, India, and South Africa. The persistent pattern of expression observed in our signature across all regions corroborates its consistency. Future studies are required to validate our findings. Despite that, the role of lncRNA emerges as a possible component promoting changes in the immunopathogenesis associated with the increased risk of persons with DM to develop active TB.[35] Once they are infected with *Mycobacterium tuberculosis*, higher is the transmission of TB among DM person[36] and more severe is the presentation of TB/DM, followed by an increased risk of unfavorable TB outcomes.[37]

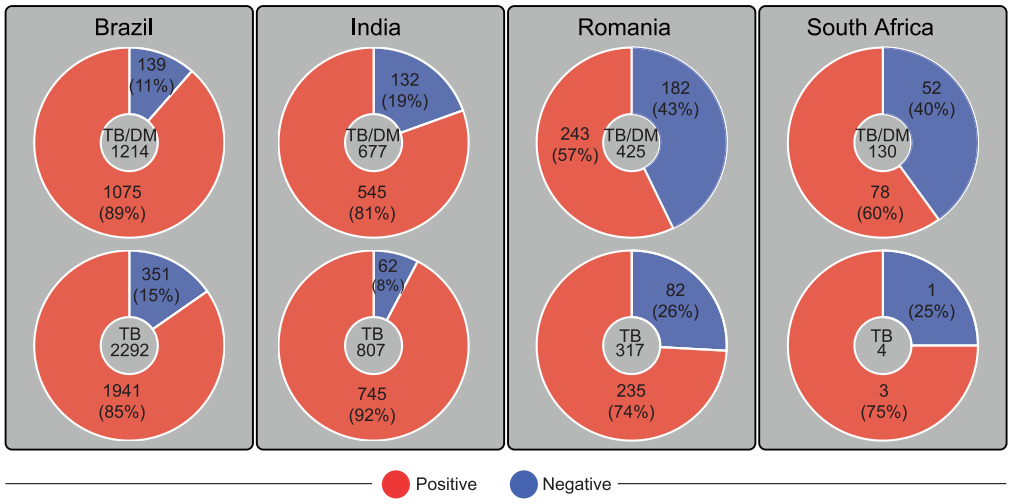


Figure 3. Correlations per region

Number of strong correlations between the lncRNA signature and overall mRNA gene expression in each region, featuring TB and TB/DM groups.

To gain insights about the role of lncRNA in the biological pathways, we used an enrichment analysis of the strongly correlated genes in Brazil and India. Most of the genes associated with the pathways were correlated to three lncRNAs (*LINC02471*, *ADM-DT* and *GK-AS1*), while the other two presented minor influence on the enrichment. Our results revealed pathways that are related to the host's immune response against Mtb. Neutrophil degranulation is one of the main neutrophil activities when facing Mtb, as they release proteins which are antimicrobial, proteolytic or even structural. These proteins are incorporated into the neutrophil's membrane to change the cellular response toward the environment.[38] The granule released molecules can inhibit the bacterial replication within the contacted macrophages,[39] but can also harm the host, as they damage both bacterial and host cells.[40] Neutrophils are the first immune cells to enter the lungs during Mtb infection, in the immunopathological side, and they are critical cells for granuloma cavitation in the active TB.[41] On the other hand, they are indispensable to control the Mtb and to induce the anti-Mtb adaptive immune response.[42] Both Signaling by and Inactivation of CSF3 (G−CSF) pathways regulate the hematopoietic proliferation of neutrophils, by the cytokine Granulocyte colony-stimulating factor (G-CSF).[43] During infections, G-CSF is induced by inflammatory cytokines, such as IL-1, TNFα and lipopolysaccharide (LPS).[44] This pathway causes its own inactivation to prevent an overpopulation of neutrophils, explaining both pathways representing its activation and inactivation being enriched simultaneously.[45]

The role of interleukins (IL) in modulating the inflammatory response toward Mtb have been largely explored.[46,47] It is known that IL-12 and IFN-γ play a crucial role in protecting the host against Mtb infection, as both molecules and their induced Th1 immune response have been extensively explored in TB.[48] There are also interleukins that can be induced by Mtb to impair the host Th1 response, such as IL-10, a potent immune regulatory interleukin. This interleukin reduces the antigen presentation and IL-12 production, enhancing intracellular bacteria survivability by inhibiting macrophages phagosomal maturation and cellular apoptosis.[49] Regarding the second enriched pathway, both IL-4 and IL-13 are associated with either the Th2 arm and have been associated with lung damage in TB.[50] IL-4 enhances macrophage endocytosis by mannose receptor, a major route of Mtb infection[51] and its suppression enhances the host resistance against Mtb in mice animal models.[52] Moreover, IL-13 upregulation in TB enhances Mtb replication and necrotizing granulomas in TB mice experimental model.[53]

The Interferon signaling is crucial for anti-TB immune response,[54] as known by exploring the IL-12/IFN-γ-mediated Th1 immune response,[55] IFN-γ macrophage and $CD8^+$ T cells activation to kill intracellular Mtb and to lyse host infected cells, respectively. Nevertheless, the excessive Th1 response activation through IFN-γ can be detrimental to the host, often leading to tecidual damage and necrosis.[56] Thus, the Interferon signaling is related to the commonly enriched pathway of Regulated Necrosis, which is mainly induced by the Th1 response byproduct and Neutrophil-produced reactive oxygen species.[57]

Regarding the last four commonly enriched pathways, MyD88 deficiency (TLR2/4) and IRAK4 deficiency (TLR2/4) are part of the Diseases associated with the TLR signaling cascade, which is a participant of the Diseases of Immune System pathway (stable ID R-HSA-5260271 in the Reactome database). Thus, all four pathways are related to Diseases of Immune System, perhaps this major pathway has been enriched due to the alterations in the host's immune system caused by the Mtb, as this pathogen impairs the host's adaptive immunological response.[58] Despite the lack of information about the biological functions regarding the five selected lncRNAs, the enrichment of important pathways to the immune response toward TB by the correlated genes is a great indicator of biological consistency in our findings.

In the present study, we have analyzed data from samples collected in Brazil, India, Romania, and South Africa and identified a set of lncRNAs with consistent accuracy at all four countries' study populations. Despite the lack of information regarding its biological functions in the literature, the five most informative lncRNAs were strongly correlated with genes associated with pathways related to immune response regulation against TB. This suggests that these lncRNAs may play an important role in the regulation of genes related to TB response, but

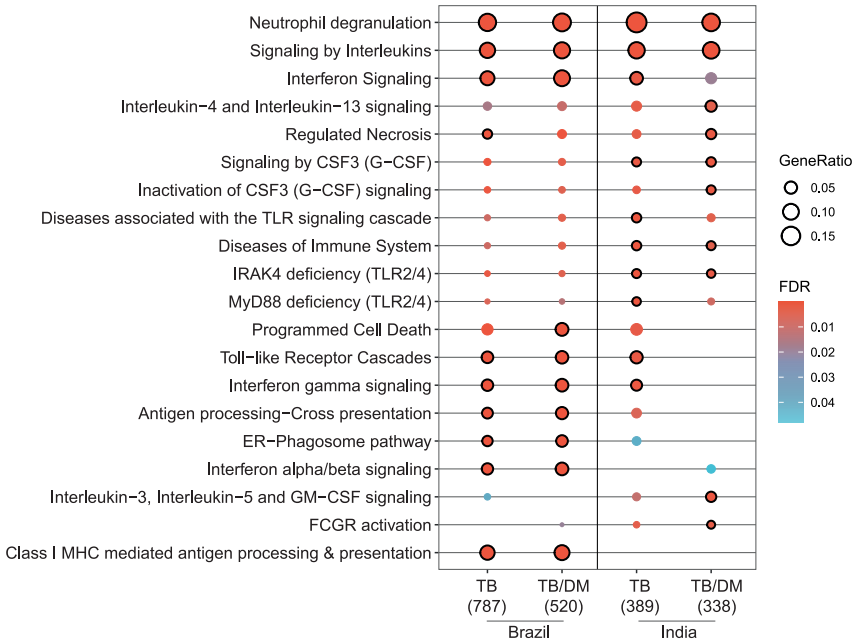


Figure 4. Correlated genes' enrichment

Dot plot displaying the top 10 gene ratio enriched pathways in each region, using the REACTOME database. Dot colors represent the statistical significance (FDR), while dot size represents the gene ratio of each enriched pathway. Dots with black circles around represent the top 10 gene ratio pathway for its respective region. The pathway names are displayed at the Y axis, while the region and group is displayed at the X axis.

further studies are still required to enlighten their biological functions and regulation mechanisms. We propose this highly consistent set of lncRNAs as biomarkers for TB, regardless of DM status.

Limitations of study

This work has some limitations, starting with the methodology used in the RNA-sequencing, as the data from South Africa and Romania were sequenced in single-end platforms, while the data from the India region has been sequenced in a paired-end platform. The employed negative binomial model for differential gene expression analysis has limitations related to multiple variable adjustments. Furthermore, as this present work employs the same samples as our previous work, all limitations regarding the metadata related to samples have been inherited as well, such as the observed differences in BMI, age, sex, smoking and alcohol use between the Brazilian and Indian populations.[24] Moreover, some patients were under treatment with metformin and statin, which could affect the inflammatory responses.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Participants enrollment and data acquisition
- METHOD DETAILS
  - Data preprocessing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Differential expression analysis and ncRNAs filtering
  - Machine learning - Random forest application and validation on independent datasets
  - Correlations lncRNAs - mRNAs and enrichment analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108662.

ACKNOWLEDGMENTS

We thank Ms. Daphne Martin and Ms. Samyra Cox for outstanding administrative support, Mr. Paul Simon and Mr. Art Garfunkel for the amazing inspiration they give to us for reporting the poetry of silent RNAs. This work was funded by: OISE-17-63459-1 from the National Institutes of Health, administered by CRDF Global; DAA3-18-64718-1, formerly USB1-31149-XX-13 from the Indo-US Vaccine Action Initiative on TB Research, administered by CRDF Global. The Brazilian site was supported by the National Institutes of Health (NIH U01AI069923 and R01AI120790), CCASAnet, RePORT-Brazil Tennessee Center for AIDS Research (TN-CFAR). The study was also supported by the Intramural Research Program of the Fundação José Silveira, and the Intramural Research Program of the Oswaldo Cruz Foundation, Brazil. BBA and ATLQ are senior investigators from the Brazilian Council for Science and Technology (CNPq). ERF and MBA received research fellowship from the Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB).

AUTHOR CONTRIBUTIONS

C.L.V., E.R.F. T.F.M, M.A.P, and M.B.A. performed the data curation, analysis, interpretation, and draft of the first version of the article. A.N.G., N.G., V.K, S.G., R.K., V.M., M.P., and A.G. performed the data interpretation, revising article critically for important intellectual content, final approval of the article. N.P.K. performed sample preparation and curation for the Chennai cohort. T.R.S. supervised the Brazilian study and helped with data interpretation. S.B. and V.V. coordinated the clinical study in Chennai. A.T.Q.L. and B.B.A. performed the study conceptualization, data analysis and interpretation, and draft of the article. H.K. performed data curation, analysis and interpretation, revising article critically for important intellectual content and coordinated all sites studies.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 5, 2023
Revised: November 27, 2023
Accepted: December 5, 2023
Published: December 8, 2023

REFERENCES

1. Global Tuberculosis Report 2021 (2021 (World Health Organization).
2. Houben, R.M.G.J., and Dodd, P.J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. PLoS Med. *13*, e1002152.
3. Narasimhan, P., Wood, J., Macintyre, C.R., and Mathai, D. (2013). Risk factors for tuberculosis. Pulm. Med. *2013*, 828939.
4. Zimmet, P., Alberti, K.G., and Shaw, J. (2001). Global and societal implications of the diabetes epidemic. Nature *414*, 782–787.
5. American Diabetes Association (2013). Diagnosis and classification of diabetes mellitus. Diabetes Care *36*, S67–S74.
6. Ruslami, R., Aarnoutse, R.E., Alisjahbana, B., van der Ven, A.J.A.M., and van Crevel, R. (2010). Implications of the global increase of diabetes for tuberculosis control and patient care. Trop. Med. Int. Health *15*, 1289–1299.
7. Mantovani, A., and Garlanda, C. (2023). Humoral Innate Immunity and Acute-Phase Proteins. N. Engl. J. Med. *388*, 439–452.
8. Jiménez-Corona, M.E., Cruz-Hervert, L.P., García-García, L., Ferreyra-Reyes, L., Delgado-Sánchez, G., Bobadilla-Del-Valle, M., Canizales-Quintero, S., Ferreira-Guerrero, E., Báez-Saldaña, R., Téllez-Vázquez, N., et al. (2013). Association of diabetes and tuberculosis: impact on treatment and post-treatment outcomes. Thorax *68*, 214–220.
9. Liu, T., Wang, Y., Gui, J., Fu, Y., Ye, C., Hong, X., Chen, L., Li, Y., Zhang, X., and Hong, W. (2022). Transcriptome analysis of the impact of diabetes as a comorbidity on tuberculosis. Medicine *101*, e31652.
10. Anastasiadou, E., Jacob, L.S., and Slack, F.J. (2018). Non-coding RNA networks in cancer. Nat. Rev. Cancer *18*, 5–18.
11. Kazimierczyk, M., Kasprowicz, M.K., Kasprzyk, M.E., and Wrzesinski, J. (2020). Human Long Noncoding RNA Interactome: Detection, Characterization and Function. Int. J. Mol. Sci. *21*, 1027.
12. Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.
13. Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.
14. Ulitsky, I. (2018). Interactions between short and long noncoding RNAs. FEBS Lett. *592*, 2874–2883.
15. Romero-Barrios, N., Legascue, M.F., Benhamed, M., Ariel, F., and Crespi, M. (2018). Splicing regulation by long noncoding RNAs. Nucleic Acids Res. *46*, 2169–2184.
16. Gott, J.M., and Emeson, R.B. (2000). Functions and mechanisms of RNA editing. Annu. Rev. Genet. *34*, 499–531.
17. Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell *39*, 925–938.
18. Fathizadeh, H., Hayat, S.M.G., Dao, S., Ganbarov, K., Tanomand, A., Asgharzadeh, M., and Kafil, H.S. (2020). Long non-coding RNA molecules in tuberculosis. Int. J. Biol. Macromol. *156*, 340–346.
19. Wang, Y., Zhong, H., Xie, X., Chen, C.Y., Huang, D., Shen, L., Zhang, H., Chen, Z.W., and Zeng, G. (2015). Long noncoding RNA derived from CD244 signaling epigenetically controls CD8+ T-cell immune responses in tuberculosis infection. Proc. Natl. Acad. Sci. USA *112*, E3883–E3892.
20. Zhang, Q., Chao, T.-C., Patil, V.S., Qin, Y., Tiwari, S.K., Chiou, J., Dobin, A., Tsai, C.-M., Li, Z., Dang, J., et al. (2019). The long noncoding RNA regulates inflammatory gene expression. EMBO J. *38*, e100041.
21. Ismail, N., Abdullah, N., Abdul Murad, N.A., Jamal, R., and Sulaiman, S.A. (2021). Long Non-Coding RNAs (lncRNAs) in Cardiovascular Disease Complication of Type 2 Diabetes. Diagnostics *11*, 145.
22. Yang, X., Yang, J., Wang, J., Wen, Q., Wang, H., He, J., Hu, S., He, W., Du, X., Liu, S., and Ma, L. (2016). Microarray analysis of long noncoding RNA and mRNA expression profiles in human macrophages infected with. Sci. Rep. *6*, 38963.

23. Hu, X., Liao, S., Bai, H., Gupta, S., Zhou, Y., Zhou, J., Jiao, L., Wu, L., Wang, M., Chen, X., et al. (2020). Long Noncoding RNA and Predictive Model To Improve Diagnosis of Clinically Diagnosed Pulmonary Tuberculosis. J. Clin. Microbiol. *58*, e01973-19.
24. Queiroz, A.T.L., Vinhaes, C.L., Fukutani, E.R., Gupte, A.N., Kumar, N.P., Fukutani, K.F., Arriaga, M.B., Sterling, T.R., Babu, S., Gaikwad, S., et al. (2023). A multi-center, prospective cohort study of whole blood gene expression in the tuberculosis-diabetes interaction. Sci. Rep. *13*, 7769.
25. van der Heijden, Y.F., Abdullah, F., Andrade, B.B., Andrews, J.R., Christopher, D.J., Croda, J., Ewing, H., Haas, D.W., Hatherill, M., Horsburgh, C.R., Jr., et al. (2018). Building capacity for advances in tuberculosis research; proceedings of the third RePORT international meeting. Tuberculosis *113*, 153–162.
26. Kornfeld, H., West, K., Kane, K., Kumpatla, S., Zacharias, R.R., Martinez-Balzano, C., Li, W., and Viswanathan, V. (2016). High Prevalence and Heterogeneity of Diabetes in Patients With TB in South India: A Report from the Effects of Diabetes on Tuberculosis Severity (EDOTS) Study. Chest *149*, 1501–1508.
27. Gupte, A., Padmapriyadarsini, C., Mave, V., Kadam, D., Suryavanshi, N., Shivakumar, S.V.B.Y., Kohli, R., Gupte, N., Thiruvengadam, K., Kagal, A., et al. (2016). Cohort for Tuberculosis Research by the Indo-US Medical Partnership (CTRIUMPH): protocol for a multicentric prospective observational study. BMJ Open *6*, e010542.
28. Eckold, C., Kumar, V., Weiner, J., Alisjahbana, B., Riza, A.-L., Ronacher, K., Coronel, J., Kerry-Barnard, S., Malherbe, S.T., Kleynhans, L., et al. (2021). Impact of Intermediate Hyperglycemia and Diabetes on Immune Dysfunction in Tuberculosis. Clin. Infect. Dis. *72*, 69–78.
29. Fukutani, E.R., Ramos, P.I.P., Kasprzykowski, J.I., Azevedo, L.G., Rodrigues, M.M.d.S., Lima, J.V.d.O.P., de Araújo Junior, H.F.S., Fukutani, K.F., and de Queiroz, A.T.L. (2019). Meta-Analysis of HTLV-1-Infected Patients Identifies CD40LG and GBP2 as Markers of ATLL and HAM/TSP Clinical Status: Two Genes Beat as One. Front. Genet. *10*, 1056.
30. Fukutani, K.F., Kasprzykowski, J.I., Paschoal, A.R., Gomes, M.d.S., Barral, A., de Oliveira, C.I., Ramos, P.I.P., and de Queiroz, A.T.L. (2017). Meta-Analysis of Expression Datasets: Comparing Virus Infection and Blood-Fed Transcriptomes to Identify Markers of Virus Presence. Front. Bioeng. Biotechnol. *5*, 84.
31. Fukutani, E., Rodrigues, M., Kasprzykowski, J.I., Araujo, C.F.d., Paschoal, A.R., Ramos, P.I.P., Fukutani, K.F., and Queiroz, A.T.L.d. (2018). Follow up of a robust meta-signature to identify Zika virus infection in Aedes aegypti: another brick in the wall. Mem. Inst. Oswaldo Cruz *113*, e180053.
32. Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., and Yan, J. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Sci. Rep. *10*, 5245.
33. Prada-Medina, C.A., Fukutani, K.F., Pavan Kumar, N., Gil-Santana, L., Babu, S., Lichtenstein, F., West, K., Sivakumar, S., Menon, P.A., Viswanathan, V., et al. (2017). Systems Immunology of Diabetes-Tuberculosis Comorbidity Reveals Signatures of Disease Complications. Sci. Rep. *7*, 1999.
34. van Doorn, C.L.R., Eckold, C., Ronacher, K., Ruslami, R., van Veen, S., Lee, J.-S., Kumar, V., Kerry-Barnard, S., Malherbe, S.T., Kleynhans, L., et al. (2022). Transcriptional profiles predict treatment outcome in patients with tuberculosis and diabetes at diagnosis and at two weeks after initiation of anti-tuberculosis treatment. EBioMedicine *82*, 104173.
35. Restrepo, B.I. (2016). Diabetes and Tuberculosis. Microbiol. Spectr. *4*.
36. Arriaga, M.B., Araújo-Pereira, M., Barreto-Duarte, B., Nogueira, B., Freire, M.V.C.N.S., Queiroz, A.T.L., Rodrigues, M.M.S., Rocha, M.S., Souza, A.B., Spener-Gomes, R., et al. (2022). The Effect of Diabetes and Prediabetes on Antituberculosis Treatment Outcomes: A Multicenter Prospective Cohort Study. J. Infect. Dis. *225*, 617–626.
37. Calderon, R.I., Arriaga, M.B., Aliaga, J.G., Barreda, N.N., Sanabria, O.M., Barreto-Duarte, B., Franco, J.P.D., Lecca, L., Andrade, B.B., Carvalho, A.C.C., and Kritski, A.L. (2022). Persistent dysglycemia is associated with unfavorable treatment outcomes in patients with pulmonary tuberculosis from Peru. Int. J. Infect. Dis. *116*, 293–301.
38. Borregaard, N., Sørensen, O.E., and Theilgaard-Mönch, K. (2007). Neutrophil granules: a library of innate immunity proteins. Trends Immunol. *28*, 340–345.
39. Tan, B.H., Meinken, C., Bastian, M., Bruns, H., Legaspi, A., Ochoa, M.T., Krutzik, S.R., Bloom, B.R., Ganz, T., Modlin, R.L., and Stenger, S. (2006). Macrophages acquire neutrophil granules for antimicrobial activity against intracellular pathogens. J. Immunol. *177*, 1864–1871.
40. Dallenga, T., and Schaible, U.E. (2016). Neutrophils in tuberculosis–first line of defence or booster of disease and targets for host-directed therapy? Pathog. Dis. *74*, ftw012.
41. Ong, C.W.M., Elkington, P.T., Brilha, S., Ugarte-Gil, C., Tome-Esteban, M.T., Tezera, L.B., Pabisiak, P.J., Moores, R.C., Sathyamoorthy, T., Patel, V., et al. (2015). Neutrophil-Derived MMP-8 Drives AMPK-Dependent Matrix Destruction in Human Pulmonary Tuberculosis. PLoS Pathog. *11*, e1004917.
42. Blomgran, R., and Ernst, J.D. (2011). Lung neutrophils facilitate activation of naive antigen-specific CD4+ T cells during Mycobacterium tuberculosis infection. J. Immunol. *186*, 7110–7119.
43. Roberts, A.W. (2005). G-CSF: a key regulator of neutrophil production, but that's not all. Growth Factors *23*, 33–41.
44. Demetri, G.D., and Griffin, J.D. (1991). Granulocyte colony-stimulating factor and its receptor. Blood *78*, 2791–2808.
45. Beekman, R., and Touw, I.P. (2010). G-CSF and its receptor in myeloid malignancy. Blood *115*, 5131–5136.
46. Kalsum, S. (2019). Characterizing Phenotypes of Mycobacterium tuberculosis and Exploring Anti-mycobacterial Compounds through High Content Screening (Linköping University Electronic Press).
47. He, X.-Y., Xiao, L., Chen, H.-B., Hao, J., Li, J., Wang, Y.-J., He, K., Gao, Y., and Shi, B.-Y. (2010). T regulatory cells and Th1/Th2 cytokines in peripheral blood from tuberculosis patients. Eur. J. Clin. Microbiol. Infect. Dis. *29*, 643–650.
48. Cooper, A.M., Kipnis, A., Turner, J., Magram, J., Ferrante, J., and Orme, I.M. (2002). Mice lacking bioactive IL-12 can generate protective, antigen-specific cellular responses to mycobacterial infection only if the IL-12 p40 subunit is present. J. Immunol. *168*, 1322–1327.
49. Abdalla, A.E., Lambert, N., Duan, X., and Xie, J. (2016). Interleukin-10 Family and Tuberculosis: An Old Story Renewed. Int. J. Biol. Sci. *12*, 710–717.
50. van Crevel, R., Karyadi, E., Preyers, F., Leenders, M., Kullberg, B.J., Nelwan, R.H., and van der Meer, J.W. (2000). Increased production of interleukin 4 by CD4+ and CD8+ T cells from patients with tuberculosis is related to the presence of pulmonary cavities. J. Infect. Dis. *181*, 1194–1197.
51. Ernst, J.D. (1998). Macrophage receptors for Mycobacterium tuberculosis. Infect. Immun. *66*, 1277–1281.
52. Buccheri, S., Reljic, R., Caccamo, N., Ivanyi, J., Singh, M., Salerno, A., and Dieli, F. (2007). IL-4 depletion enhances host resistance and passive IgA protection against tuberculosis infection in BALB/c mice. Eur. J. Immunol. *37*, 729–737.
53. Heitmann, L., Abad Dar, M., Schreiber, T., Erdmann, H., Behrends, J., Mckenzie, A.N.J., Brombacher, F., Ehlers, S., and Hölscher, C. (2014). The IL-13/IL-4Rα axis is involved in tuberculosis-associated pathology. J. Pathol. *234*, 338–350.
54. Chin, K.L., Anis, F.Z., Sarmiento, M.E., Norazmi, M.N., and Acosta, A. (2017). Role of Interferons in the Development of Diagnostics, Vaccines, and Therapy for Tuberculosis. J. Immunol. Res. *2017*, 5212910.
55. O'Garra, A., Redford, P.S., McNab, F.W., Bloom, C.I., Wilkinson, R.J., and Berry, M.P.R. (2013). The immune response in tuberculosis. Annu. Rev. Immunol. *31*, 475–527.
56. Sakai, S., Kauffman, K.D., Sallin, M.A., Sharpe, A.H., Young, H.A., Ganusov, V.V., and Barber, D.L. (2016). CD4 T Cell-Derived IFN-γ Plays a Minimal Role in Control of Pulmonary Mycobacterium tuberculosis Infection and Must Be Actively Repressed by PD-1 to Prevent Lethal Disease. PLoS Pathog. *12*, e1005667.
57. Dallenga, T., Repnik, U., Corleis, B., Eich, J., Reimer, R., Griffiths, G.W., and Schaible, U.E. (2017). M. tuberculosis-Induced Necrosis of Infected Neutrophils Promotes Bacterial Growth Following Phagocytosis by Macrophages. Cell Host Microbe *22*, 519–530.e3.
58. Chandra, P., Grigsby, S.J., and Philips, J.A. (2022). Immune evasion and provocation by Mycobacterium tuberculosis. Nat. Rev. Microbiol. *20*, 750–766.
59. Hamilton, C.D., Swaminathan, S., Christopher, D.J., Ellner, J., Gupta, A., Sterling, T.R., Rolla, V., Srinivasan, S., Karyana, M., Siddiqui, S., et al. (2015). RePORT International: Advancing Tuberculosis Biomarker Research Through Global Collaboration. Clin. Infect. Dis. 61Suppl *3*, S155–S159.
60. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

61. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
62. Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. *4*, 1521. F1000Res.
63. Jeffrey, T.L., Evan, W.J., Hilary, S.P., Elana, J.F., Andrew, E.J., John, D.S., Yuqing, Z., and Leonardo, C.T. (2017). sva: Surrogate Variable Analysis. Bioconductor R package.
64. Melissa, L., Pedro, R., and Helder, N. (2018). mdp: Molecular Degree of Perturbation calculates scores for transcriptome data samples based on their perturbation from controls. Bioconductor R package.
65. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.
66. Rainer, J., Gatto, L., and Weichenberger, C.X. (2019). ensembldb: an R package to create and use Ensembl-based annotation resources. Bioinformatics *35*, 3151–3153.
67. Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest.
68. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*.
69. Johnson, W.E., Odom, A., Cintron, C., Muthaiah, M., Knudsen, S., Joseph, N., Babu, S., Lakshminarayanan, S., Jenkins, D.F., Zhao, Y., et al. (2021). Comparing tuberculosis gene signatures in malnourished individuals using the TBSignatureProfiler. BMC Infect. Dis. *21*, 106.
70. (2023). Hmisc. https://hbiostat.org/R/Hmisc/.
71. Mukaka, M.M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med. J. *24*, 69–71.
72. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141.
73. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al. (2022). The reactome pathway knowledgebase 2022. Nucleic Acids Res. *50*, D687–D692.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw RNAseq data from TANDEM	Eckold et al.[28]	Bioproject ID: PRJNA470512
Raw RNAseq from RePORT	Kornfeld et al.,[26] Gupte et al.[27], Hamilton et al.[59]	GEOncbi ID: GSE181143
Software and algorithms		
R version 4.2.2.	R Core Team	https://cran.r-project.org/
sra-tools version 3.0.6	Available at https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software	https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view = software
Trimmomatic version 0.32	https://doi.org/10.1093/bioinformatics/btu170	http://www.usadellab.org/cms/index.php?page = trimmomatic
STAR version 2.7.10	https://doi.org/10.1093/bioinformatics/bts635	https://code.google.com/archive/p/rna-star/
tximportv version 1.28.0	https://doi.org/10.18129/B9.bioc.tximport	https://bioconductor.org/packages/release/bioc/html/tximport.html
Sva version 3.48.0	https://doi.org/10.18129/B9.bioc.sva	https://bioconductor.org/packages/release/bioc/html/sva.html
mdp version 1.20.0	https://doi.org/10.18129/B9.bioc.mdp	https://bioconductor.org/packages/release/bioc/html/mdp.html
DESeq2 version 1.40.2	https://doi.org/10.1186/s13059-014-0550-8	https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html
ensembldb version 2.24.0	https://doi.org/10.1093/bioinformatics/btz031	https://bioconductor.org/packages/release/bioc/html/ensembldb.html
randomForest version 4.7–1.1	Liaw et al.[67]	https://cran.r-project.org/web/packages/randomForest/index.html
caret version 6.0–94	Kuhn et al.[68]	https://cran.r-project.org/web/packages/caret/index.html
clusterProfiler version 4.8.2	https://doi.org/10.18129/B9.bioc.clusterProfiler	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests regarding the packages employed for the analysis performed in this study should be directed to and will be fulfilled, if possible, by the lead contact, Artur Trancoso Lopo de Queiroz (arturlopo@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

(1)The data from TANDEM have been previously deposited at the SRA database and are publicly available as of the date of publication (BioProject: PRJNA470512). The MSTDI gene expression data have been deposited at the GEO ncbi database and is publicly available as of the date of publication (GEOncbi: GSE181143, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE181143).

(2) All employed packages' references are available at the key resources table and methodology.

(3) This paper does not report original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants enrollment and data acquisition

The current study used an already published data (GEO NCBI accession number GSE181143), Regional Prospective Observational Research in Tuberculosis (RePORT), from India and Brazil consortia. Protocols have been approved by the Ethics Committees of the Prof. M. Viswanathan Diabetes Research Center and the Institutional Review Boards of Byramjee Jeejeebhoy Government Medical College, Pune and National Institute for Research in Tuberculosis and Johns Hopkins University. Participants enrolled from the RePORT Brazil had their protocols approved by Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, as well as Vanderbilt University Medical Center institutional review boards. Written informed consent was obtained from all participants. The enrollment of participants was prospective at two sites of the RePORT-India Consortium and one site of RePORT-Brazil, with organizational support provided by RePORT-International.[25] The Indian sites were located in Chennai (EDOTS study)[26] and Pune (CTRIUMPh study),[27] while the Brazilian site was in Salvador (RePORT International Common Protocol).[59] Furthermore, data from samples enrolled at the TANDEM study[28] were also retrieved, featuring samples from Indonesia, Peru, Romania and South Africa. However, only the Romania and South Africa regions had site-specific control patient data, being the only regions included in this study. Participant groups included active pulmonary TB disease, with or without DM (TB and TB/DM groups, respectively) and one control group, composed of healthy controls (HC). Inclusion criteria were age 18–65 and new diagnosis of active pulmonary TB (or absence of pulmonary TB for the control group participants). Drug-resistant TB, retreatment, treatment of incident TB for >7 days prior to enrollment, pregnancy, immunosuppressive medications and HIV infection were the exclusion criteria.

The combined 322 cohort comprised 160 participants from India, being 90 participants from Chennai and 70 from Pune (60 HC, 60 TB and 40 TB/DM), 75 participants from Brazil (15 HC, 29 TB and 31 TB/DM), 37 participants from Romania (12 HC, 10 TB and 15 TB/DM) and 50 participants from South Africa (24 HC, 11 TB and 15 TB/DM).

METHOD DETAILS

Data preprocessing

Raw RNA-seq data from the MSTDI cohort were retrieved from *Illumina HiSeq 2500* platform.[24] Sequence data from the TANDEM cohort was retrieved from the SRA database using BioProject PRJNA470512. Both MSTDI and TANDEM raw data were retrieved using the SRA tools and fastq files were processed identically to MSTDI data. Sequence data from MSTDI and TANDEM were prepared by removing low-quality bases and trimming adapters using *Trimmomatic* V0.32.[60] After the quality check, sequences were aligned against the human transcriptome (GRCh38 version) comprising both mRNA and ncRNA with the STAR algorithm v2.7.10.[61] After mapping, the outputs were converted to count tables using *tximport* package.[62] The India dataset was obtained by merging the Chennai and Pune individual datasets, followed by batch effect correction using the *sva* package.[63] Outliers detection among the samples in each individual dataset was performed by the *mdp* package,[64] after performing data normalization in each dataset using the variance stabilizing transformation, from the *DESeq2* package.[65] All data processing, post mapping, and downstream analysis have been performed in R environment v4.2.2.

QUANTIFICATION AND STATISTICAL ANALYSIS

Differential expression analysis and ncRNAs filtering

The data analysis and biomarker discovery were performed in the Brazilian dataset, consisting of 89 non-outlier samples (15 DM, 14 HC, 29 TB and 31 TB/DM), whereas the other datasets were used as validation dataset. Differential expression analysis has been employed to compare the gene expressions between the groups: HC vs. TB and HC vs. TB/DM, identifying the differentially expressed genes (DEGs). This analysis has been performed using the *DESeq2* package using the raw transcriptomic count tables. The log2 fold change and significance values of all genes were calculated by applying generalized linear models to the data, considering the mean and dispersion values of each gene. For a gene to be considered a DEG, we used the threshold of ± 1.4 log2 fold change and false discovery rate (FDR) < 0.05. Afterward, the non-coding RNAs (miRNAs and lncRNAs) were identified within the DEGs, using the *ensembldb* package query with the *homo sapiens ensdb* version AH109336 [66]. The miRNAs and lncRNAs were retrieved using the gene biotype variable and the assessed database was the Ensembl 108 EnsDb for Homo sapiens.[66]

Machine learning - Random forest application and validation on independent datasets

Afterward, the transcriptomic data containing miRNAs and lncRNAs was normalized using the *varianceStabilizingTransformation* function from the DESeq2 package. Following, we applied the random forest algorithm[67] using the DEncRNAs normalized expression data, alongside the disease categories TB and TB/DM, plus the healthy (HC) as factors for performing the classification. This algorithm aims to identify the best variables to distinguish the sample groups and has been employed due to its ability to handle multicollinearity better than linear models, such as Lasso in example. Collinearity is frequently observed in gene expression data, since different genes can be associated with the same pathway. A total of 10000 decision trees were performed by the RF, mtry parameter was set to 50. The best variables were selected using the Mean decreasing accuracy and Mean decrease gini > third quartile as criteria, which are directly related to the variable importance when classifying samples. Selected variables were retrieved from the dataset and their accuracy was evaluated using the area under the curve (AUC) value, using receiver operating characteristic (ROC) curves. The k-fold cross validation was also performed to evaluate the RF model, using the caret package's confusion matrix to assess the model's overall performance, with 100-folds and 25 repetitions.[68] The inputs for the confusion matrix were the real classes for each sample, alongside with predictions performed by the model trained with the "rf" method. The

trControl parameter was set using the trainControl function with method "repeatedcv", fold of 100 and repeating 25 times. Furthermore, the sample overall dispersion among the groups was assessed in a heatmap using the biomarkers Z-scaled expression values with Manhattan distance calculation and Ward test.

To validate the random forest model's accuracy and consistency in other populations, independent datasets with samples from India, Romania and South Africa were employed. Thus, the biomarker genes expression values were used to classificate the samples in each dataset. In order to assess the classification overall performance, ROC curves were employed for each independent dataset. The previously proposed TB signatures were retrieved from the TBSignatureProfiles package,[69] and the same method was used in order to evaluate the performance of our identified lncRNA signature in comparison to the previous TB signatures.

Correlations lncRNAs - mRNAs and enrichment analysis

To assess how the lncRNAs selected by RF could impact the overall gene expression in each region and group, we performed a correlation analysis using the Spearman rho rank coefficient between these selected lncRNAs and all mRNA genes. The correlations have been made for each TB infected group (TB and TB/DM) in each region datasets (Brazil, India, Romania and South Africa). Only highly positive/negative correlations ($|rho| > 0.7$) and False Discovery Rate (FDR) < 0.05 were considered.[70,71] Afterward, the fold changes and entrez IDs of each correlated transcript were used as input to perform the enrichment analysis, using the clusterProfiler package.[72] The enrichment analysis was performed with the REACTOME database,[73] other parameters were: Minimum gene set size = 10, Maximum gene set size = 500, q value cutoff = 0.2, p value cutoff = 0.05, with FDR as p value adjustment method.