The NEW ENGLAND JOURNAL of MEDICINE

Please review the Supplemental Files folder to review documents not compiled in the PDF.

A 29-Gene Signature Predicts Progression to Active Tuberculosis in Exposed Household Contacts

Journal:	*New England Journal of Medicine*
Manuscript ID	18-06385
Article Type:	Original Article
Date Submitted by the Author:	08-May-2018
Complete List of Authors:	Leong, Samantha; Rutgers New Jersey Medical School Zhao, Yue; Boston University Rodrigues, Rodrigo; Universidade Federal do Espirito Santo Nucleo de Doencas Infecciosas, Jones, Edward; Boston Medical Center Palaci, Moises; Núcleo de Doenças Infecciosas – Centro de Ciencias da Saúde – UFES, Alland, David; New Jersey Medical School, Medicine Dietze, Reynaldo; Núcleo de Doenças Infecciosas – Centro de Ciencias da Saúde – UFES, Ellner, Jerrold; Boston Medical Center; Boston University School of Medicine Johnson, Evan; Boston University Salgame, Padmini; Rutgers New Jersey Medical SAchool, Medicine
Abstract:	Background: One third of the world's population is latently infected with Mycobacterium tuberculosis (Mtb). Clinical management of latent tuberculosis (TB) infection (LTBI) remains a challenge due to lack of diagnostic biomarkers that can predict which individuals will progress to active TB disease. Methods: RNA-sequencing was performed on peripheral blood mononuclear cell samples from a prospective Brazilian household contact (HHC) cohort consisting of individuals with LTBI that either progressed to TB disease (progressors) or not during long-term follow-up (non-progressors). We used the RNA-Seq data from GSE79362 dataset to train a baseline biomarker, using an ensemble feature selection pipeline, and then validated this biomarker on the RNA-seq data from our Brazilian cohort. Results: The published 16-gene correlate of risk signature derived from an adolescent cohort (ACS-COR)1 performed poorly in predicting progressors from non-progressors in the Brazilian dataset. Since the performance of the ACS-COR signature improved closer to time of TB diagnosis1, we used the earliest time-point samples from progressors and non-progressors from the African dataset to derive a novel 29-gene signature that was significantly better at predicting progressors from non-progressors in the Brazilian dataset with a sensitivity of 74.2% and a specificity of 84.8%. Conclusions: PREDICT29 signature offers strong predictive performance in distinguishing non-progressors from progressors at early time-points,

	several years prior to TB diagnosis, and across geographical regions and despite ethnically distinct host and bacterial genetics. Therefore, this signature is a potential tool for risk stratification of infected individuals for targeted anti-TB preventive therapy.

"A 29-gene signature predicts progression to active tuberculosis in exposed household contacts"

Samantha Leong, PhD*[a], Yue Zhao, BS*[b], Rodrigo R. Rodriguez, PhD[c], Edward Lopez Jones, MD[d], Moises Palaci, PhD[c], David Alland, MD[a], Reynaldo Dietze, MD[c], Jerrold J. Ellner, MD[d], W. Evan Johnson, PhD[e,†] and Padmini Salgame, PhD[a†]

*Equal contribution

† Corresponding authors

[a]Centre for Emerging Pathogens, Department of Medicine, Rutgers-New Jersey Medical School, Newark, NJ, USA

[b]Division of Computational Biomedicine and Bioinformatics Program, Boston University, Boston, MA, USA.

[c]Núcleo de Doenças Infecciosas – UFES, Vitoria, Brazil

[d]Boston Medical Centre, Boston, MA, USA

[e]Department of Biostatistics, Boston University, Boston, MA

Address correspondence and reprint requests to Dr. Padmini Salgame, Center for Emerging Pathogens, Department of Medicine, MSB A901, Rutgers New Jersey Medical School, 185 South Orange Avenue, MSB Room A-902, Newark, NJ 07101. E-mail address: padmini.salgame@rutgers.edu

ABSTRACT

Background: One third of the world's population is latently infected with *Mycobacterium tuberculosis* (Mtb). Clinical management of latent tuberculosis (TB) infection (LTBI) remains a challenge due to lack of diagnostic biomarkers that can predict which individuals will progress to active TB disease.

Methods: RNA-sequencing was performed on peripheral blood mononuclear cell samples from a prospective Brazilian household contact (HHC) cohort consisting of individuals with LTBI that either progressed to TB disease (progressors) or not during long-term follow-up (non-progressors). We used the RNA-Seq data from GSE79362 dataset to train a baseline biomarker, using an ensemble feature selection pipeline, and then validated this biomarker on the RNA-seq data from our Brazilian cohort.

Results: The published 16-gene correlate of risk signature derived from an adolescent cohort (ACS-COR)[1] performed poorly in predicting progressors from non-progressors in the Brazilian dataset. Since the performance of the ACS-COR signature improved closer to time of TB diagnosis[1], we used the earliest time-point samples from progressors and non-progressors from the African dataset to derive a novel 29-gene signature that was significantly better at predicting progressors from non-progressors in the Brazilian dataset with a sensitivity of 74.2% and a specificity of 84.8%.

Conclusions: PREDICT29 signature offers strong predictive performance in distinguishing non-progressors from progressors at early time-points, several years prior to TB diagnosis, and across geographical regions and despite ethnically distinct host and bacterial genetics. Therefore, this signature is a potential tool for risk stratification of infected individuals for targeted anti-TB preventive therapy.

INTRODUCTION

With one-third of the world's population estimated to be latently infected with *Mycobacterium tuberculosis* (Mtb), the World Health Organization's guidelines on management of latent tuberculosis infection (LTBI) call for better strategies for testing and treating LTBI, particularly pointing out the need for methods to determine risk of LTBI progression to active tuberculosis (TB) disease.[2] Administration of preventive anti-tuberculous treatment to persons with LTBI can reduce the subsequent number of TB disease diagnoses.[5,6] However, in TB endemic regions, mass treatment of all latently-infected individuals is not practical nor cost-effective. Therefore, the identification of Mtb-infected subjects most likely to progress to disease could enable the targeting of anti-TB preventive therapy to those most likely to benefit.

Several blood signatures differentiating TB from LTBI have been reported[7-14], but studies centered on predicting the outcome of Mtb infection are limited.[15] One study found that the expression of *IL-13* and *AIRE* could identify individuals that progressed to TB within 8 months prior to disease.[16] In addition, a large prospective biomarker study of African cohorts identified a 16-gene (henceforth denoted ACS-COR) whole blood signature capable of predicting TB progression with 53.7% sensitivity and 82.8% specificity when disease diagnosis occurred within 12 months of sample collection.[1] Whereas this study demonstrates the utility of blood transcriptomics as a biomarker for progression risk, sensitivity is a key limitation to its conclusions.[17] Importantly, this signature's predictive performance, most notably its sensitivity, decreases with increasing time to disease onset. For example, the sensitivity of their approach decreased to 39.3% when disease diagnosis occurred between 361 and 720 days of sample collection.[1] Thus, further studies in additional cohorts to identify a more stable and broadly applicable signature that predicts progression to TB disease are required. Therefore, the goal of this study was to use differential transcriptomic profiling in peripheral blood samples of Mtb-infected progressors prior to their TB diagnoses versus long-term persistent non-progressors to

identify a transcriptomic signature capable of predicting risk of TB progression at much earlier time points, even several years prior to disease.

METHODS

Household contact study design and subject inclusion criteria

Subject groups in the International Collaboration for Infectious Disease Research (ICIDR) household contact (HHC) study included index TB cases and household contacts, as described previously.[34,35] Briefly, index cases were eligible for enrollment if they were consenting HIV-negative adult (≥ 18 years old) pulmonary TB cases living in a household with 3 or more contacts and had a 2+ or greater sputum acid fast bacilli (AFB) smear, positive AFB culture, and a history of cough ≥ 3 months. Enrolled household contacts (HHCs) included consenting HIV-negative individuals of all ages who had close contact with the index case for at least 3 months. Close contact was defined as meeting at least one of the following criteria: sleeping under the same roof ≥ 5 days/week, sharing meals ≥ 5 days/week, watching TV nights or weekends, or other significant contact, such as visiting the household > 18 days/month. Any subjects known to be HIV-positive or unable to provide consent were excluded.

At the time of enrollment, all HHCs were screened via tuberculin skin testing. Additional subject data, including age, gender, socioeconomic information, and health history, as well as household environmental evaluation were also collected. HHCs were passively followed post-enrollment to monitor for changes in TB disease status denoted in their medical records. Individuals who were diagnosed with TB disease during follow-up have been retrospectively classified as progressors. Individuals who have remained healthy during long-term follow-up (>4 years) were considered non-progressors.

Ethics Statement

The study was approved by the Comite de Ética em Pesquisa do Hospital Universitário Cassiano Antonio de Morais, the Rutgers Biomedical Health Sciences Institutional Review Board, and the Boston Medical Center Institutional Review Board. Written informed consent and assent was obtained from all study participants as per the consent procedure approved by IRBs from all participating institutions.

RESULTS

HHC follow up for progression to TB disease

From March 2008 to May 2015, 410 index cases and 1203 of their household contacts (HHCs) were enrolled in a prospective observational cohort study in Vitória, Brazil (Figure S1). At the time of enrollment, HHCs were screened for tuberculin skin test (TST) reactivity; 573 HHCs had a positive TST, indicating infection with Mtb. Of these TST-positive contacts, 6 (1%) were clinically diagnosed with TB disease at the time of enrollment (denoted co-prevalent TB). During post-enrollment follow-up, 28 (5%) additional TST-positive contacts were later diagnosed with TB disease (progressors). All samples in one of the many shipments received from Brazil were of poor quality. This shipment included PBMCs from 12 progressors and these had to be excluded from the study. For this study, we therefore profiled the baseline (time of enrollment) gene expression in peripheral blood mononuclear cells (PBMC) of 16 of the progressors. These individuals developed TB disease between 11 and 1795 days after enrollment and baseline blood collection. Five of the 16 progressors were diagnosed within the first three months of sample collection (Tables S1-S2). TST-positive contacts who were not diagnosed with TB disease long-term (>4 years) were considered latent TB infected (LTBI) non-progressors. We selected 21 age- and gender-matched non-progressors as controls for this study. In addition, randomly selected PBMC from 14 TB index patients from the cohort were also studied. PBMC obtained at baseline from the 16 progressors, 21 non-progressors, and 14 TB patients were used for RNA sequencing (RNA-seq) analysis (Tables S1-S2, Protocol S15).

Derivation of a 29-gene signature for predicting progressors from non-progressors

The ACS-COR progression risk signature was initially derived from genes differentially expressed between non-progressors and progressors at the time-point most proximal to their TB diagnosis.[1,18] Not surprisingly, the signature shows strong predictive performance closer to TB diagnosis and also distinguishes TB patients from LTBI individuals with high accuracy[1,19], We hypothesized that if we selectively used 'baseline' progressor samples (from the African RNA-seq dataset) at the time point furthest from their eventual TB diagnosis, we could derive a signature that is specific to risk of progression in the early stages of infection. This might lead to a signature that is independent of inflammatory responses that may occur proximally to or during the clinical expression of TB disease. By this logic, we used the existing RNA-seq dataset from the Zak et al. (2016) study (GSE79362) to train a baseline biomarker, using an ensemble feature selection pipeline, and then validated this biomarker on the RNA-seq data from our Brazilian cohort. The training RNA-seq dataset profiles subjects from the prospective South African adolescent cohort study (ACS cohort) containing 46 progressors and 107 matched controls; samples were collected every six months, ranging from baseline to up to two years per subject.[1] For this analysis, available sequencing data for 39 progressor samples from time-points furthest from their TB diagnosis dates as well as 103 non-progressors were used for predictive model training based on gene signatures (Figure S2). Then our Brazilian dataset, which was smaller in size, provided a new independent testing set of LTBI progressors and non-progressors for assessing predictive performance of gene signatures. This approach of independent training and validation in an ethnically distinct cohort should yield highly robust biomarkers of disease, and has not been previously applied to genomic biomarkers of progression.

Briefly, initial identification of potential genes (features) of interest involved applying an interquartile range filter, days to progression correlation filter, and differential gene expression

filter, after which we identified 639 putative biomarker genes. The next step for feature selection used an ensemble model combining random forest, lasso logistic regression, and gradient boosting, which resulted in selection of 89 genes. Then, 40 of these 89 genes were selected based on a single lasso logistic regression classifier (Figure S2; Tables S2-S4; Protocol S15). Finally, 29 out of these 40 genes were selected based on their mapping to protein-coding genes (Figure S2). As shown in the heatmap (Figure 1A), the 29-gene signature, henceforth designated PREDICT29, discriminated progressors from non-progressors. The data were further analyzed using Principal Component Analysis (PCA) based on these 29 genes. Consistent with the heat map, the PCA plot also corroborated that progressors and non-progressors segregated into two main clusters (Figure 1B). The resultant 29 genes contained no overlapping genes with the TB or progression risk signatures tested, except for four genes (*SRBD1*, *WARS*, *APOL6*, and *TCN2*) of the Berry 393-gene signature derived from TB versus LTBI individuals.

Validation of the PREDICT29 signature in predicting Brazilian progressors from non-progressors

We next aimed to test the predictive performance of our new PREDICT29 signature via various predictive model methods. Model training for four methods (ridge logistic regression, SVM, random forest, gradient boosting) was performed using the African training dataset and tested in the Brazilian dataset (Table 1A and Figure 2A). The best predictive model (XGBoost method) derived from the PREDICT29 signature yielded a mean AUC of 0.915 (0.900, 0.929) with a sensitivity of 0.785 (0.749, 0.821) and specificity of 0.844 (0.808, 0.880). Additionally, PREDICT29 performed well across all four models used, suggesting its reproducibility and versatility across modeling methods. When averaging performance across the four models tested, the PREDICT29 signature resulted in a mean AUC of 0.890 (0.876, 0.905) with a sensitivity of 0.760 (0.725, 0.794) and specificity of 0.825 (0.792, 0.857). Of the 16 progressors,

5 were diagnosed with TB within three months of sample collection and are likely subclinical disease. We therefore tested whether the fidelity of the PREDICT29 signature would improve if the subclinical cases were omitted from the analysis. The performance did improve, although not significantly: mean AUC 0.911 (0.894, 0.928), sensitivity of 0.742 (0.704, 0.780) and specificity of 0.848 (0.816, 0.880) (Table 1B and Figure 2B).

Predicting Brazilian progressors from non-progressors using the ACS-COR gene signature

Performance of the ACS-COR signature improves closer to time of TB diagnosis.[1] In addition, the ACS-COR signature also segregates with high accuracy TB disease from LTBI.[1,20] We wanted to determine whether the ACS-COR signature would work on the Brazilian samples where TB disease developed in the progressors after greater than three years following exposure to the index case (Table 2A and Figure 2D). In evaluating the ACS-COR progression risk signature[1], we observed that an SVM-based model offered some predictive value in the Brazilian dataset that was comparable to the predictive performance reported in the original publication. The best ACS-COR progression risk signature-based model (SVM method) yielded a mean AUC of 0.792 (0.768, 0.817), sensitivity of 0.741 (0.698, 0.783) and specificity of 0.687 (0.643, 0.730). In contrast, the ACS-COR signature performed significantly less better in the other three models tested, which is not surprising given that the ACS-COR signature was originally derived using an SVM-based method.[1] Thus, when averaging the ACS-COR signature's performance across the four models, this yielded a lower mean AUC of 0.670 (0.640, 0.700) with a sensitivity of 0.515 (0.460, 0.571)) and a specificity of 0.774 (0.728, 0.820). Even after omitting the subclinical cases from the analysis, the ACS-COR gene signature still under-performed in classifying progressors from non-progressors (Table 2B and Figure 2E).

Predictive performance was also evaluated for models of the existing ACS-COR and novel PREDICT29 progression signatures in the African training dataset via 100-fold cross-validation,

and both signatures performed roughly equivalently (Tables S6-S7). Thus, whereas the ACS-COR signature did not perform better than the PREDICT29 signature in the African dataset, it was noted that the ACS-COR signature performed better in the African dataset than in the Brazilian dataset.

Predictive performance of PREDICT29 and ACS-COR signatures in classifying TB from LTBI

We evaluated the predictive performance of the novel PREDICT29 signature in predicting TB patients from LTBI non-progressors to distinguish its specificity for predicting progression risk as opposed to being a TB disease signature (Table 3A and Figure 2C). The best model derived from the PREDICT29 signature (SVM method) yielded a mean AUC of 0.764 (0.740, 0.787), sensitivity of 0.651 (0.612, 0.691) and specificity of 0.810 (0.775, 0.844). Overall, whereas the PREDICT29 signature performed worse in predicting TB patients from LTBI non-progressors compared to its performance in predicting progressors from non-progressors, the reverse pattern was seen for the ACS-COR signature, which performed better in TB disease prediction. Given the strong predictive performance of the ACS-COR signature for distinguishing TB patients from LTBI individuals[1,19], it was not surprising that it also performed well in classification of TB disease from LTBI non-progressors in the Brazilian dataset (Table 3B and Figure 2F); the best model (ranger method) yielding a mean AUC of 0.961 (0.954, 0.968), sensitivity of 0.780 (0.755, 0.806) and specificity of 0.953 (0.935, 0.970). Together, the data suggest that the PREDICT29 signature may offer a better predictor of eventual progression risk as opposed to subclinical TB disease.

Several existing TB disease signatures and the 4-gene pan-African risk to progression signature perform poorly in distinguishing progressors from non-progressors

Evaluation of predictive performance of four existing TB disease signatures revealed that none of the signatures performed well in classifying progressors from non-progressors with all average AUC values <0.65 (Tables S8-S11). Thus, these results suggest that the existing TB disease signatures do not offer discriminatory ability between progressors and non-progressors.

A 4-transcript signature (RISK4) derived from a HHC cohort capable of predicting progression up to 2 years prior to disease onset in HHC[21] also performed poorly in segregating progressors from non-progressors in the Brazilian cohort (Table S12).

Functional characterization of the PREDICT29 signature

KEGG pathway analysis identified 25 pathways from the PREDICT29 signature (Table S13), and the majority were related to cellular metabolism or related to cell homeostasis. This was followed by enrichment in pathways broadly categorized as physiological processes. In addition, an immunological pathway, natural killer cell mediated cytotoxicity, was identified. Ingenuity Pathway Analysis also yielded findings consistent with KEGG pathway analysis (Table S14).

DISCUSSION

In this study, we aimed to identify predictive blood-based signature that can accurately indicate an individual's risk of progression from Mtb infection to disease. Evaluation of predictive performance of the existing ACS-COR and RISK4 signatures for TB progression risk demonstrated its moderate ability to distinguish progressors in our new independent RNA-seq dataset derived from our prospective observational cohort study in Brazil. However, our novel PREDICT29 signature offered superior performance in predicting progressors in the Brazilian dataset. A follow-up to the ACS-COR study, described the sequential inflammatory processes involved in the progression to TB disease, suggesting elevated type I/II IFN-signaling at 18 months before diagnosis followed by changes in immune cell subset profiles more proximal to

disease.[18] The PREDICT29 signature's lack of inflammatory gene or pathway enrichment and weak ability for distinguishing between TB disease and latent TB infection suggests that this signature provides specific detection of progression risk at early time-points prior to eventual TB diagnosis. Characterization of the PREDICT29 individual genes suggests that most of these genes appear to be related to gene expression regulation and cellular metabolism and homeostasis, which was also reflected by KEGG pathway enrichment analysis of the gene set. PREDICT29 signature may have captured processes occurring in the early host response to Mtb that could dictate successful long-term pathogen control or permit progressive disease.

Although the same RNA-seq prospective African cohort dataset was used to derive both the ACS-COR signature and our novel PREDICT29 signature, markedly different signatures resulted. This is largely attributable to the differing methodologies for RNA-seq data processing, biomarker selection, and sample utilization. Most critical to our methodology was our selective use of samples obtained from the earliest time-point available for each individual so as to capture transcriptional responses as distal from the date of TB disease diagnosis as possible. The gene selection by Zak and colleagues, based on samples most proximal to diagnosis likely biased their ACS-COR signature to detect subclinical TB, during which TB disease-related inflammatory processes are likely already occurring although clinical manifestations of the disease may not be present yet.[18,22,23] Consistent with this, the ACS-COR signature performed well in distinguishing active TB disease from LTBI in multiple datasets[1,19] as well as in the present study using the Brazilian dataset. Another difference in our methodologies was that we adopted a classic approach of ensemble feature selection, which has been used successfully in cancer biomarker studies.[24,25] On the other hand, Zak et al. employed a method that involved measuring gene expression abundance at the level of splice junction counts by quantifying frequency of mRNA splicing events.[1]

The PREDICT29 signature derived from the African dataset was obtained from whole blood RNA-seq while validation was performed on the Brazil dataset that was derived from RNA-seq performed with PBMC. Thus the PREDICT29 signature works on PBMC that do not contain neutrophils. This further supports that the PREDICT29 signature is not related to the interferon-inducible neutrophil-driven blood transcriptional signature that segregates TB from LTBI [7]. In addition, the fidelity of the PREDICT29 genes did not change when subclinical cases were removed from the analysis further highlighting that the genes and pathways associated with this signature could provide insights to the immune mechanisms regulating protection from progression to disease. Overall, our novel PREDICT29 signature yields robust predictive performance to discriminate progressors from non-progressors prior to TB diagnosis, and it therefore offers a potential clinical screening tool to identify infected individuals with increased risk for disease progression. Additional testing in larger validation cohorts would need to be performed prior to possible implementation of this PREDICT29 panel as a clinical tool.

Acknowledgements

We acknowledge the dedication and hard work of the field staff and study team. This work was funded by the National Institute of Allergy and Infectious Diseases, National Institutes of Health grants U19 AI111276 and U01AI065663.

REFERENCES

1. Zak DE, Penn-Nicholson A, Scriba TJ, et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. The Lancet 2016;387(10035):2312-22.

2. World Health Organization. Guidelines on the management of latent tuberculosis infection. 2015.

3. Salgame P, Geadas C, Collins L, Jones-Lopez E, Ellner JJ. Latent tuberculosis infection--Revisiting and revising concepts. Tuberculosis (Edinb) 2015;95:373-84.

4. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MP. The immune response in tuberculosis. Annu Rev Immunol 2013;31:475-527.

5. Lobue P, Menzies D. Treatment of latent tuberculosis infection: An update. Respirology 2010;15:603-22.

6. Denholm JT, McBryde ES. The use of anti-tuberculosis therapy for latent TB infection. Infection and Drug Resistance 2010;3:63-72.

7. Berry MP, Graham CM, McNab FW, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 2010;466:973-7.

8. Jacobsen M, Repsilber D, Gutschmidt A, et al. Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis. J Mol Med 2007;85:613-21.

9. Kaforou M, Wright VJ, Oni T, et al. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. PLoS Med 2013;10:e1001538.

10. Sambarey A, Devaprasad A, Mohan A, et al. Unbiased Identification of Blood-based Biomarkers for Pulmonary Tuberculosis by Modeling and Mining Molecular Interaction Networks. EBioMedicine 2017;15:112-26.

11. Maertzdorf J, McEwen G, Weiner J, 3rd, et al. Concise gene signature for point-of-care classification of tuberculosis. EMBO Mol Med 2016;8:86-95.

12. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Respir Med 2016;4:213-24.

13. Sutherland JS, Loxton AG, Haks MC, et al. Differential gene expression of activating Fcgamma receptor classifies active tuberculosis regardless of human immunodeficiency virus status or ethnicity. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 2014;20:O230-8.

14. Roe JK, Thomas N, Gil E, et al. Blood transcriptomic diagnosis of pulmonary and extrapulmonary tuberculosis. JCI Insight 2016;1:e87238.

15. Petruccioli E, Scriba TJ, Petrone L, et al. Correlates of tuberculosis risk: predictive biomarkers for progression to active tuberculosis. Eur Respir J 2016;48:1751-63.

16. Sloot R, Schim van der Loeff MF, van Zwet EW, et al. Biomarkers Can Identify Pulmonary Tuberculosis in HIV-infected Drug Users Months Prior to Clinical Diagnosis. EBioMedicine 2015;2:172-9.

17. Levin M, Kaforou M. Predicting active tuberculosis progression by RNA analysis. The Lancet 2016; 4;387(10035):2268-2270.

18. Scriba TJ, Penn-Nicholson A, Shankar S, et al. Sequential inflammatory processes define human progression from M. tuberculosis infection to tuberculosis disease. PLoS Pathog 2017;13:e1006687.

19. Leong S, Zhao Y, Joseph NM, et al. Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India. Tuberculosis (Edinb) 2018;109:41-51.

20. Leong S, Zhao Y, Joseph NM, et al. Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India. Tuberculosis 2018;109:41-51.

21. Suliman S, Thompson E, Sutherland J, et al. Four-gene Pan-African Blood Signature Predicts Progression to Tuberculosis. Am J Respir Crit Care Med 2018;doi: 10.1164/rccm.201711-2340OC

22. Robertson BD, Altmann D, Barry C, et al. Detection and treatment of subclinical tuberculosis. Tuberculosis 2012;92:447-52.

23. Pai M, Behr MA, Dowdy D, et al. Tuberculosis. Nat Rev Dis Primers 2016;2:16076.

24. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 2010;26:392-8.

25. Moon H, Ahn H, Kodell RL, Baek S, Lin CJ, Chen JJ. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. Artificial intelligence in medicine 2007;41:197-207.

26. Morra M, Lu J, Poy F, et al. Structural basis for the interaction of the free SH2 domain EAT-2 with SLAM receptors in hematopoietic cells. Embo j 2001;20:5840-52.

27. Reich M, Spindler KD, Burret M, Kalbacher H, Boehm BO, Burster T. Cathepsin A is expressed in primary human antigen-presenting cells. Immunology letters 2010;128:143-7.

28. Ghilardi N, Li J, Hongo JA, Yi S, Gurney A, de Sauvage FJ. A novel type I cytokine receptor is expressed on monocytes, signals proliferation, and activates STAT-3 and STAT-5. J Biol Chem 2002;277:16831-6.

29. Dillon SR, Sprecher C, Hammond A, et al. Interleukin 31, a cytokine produced by activated T cells, induces dermatitis in mice. Nat Immunol 2004;5:752-60.

30. Lemberg MK, Bland FA, Weihofen A, Braud VM, Martoglio B. Intramembrane proteolysis of signal peptides: an essential step in the generation of HLA-E epitopes. J Immunol 2001;167:6441-6.

31. Joosten SA, van Meijgaarden KE, van Weeren PC, et al. Mycobacterium tuberculosis peptides presented by HLA-E molecules are targets for human CD8 T-cells with cytotoxic as well as regulatory activity. PLoS Pathog 2010;6:e1000782.

32. Harriff MJ, Wolfe LM, Swarbrick G, et al. HLA-E Presents Glycopeptides from the Mycobacterium tuberculosis Protein MPT32 to Human CD8(+) T cells. Sci Rep 2017;7:4622.

33. Sitkovsky M, Lukashev D. Regulation of immune cells by local-tissue oxygen tension: HIF1 alpha and adenosine receptors. Nat Rev Immunol 2005;5:712-21.

34. Ribeiro-Rodrigues R, Kim S, Coelho da Silva FD, et al. Discordance of tuberculin skin test and interferon gamma release assay in recently exposed household contacts of pulmonary TB cases in Brazil. PLoS One 2014;9:e96564.

35. Jones-Lopez EC, Kim S, Fregona G, et al. Importance of cough and M. tuberculosis strain type as risks for increased transmission within households. PLoS One 2014;9:e100984.

FIGURES

Figure 1- Expression of the PREDICT29 signature in Brazilian progressors and non-progressors. (A) Gene expression heatmap of the PREDICT29 signature. (B) Principal Component Analysis plot of the PREDICT29 signature. Progressors (n=16), non-progressors (n=21).

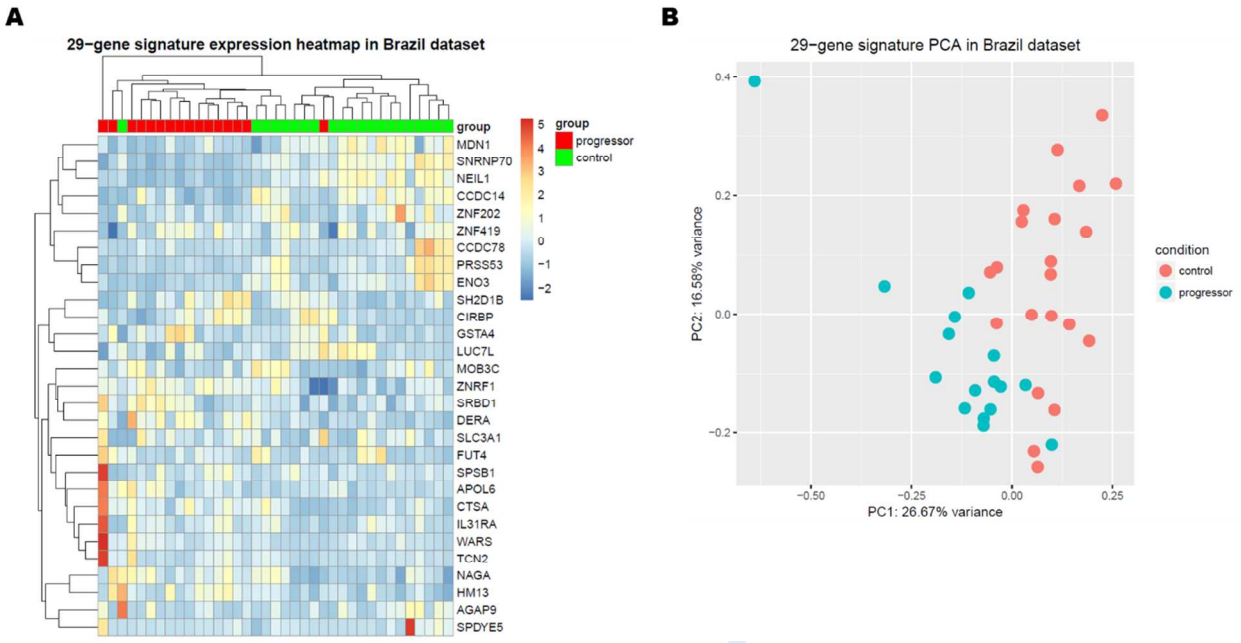
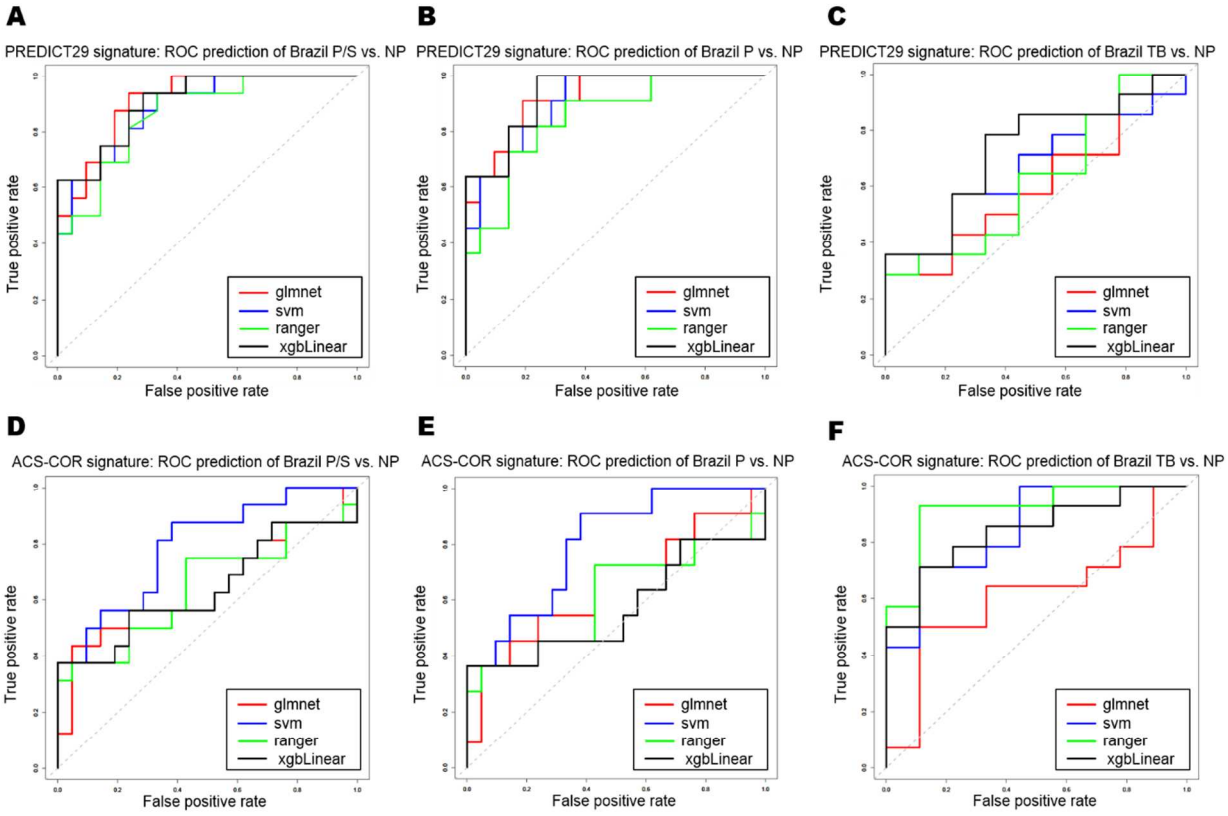


Figure 2- Receiver operating characteristic (ROC) curves for predicting clinical classification in Brazilian cohort using PREDICT29 and ACS-COR signature-based models. (A) PREDICT29 predictive performance of progressors (n=16) and non-progressors (n=21). (B) PREDICT29 predictive performance of progressors with subclinical removed (n=11) and non-progressors (n=21). (C) PREDICT29 predictive performance of active TB patients (n=14) versus non-progressors (n=21). (D) ACS-COR predictive performance of progressors (n=16) and non-progressors (n=21). (E) ACS-COR predictive performance of progressors with subclinical removed (n=11) and non-progressors (n=21). (F) ACS-COR predictive performance of active TB patients (n=14) versus non-progressors (n=21).



TABLES

Table 1- Predictive performances in Brazilian progressors versus non-progressors using PREDICT29 signature. (A) Includes progressors and subclinical (n=16) vs. non-progressors (n=21). (B) Includes progressors only (n=11) vs. non-progressors (n=21). Receiver operating characteristic (ROC) area-under-curve (AUC), sensitivity, and specificity reported as mean (95% confidence interval) for 50 iterations.

(A) PREDICT29 signature prediction on Brazil dataset: progressors/subclinical vs. non-progressors			
Model used	**AUC**	**Sensitivity**	**Specificity**
glmnet	0.903 (0.890, 0.916)	0.803 (0.771, 0.836)	0.801 (0.771, 0.832)
SVM	0.881 (0.865, 0.896)	0.720 (0.685, 0.755)	0.839 (0.804, 0.874)
ranger	0.864 (0.847, 0.880)	0.731 (0.696, 0.765)	0.814 (0.785, 0.843)
XGBoost	0.915 (0.900, 0.929)	0.785 (0.749, 0.821)	0.844 (0.808, 0.880)
AVERAGE	*0.890 (0.876, 0.905)*	*0.760 (0.725, 0.794)*	*0.825 (0.792, 0.857)*

(B) PREDICT29 signature prediction on Brazil dataset: progressors vs. non-progressors			
Model used	**AUC**	**Sensitivity**	**Specificity**
glmnet	0.929 (0.914, 0.944)	0.741 (0.703, 0.779)	0.867 (0.836, 0.897)
SVM	0.915 (0.900, 0.930)	0.756 (0.716, 0.796)	0.836 (0.807, 0.866)
ranger	0.867 (0.843, 0.891)	0.679 (0.641, 0.717)	0.830 (0.795, 0.866)
XGBoost	0.932 (0.919, 0.945)	0.794 (0.757, 0.830)	0.860 (0.828, 0.893)
AVERAGE	*0.911 (0.894, 0.928)*	*0.742 (0.704, 0.780)*	*0.848 (0.816, 0.880)*

Table 2- Predictive performances in Brazilian progressors versus non-progressors using ACS-COR signature. (A) Includes progressors and subclinical (n=16) vs. non-progressors (n=21). (B) Includes progressors only (n=11) vs. non-progressors (n=21). Receiver operating characteristic (ROC) area-under-curve (AUC), sensitivity, and specificity reported as mean (95% confidence interval) for 50 iterations.

(A) ACS-COR signature prediction on Brazil dataset: progressors/subclinical vs. non-progressors			
Model used	**AUC**	**Sensitivity**	**Specificity**
glmnet	0.714 (0.688, 0.739)	0.494 (0.447, 0.542)	0.864 (0.823, 0.905)
SVM	0.810 (0.787, 0.833)	0.710 (0.663, 0.757)	0.762 (0.719, 0.804)
ranger	0.727 (0.702, 0.752)	0.512 (0.458, 0.566)	0.867 (0.821, 0.914)
XGBoost	0.649 (0.622, 0.677)	0.506 (0.465, 0.547)	0.880 (0.853, 0.907)
AVERAGE	*0.725 (0.700, 0.750)*	*0.556 (0.508, 0.603)*	*0.843 (0.804, 0.882)*

(B) ACS-COR signature prediction on Brazil dataset: progressors vs. non-progressors			
Model used	**AUC**	**Sensitivity**	**Specificity**
glmnet	0.661 (0.628, 0.694)	0.431 (0.372, 0.490)	0.809 (0.760, 0.857)
SVM	0.792 (0.768, 0.817)	0.741 (0.698, 0.783)	0.687 (0.643, 0.730)
ranger	0.694 (0.666, 0.721)	0.561 (0.487, 0.634)	0.714 (0.652, 0.775)
XGBoost	0.533 (0.498, 0.568)	0.328 (0.282, 0.374)	0.886 (0.855, 0.917)
AVERAGE	*0.670 (0.640, 0.700)*	*0.515 (0.460, 0.571)*	*0.774 (0.728, 0.820)*

Table 3- Predictive performances in Brazilian active TB patients (n=14) versus non-progressors (n=21) using: (A) PREDICT29 signature, (B) ACS-COR signature. Receiver operating characteristic (ROC) area-under-curve (AUC), sensitivity, and specificity reported as mean (95% confidence interval) for 50 iterations.

(A) PREDICT29 signature prediction on Brazil dataset: TB vs. non-progressors			
Model used	**AUC**	**Sensitivity**	**Specificity**
glmnet	0.758 (0.734, 0.782)	0.637 (0.583, 0.692)	0.735 (0.679, 0.790)
SVM	0.764 (0.740, 0.787)	0.651 (0.612, 0.691)	0.810 (0.775, 0.844)
ranger	0.756 (0.731, 0.781)	0.672 (0.629, 0.716)	0.767 (0.734, 0.800)
XGBoost	0.752 (0.726, 0.778)	0.610 (0.566, 0.653)	0.781 (0.743, 0.819)
AVERAGE	*0.757 (0.732, 0.782)*	*0.643 (0.597, 0.688)*	*0.773 (0.733, 0.813)*

(B) ACS-COR signature prediction on Brazil dataset: TB vs. non-progressors			
Model used	**AUC**	**Sensitivity**	**Specificity**
glmnet	0.781 (0.759, 0.803)	0.608 (0.562, 0.655)	0.789 (0.732, 0.846)
SVM	0.910 (0.896, 0.923)	0.741 (0.706, 0.777)	0.846 (0.807, 0.885)
ranger	0.961 (0.954, 0.968)	0.780 (0.755, 0.806)	0.953 (0.935, 0.970)
XGBoost	0.917 (0.902, 0.932)	0.754 (0.722, 0.785)	0.970 (0.958, 0.982)
AVERAGE	*0.892 (0.878, 0.907)*	*0.721 (0.686, 0.755)*	*0.890 (0.858, 0.921)*