Tuberculosis 109 (2018) 41–51

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: www.elsevier.com/locate/tube

Diagnostics

Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India

Samantha Leong[a,1], Yue Zhao[b,1], Noyal M. Joseph[c], Natasha S. Hochberg[d,e], Sonali Sarkar[c], Jane Pleskunas[d], David Hom[d], Subitha Lakshminarayanan[c], C. Robert Horsburgh Jr.[e], Gautam Roy[c], Jerrold J. Ellner[d], W. Evan Johnson[b,**], Padmini Salgame[a,*]

[a] *Centre for Emerging Pathogens, Department of Medicine, Rutgers-New Jersey Medical School, Newark, NJ, USA*
[b] *Division of Computational Biomedicine, Boston University School of Medicine and Bioinformatics Program, Boston University, Boston, MA, USA*
[c] *Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India*
[d] *Boston Medical Centre, Boston, MA, USA*
[e] *Boston University, School of Public Health, Boston, MA, USA*

ARTICLE INFO

Keywords:
Tuberculosis
Biomarker
Gene signature
Diagnostic

ABSTRACT

Several studies have identified blood transcriptomic signatures that can distinguish active from latent Tuberculosis (TB). The purpose of this study was to assess how well these existing gene profiles classify TB disease in a South Indian population. RNA sequencing was performed on whole blood PAXgene samples collected from 28 TB patients and 16 latently TB infected (LTBI) subjects enrolled as part of an ongoing household contact study. Differential gene expression and clustering analyses were performed and compared with explicit predictive testing of TB and LTBI individuals based on established gene signatures. We observed strong predictive performance of TB disease states based on expression of known gene sets (ROC AUC 0.9007–0.9879). Together, our findings indicate that previously reported classifiers generated from different ethnic populations can accurately discriminate active TB from LTBI in South Indian patients. Future work should focus on converting existing gene signatures into a universal TB gene signature for diagnosis, monitoring TB treatment, and evaluating new drug regimens.

1. Introduction

Tuberculosis (TB), an infectious disease caused by *Mycobacterium tuberculosis* (Mtb), is a global health concern with 10.4 million new cases estimated in 2015 [1]. Untreated TB has a high mortality rate, estimated at a 70% 10-year case fatality rate in smear-positive pulmonary tuberculosis [2]. The WHO estimates that depending on context, 15–50% of TB cases are unreported or undiagnosed [1]. Further, the limitations of bacterial diagnosis in paucibacillary TB that is pediatric, extrapulmonary, or smear-negative pulmonary TB, often lead to empirical treatment. This approach exposes some individuals unnecessarily to side effects of the TB treatment while delaying effective therapy of the actual cause of the disease. Further, patients with TB may be denied appropriate treatment if bacteriologic testing is negative.

Currently, though nucleic acid amplification tests approach the sensitivity of Mtb culture [3], more sensitive diagnostics may need to be based on host biomarkers. These tests are inherently nonspecific, and it is unclear how they will perform in different populations. Several transcriptomic studies of TB cases and individuals infected with Mtb have been performed to characterize systemic gene expression (reviewed in Ref. [4]). Jacobsen et al. (2007) performed a microarray analysis and identified a minimal set of three genes in peripheral blood mononuclear cells that allowed distinction between TB disease and healthy individuals with latent TB infection (LTBI) [5]. The Berry et al. (2010) microarray studies demonstrated a 393-gene whole blood signature that discriminated subjects with active TB disease from those with LTBI, as well as an 86-gene set discriminating TB disease from other inflammatory and infectious diseases [6]. In this study, they also noted that this active TB signature was extinguished in patients following anti-TB treatment [6]. Kaforou et al. (2013) proposed a 27-gene whole blood signature that could distinguish TB and latent TB, regardless of HIV infection status [7]. Zak et al. (2016) recently identified

[*] Corresponding author. 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, USA.
[**] Corresponding author. Division of Computational Biomedicine and Bioinformatics Program, 72 E Concord Street, E645, Boston University, Boston, MA, USA.
E-mail address: padmini.salgame@rutgers.edu (P. Salgame).
[1] Equal contribution.

https://doi.org/10.1016/j.tube.2018.01.002
Received 2 August 2017; Received in revised form 9 January 2018; Accepted 15 January 2018

a 16-gene signature for predicting TB disease risk through sequencing analysis of whole blood PAXgene samples from a prospective cohort [8]. Given this recent proliferation of transcriptional studies in the field of TB, Sweeney et al. (2016) performed a meta-analysis of 14 publicly available TB transcriptomic datasets to identify a 3-gene set that they determined was robustly diagnostic for active TB disease [9].

Strikingly, although India accounts for more than one-quarter of the world's TB cases and deaths, transcriptomic studies in an Indian Mtb-infected populations have been limited [1,10–12]. TB is hyperendemic in India with an incidence of 217 per 100,000 and a mortality rate of 36 per 100,000 individuals as estimated in 2015 [1]. Furthermore, there is a high prevalence of other non-communicable confounding conditions and risk factors complicating TB disease cases in India, including diabetes, smoking, and alcohol consumption [13–15]. Presence of such comorbidities in patients with TB can impact their disease courses and treatment responses [16–19]. Thus, while two studies have proposed TB disease gene signatures (Maertzdorf et al. 4-gene signature and Sambarey et al. 10-gene signature) that were evaluated in Indian populations, additional studies are required [10,11]. Thus, the overall goal of this study was to assess how well the published gene signatures of active TB disease classified subjects in patients from South India.

2. Materials and methods

2.1. Study subjects and inclusion criteria

Subjects were recruited into the current cross-sectional sub-study from an ongoing observational household contact study being conducted at Jawarharlal Institute of Postgraduate Medical Education and Research (JIPMER). TB cases were recruited through the Revised National TB Control Programme (RNTCP) network of clinics in the city of Pondicherry (Puducherry union territory) and the districts of Villupuram and Cuddalore (Tamil Nadu state) in South India. Through the RNTCP, symptomatic individuals attending their Primary Health Centers were subsequently referred to their District Microscopy Center for sputum smear microscopy. Pulmonary TB (PTB) cases were eligible for enrollment if they were [1]: at least 6 years old [2], sputum Ziehl-Neelsen stain positive for acid-fast bacilli (AFB) (≥1+) [3], eventually culture-confirmed TB case [4] had no prior history of treatment for a previous TB episode [5], committed to complete TB therapy for the recommended duration [6], agreed to enroll in the directly observed therapy program for treatment, and [7] planned to reside in the study area for their treatment duration. Subjects were excluded if they [1]: refused HIV testing [2], received > 1 week (five daily or three intermittent) doses of anti-TB medication [3], had known multi-drug resistant (MDR) TB at diagnosis or were a household contact of an MDR case [4], were too sick to enroll with a Karnofsky score ≤10 (moribund), or [5] were HIV-infected. Household contacts of PTB cases with no prior history of TB were also recruited to the study if they were at least 6 years old and had significant contact with the PTB case for at least 3 months before the study, defined as sleeping under the same roof or sharing at least one meal per day or watching television (or equivalent) with the PTB case on average ≥ 5 days per week. Household contacts were assessed via tuberculin skin testing (TST) and sputum AFB and culture, and latent TB infected (LTBI) individuals were identified as having positive induration upon TST but were asymptomatic. All subjects were also evaluated via clinical questionnaires at enrollment to obtain demographic, clinical, and environmental information (Supplementary Table S1). Of note, only one individual (active TB case) in this study was under 18 years (age 16).

2.2. Ethics statement

The study was approved by the Boston University Medical Campus Institutional Review Board, JIPMER Institutional Review Board, and Rutgers Biomedical Health Sciences Institutional Review Board. Written informed consent and assent was obtained via forms that were translated into Tamil and that are in accordance with FDA regulations, the International Conference on Harmonization Good Clinical Practice guidelines, and local laws. The consent procedure was approved by IRBs from all participating institutions.

2.3. RNA sample processing and sequencing

At subject enrollment, 3 ml of peripheral blood was collected into PAXgene Blood RNA tubes (Cat #762165, BD Biosciences, San Jose, CA, USA), and frozen at −80 °C for storage until analysis. RNA was extracted from thawed samples from 28 active TB cases and 16 LTBI individuals using the PAXgene Blood RNA kit (Cat #762164, Qiagen, Hilden, Germany). Library preparation and sequencing was performed at the GenoTypic Technology Pvt. Ltd. Genomics facility in Bangalore, India, using the SureSelect Strand-Specific mRNA Library Prep kit (Cat #5190–6411, Agilent, Santa Clara, USA). Briefly, 1 µg of total RNA, quantified using the Qubit fluorometer (ThermoFisher Scientific, MA, USA), was used for mRNA enrichment by poly(A) selection. Enriched mRNA was fragmented using RNASeq Fragmentation Mix containing divalent cations at 94 °C for 4 min. Single-strand cDNA was synthesized in the presence of Actinomycin D (Gibco, Life Technologies, Carlsbad, CA, USA) and purified using HighPrep magnetic beads (Magbio Genomics Inc., USA). From this, double-stranded cDNA was synthesized, and ends were repaired before subsequent purification. The 3′-ends of cDNA were adenylated prior to ligation of Illumina Universal sequencing adaptors, which were then purified and amplified via 10 PCR cycles. Final cDNA sequencing libraries were purified and assessed for quantity using Qubit and fragment size distribution using Agilent TapeStation. Libraries were pooled in equimolar amounts, and resulting multiplexed library pools were sequenced using the Illumina NextSeq 500 for 75 bp single end reads.

2.4. Data processing and analysis

2.4.1. Data processing

Raw sequencing output files were assessed for data quality control using FastQC [20]. The average mean quality score (Phred score) at each bp position of the reads were above 30. Principal component analysis of the raw data revealed four outliers that could not be corrected by various methods of normalization or transformation and were subsequently removed prior to downstream analysis of the remaining 40 samples (24 TB, 16 LTBI). Rsubread [21], a highly efficient and accurate aligner for mapping RNA sequencing reads, was used to align reads to human genome hg19 and to determine expression counts for each gene. Previous work has established it as one of the most effective alignment tools for RNA sequencing data. On average, the number of aligned reads per sample was 30.9 M (range: 21.9–42.7 M), with an average alignment proportion of 60% (range: 50–70%). The 60% of reads that did align yielded a highly robust signal that was consistent across all samples. Gene expression counts were filtered to remove genes with max read counts less than or equal to 20 prior to differential expression analysis. Differentially expressed genes (DEGs) between the active TB and latent TB groups within the Indian dataset were identified using DESeq2 [22]. The default parameters of DESeq2 were used, with the model design incorporating both individuals' gender and TB condition as variables as follows: "design = ∼ condition + gender". Differential expression of TB over LTBI was determined via contrasting based on TB condition as follows: "contrast = c("condition", "TB", "LTBI")". This analysis produced 1200 DEGs using an adjusted p-value (FDR) cutoff of 0.00001. This list was used for development of an optimally performing gene signature specific to this dataset. The sequencing data have been deposited in Gene Expression Omnibus, and R markdown code for these and all subsequent analyses is available via GitHub at: https://github.com/jasonzhao0307/TB_Indian.

2.4.2. Evaluating published signatures

The following published gene signatures were evaluated in our dataset: Jacobsen 3-gene signature [5], Berry 86-gene and Berry 393-gene signatures [6], Kaforou 27-gene signature [7], Zak 16-gene signature [8], Sweeney 3-gene signature [9], Maertzdorf 4-gene signature [10], and Sambarey 10-gene signature [11]. The Jacobsen, Sweeney, Maertzdorf, Sambarey, and Zak gene signatures were used in their entirety. The gene identifiers provided by Berry et al. and Kaforou et al. in their published supplementary materials were first mapped to hg19 gene names, but some identifiers failed to map or resulted in redundancies. Thus, only the remaining unique genes (24 of the 27 genes in the Kaforou 27-gene signature, 65 genes of the Berry 86-gene signature and 264 genes of the Berry 393-gene signature) were used for downstream analysis. The presence of overlapping genes from these signatures in the differentially expressed gene (DEG) list was tested for independence using Fisher's exact test, and the expression of these described genes was visualized by heatmaps and hierarchical clustering analysis. We used analysis by t-distributed stochastic neighbor embedding (tSNE) [23] using the tsne package in R with perplexity = 10 and theta = 0.5 to assess for potential segregation of TB versus LTBI samples by qualitative visual inspection.

To quantitatively determine performance of these published gene sets in predictively classifying the TB and LTBI samples, a single run of leave-one-out cross-validation (LOOCV) with ridge logistic regression (using the glmnet R package with default parameters) was performed for the Indian dataset [24]. To evaluate the gene signatures alone, ridge logistic regression models, as opposed to the models associated with the published signatures, were used to quantitatively compare all signatures. Briefly, we aimed to develop a logistic regression model to predict probability of being a TB case or not based on expression values of the individual genes comprising each published signature. Using LOOCV, a logistic regression model was trained on all samples in the dataset excluding one random sample (training set), and then performance was tested on the one excluded sample (testing set). This cross-validation process was repeated iteratively to test all possible training data subsets and testing subsets in order to reduce variability.

Performance was evaluated by generating receiver operating characteristic (ROC) curves and computing area-under-curve (AUC) using the ROCR package in R [25]. Bootstrapping was used to obtain multiple AUCs and confidence intervals by sampling n samples with replacement, calculating AUC using LOOCV logistic regression, and repeating this process for 100 iterations. R Markdown code and reports for these analyses is available via GitHub: https://github.com/jasonzhao0307/TB_Indian.

2.4.3. GEO accession number

GSE101705.

3. Results

3.1. Differentially expressed genes in Indian subjects with TB versus LTBI overlap with existing TB signatures

The DEGs were identified between patients with active TB disease and asymptomatic individuals with latent TB infection (LTBI) in the Indian study. More than 1200 genes were differentially expressed between TB and LTBI individuals at a 0.00001 adjusted p-value (FDR) cutoff. For visualization and discussion purposes of a shorter DEG list, a more stringent adjusted p-value (FDR) cutoff of $< 10^{-11}$ was used resulting in a list of 76 DEGs (Fig. 1, Supplementary Table S2). Many DEGs identified from analysis of this patient group have previously been noted in published gene signatures describing TB infection stages, despite initial characterization in cohorts of individuals with differing ethnicity. The signature from Berry et al. (2010) (393 genes) [6] was included since it classified active TB in intermediate and high-burden settings and the signature was present in a subset with latent TB. The Berry 86-gene signature [6] discriminates active TB from other inflammatory and infectious diseases and was therefore evaluated in our study as a TB-specific disease signature. The Kaforou signature was derived from case-control studies conducted in South Africa and Malawi and was functional even in HIV-infected individuals [7]. We evaluated the Sweeney 3-gene signature [9] since it was derived from a meta-analysis, using a total of 14 datasets. Both the Sweeney and the Jacobsen signatures [5] provide minimal gene sets that separate active TB from healthy controls and latent tuberculosis. The Maertzdorf 4-gene and Sambarey 10-gene TB disease signatures, which were mined based on existing TB datasets and evaluated in Indian populations, were also tested here. Furthermore, the Sweeney 3-gene signature separated TB from other diseases and declined during treatment of patients with active TB [9]. In addition, the Zak 16-gene risk signature [8] is present in a subset of latently infected individuals who progress to disease and was therefore included in this evaluation.

The signatures from Berry et al. (2010) (393 genes), Berry et al. (2010) (86 genes), Kaforou et al. (2013) (27 genes), Sambarey et al. (2017), and Zak et al. (2016) (16 genes) all significantly overlapped with the 76 DEG list, with Fisher's exact test p-values less than 10^{-3}. Despite being defined from a London cohort, the Berry 393-gene set distinguishing TB disease from LTBI significantly overlapped with the 76 DEG list with a p-value of 1.56×10^{-13} and 15 genes in common (*LHFPL2, UBE2L6, BATF2, PSTPIP2, DHRS9, VAMP5, ANKRD22, FCGR1B, FCGR1A, CD274, IFITM1, PSME2, SERPING1, GBP1, GBP5*). The Berry 86-gene set also significantly overlapped with the 76 DEGs with a p-value of 2.518×10^{-7} and six overlapping genes (*NPC2, GLRX, POLB, DHRS9, ATP6V0E1, TYROBP*). Eight of the 76 DEGs (p-value = 6.386×10^{-16}) also overlapped with the Zak 16-gene signature identified from a South African cohort for risk of progressive TB infection (*ANKRD22, BATF2, FCGR1A, FCGR1B, GBP1, GBP5, SEPT4, SERPING1*). The Kaforou 27-gene signature, which was also defined from an African cohort and distinguished TB from LTBI regardless of HIV status, had seven overlapping genes with the 76 DEG list (p-value = 8.533×10^{-12}; *ANKRD22, FCGR1A, VAMP5, C1QC, FCGR1B, LHFPL2, FCGR1C*). The Sambarey 10-gene TB signature, derived from an Indian population, also significantly overlapped with the 76 DEG list with a p-value of 7.18×10^{-4} and two genes in common (*FCGR1A, RAB13*). The Maertzdorf 4-gene TB signature, which was also derived from an Indian population, shared one gene (*GBP1*) with the 76 DEG list and had significant overlap despite the small signature size (p-value of 0.016). The Sweeney et al. (2016) 3-gene signature also shared *GBP5* with the 76 DEG list. At varying degrees of p-value stringency for differential expression, genes described in these published signatures continue to be represented even at p-values as low as 10^{-20} (Supplementary Fig. S1). Thus, this warranted further analysis of established gene signatures to assess their application in this Indian study population.

3.2. TB and LTBI subjects segregate based on expression of published TB signature genes

Gene expression patterns of eight published signatures [5–11] were evaluated in our RNA sequencing dataset obtained from the Indian study. These were compared against a “best case scenario” using a 24-gene biomarker set developed from the 1200 DEG gene list using a cross-validation approach of Lasso logistic regression (Supplementary Table S3). To create an optimistic upper-bound for classification performance to compare existing biomarkers in this dataset, we generated a gene signature based on this Indian dataset itself. We used multiple lasso logistic regression (fixed alpha = 1) via glmnet and built a gene signature (24 genes in total) that separates TB from LTBI. The input features that were fed into lasso logistic regression were the 1200 DEGs identified using an adjusted p-value (FDR) cutoff of 0.00001. Then, we ran 100 iterations of lasso logistic regression, in which 10-fold cross-validation was applied in each run. After 100 runs, we defined *m* as the

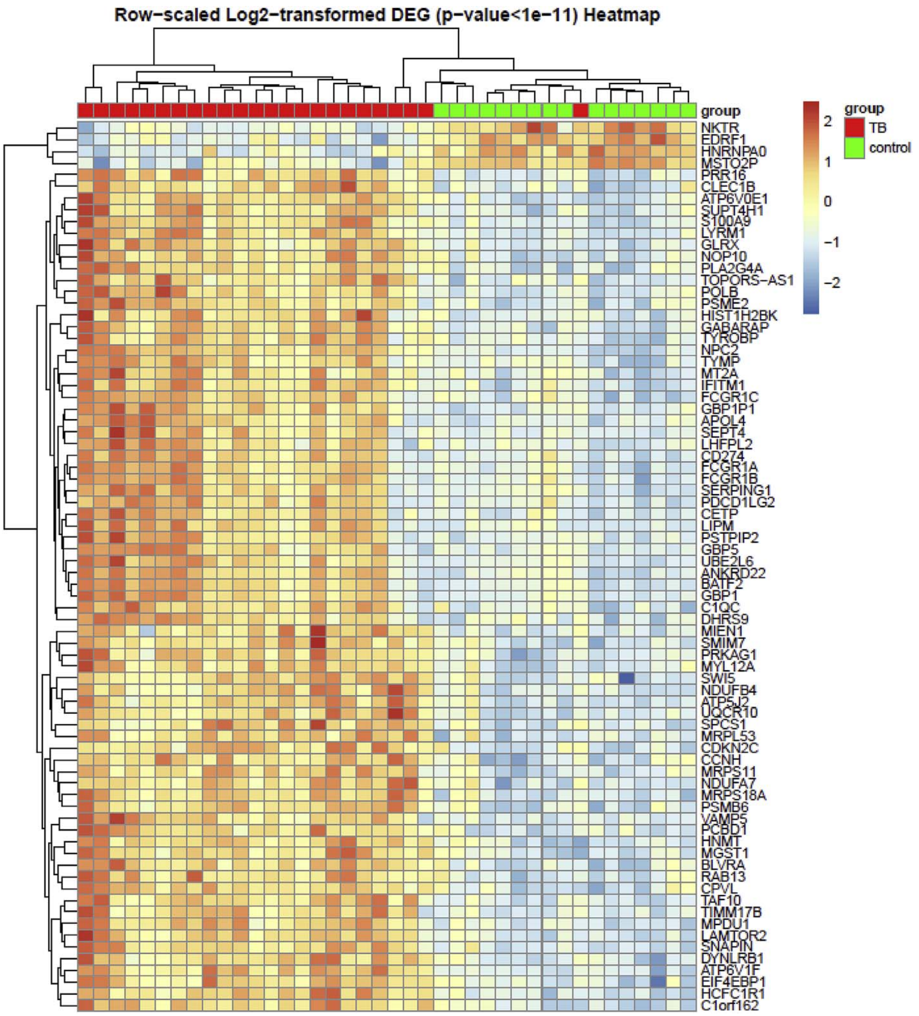

Fig. 1. Differentially expressed genes between active TB disease and LTBI subjects from a South Indian population. Gene expression heatmap of top 76 DEGs ($p < 10^{-11}$).

average signature length (number of genes in a signature) from the results of 100 runs of lasso logistic regression; then we ranked all genes based on the number of times the gene appeared in 100 runs of feature selection. Finally, we chose *m* top-ranking genes as the final gene signature, resulting in a set of 24 genes. The resulting 24-gene signature overlapped with genes in the published signatures including *BATF2* and *SERPING1* from the Zak 16-gene and Berry 393-gene signatures, as well as *LBH* from the Berry 393-gene signature and *WSB2* from the Berry 86-gene signature. Classification performance of this gene set was determined using the ridge logistic regression method. Thus, this method enabled us to generate an over-fitting reference signature that optimally performs in our Indian dataset, but not necessarily on other public datasets (Supplementary Table S4), for the comparison of the published signatures.

Heatmaps depicting expression of these gene signatures in patients with active TB disease and individuals with LTBI reflect similar patterns described in the respective originating publications (Figs. 2 and 10a). Furthermore, hierarchical clustering of subjects based on these selected genes demonstrate general clusters of individuals based on TB disease or LTBI status. Clustering analysis by tSNE was performed to identify possible segregation of samples based on gene expression levels of each of the gene signatures (Figs. 2–10b). Excluding the Berry 86-gene signature, which was described to distinguish TB from other infectious and inflammatory diseases, tSNE analysis of expression of these published gene sets revealed general segregation of subjects into two groups with one consisting of primarily TB disease patients and the other of LTBI individuals. Of the subjects that were misclassified by clustering analysis, one TB case consistently clustered amongst LTBI individuals in analysis of all of the signatures. Overall, this analysis suggests that existing published gene signatures can distinguish patients with TB versus LTBI in this Indian cohort.

3.3. Published signature genes accurately predict TB versus LTBI subjects in Indian patients

To quantitatively assess the ability of each of these published gene signatures to classify subjects as having either TB disease or LTBI, predictive classifiers were built based on the India RNA sequencing dataset using leave-out-one cross-validation with ridge logistic regression. This method produces an optimistic upper-bound for performance that results in over-fitting in this scenario where the number of variables (genes) exceeds number of participants. Receiver operating characteristic (ROC) curves of each classifier were generated, and bootstrapping was used to iteratively calculate area-under-curve (AUC) values using leave-one-out cross-validation to obtain mean AUCs and 95% confidence intervals (CI) for 100 repeats for each signature. Evaluation of AUCs provide a single quantitative measure reflective of overall predictive performance, considering both sensitivity and specificity parameters, without requiring preselection of probability cutoffs. The high mean AUC, sensitivity, and specificity values (all > 0.85) of all signatures suggest overall strong discriminatory ability between TB and LTBI (Fig. 11, Table 1). The genes from the Berry 393-gene signature performed the best of all eight known signatures tested, with the highest mean AUC (0.9879) and a lower 95% CI bound (0.9852) that

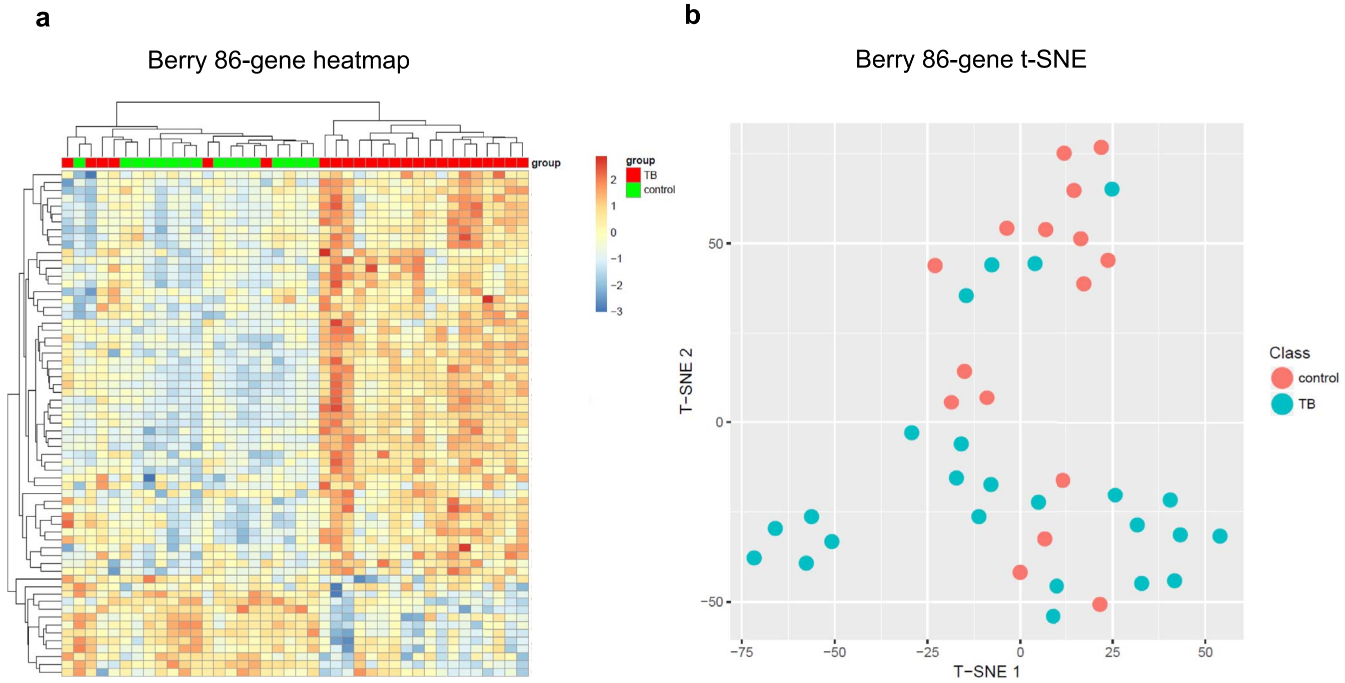

Fig. 2. Gene expression of Berry et al. (2010) 86-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

exceeded the upper bounds of the other seven known signatures. The classifier derived from the Berry 393-genes also nearly surpassed the performance of the upper-bound biomarker generated from the Indian dataset itself (mean AUC 0.9840, 95% CI [0.9802, 0.9877]). To determine if the Berry 393-gene signature's superior performance was due to its larger size, subsets of 16 genes, to match the size of the Zak et al. signature, from the 393-gene list were randomly sampled and tested, and AUC was calculated for each 16-gene set. This process was repeated 1000 times, and the mean AUC was determined to be 0.8919 with a 95% CI of [0.8885, 0.8952] (Supplementary Fig. S2). The decreased mean AUC value compared to 0.9802 obtained when using the complete 393-gene set suggests that use of a large number of genes in this case may contribute to its improved performance. Overall, these eight signatures seem to perform well in classification of TB and LTBI subjects within the dataset, despite the differing genetic backgrounds of subjects and the Mtb organisms that they were infected with.

4. Discussion

The results of this study suggest that gene expression of existing

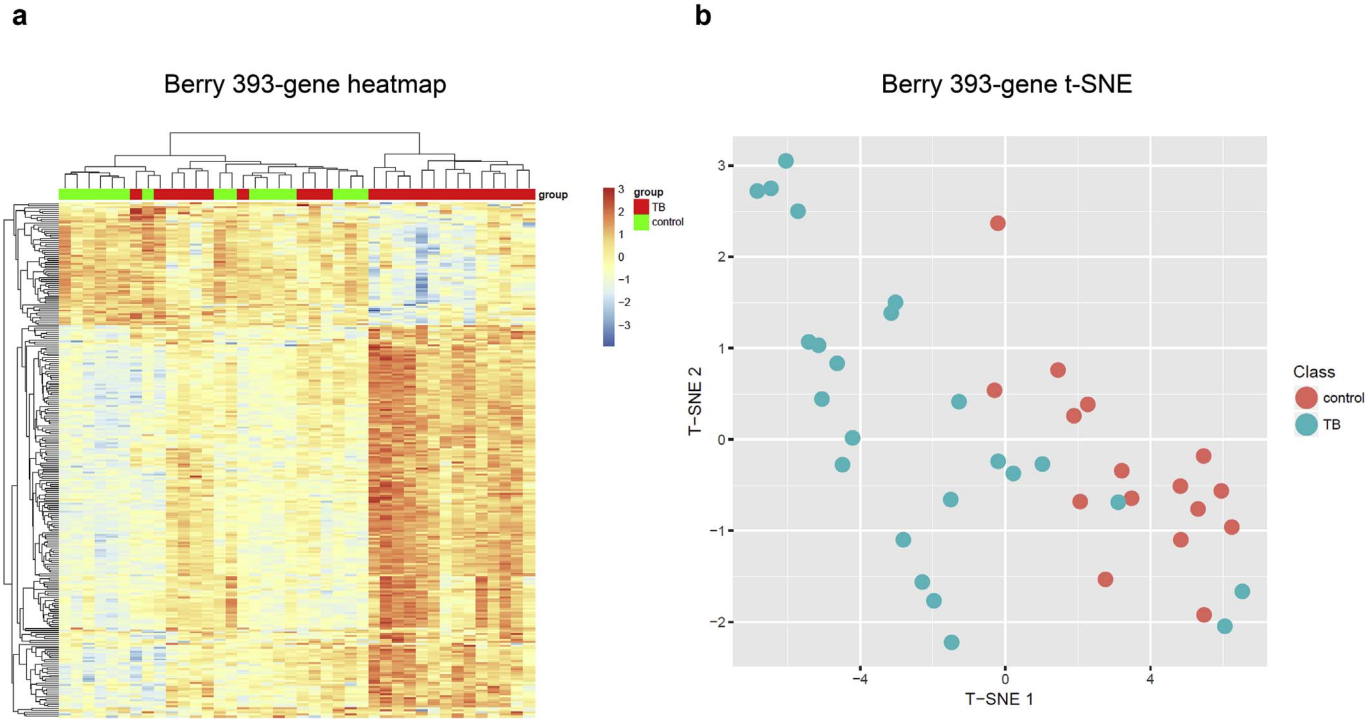

Fig. 3. Gene expression of Berry et al. (2010) 393-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

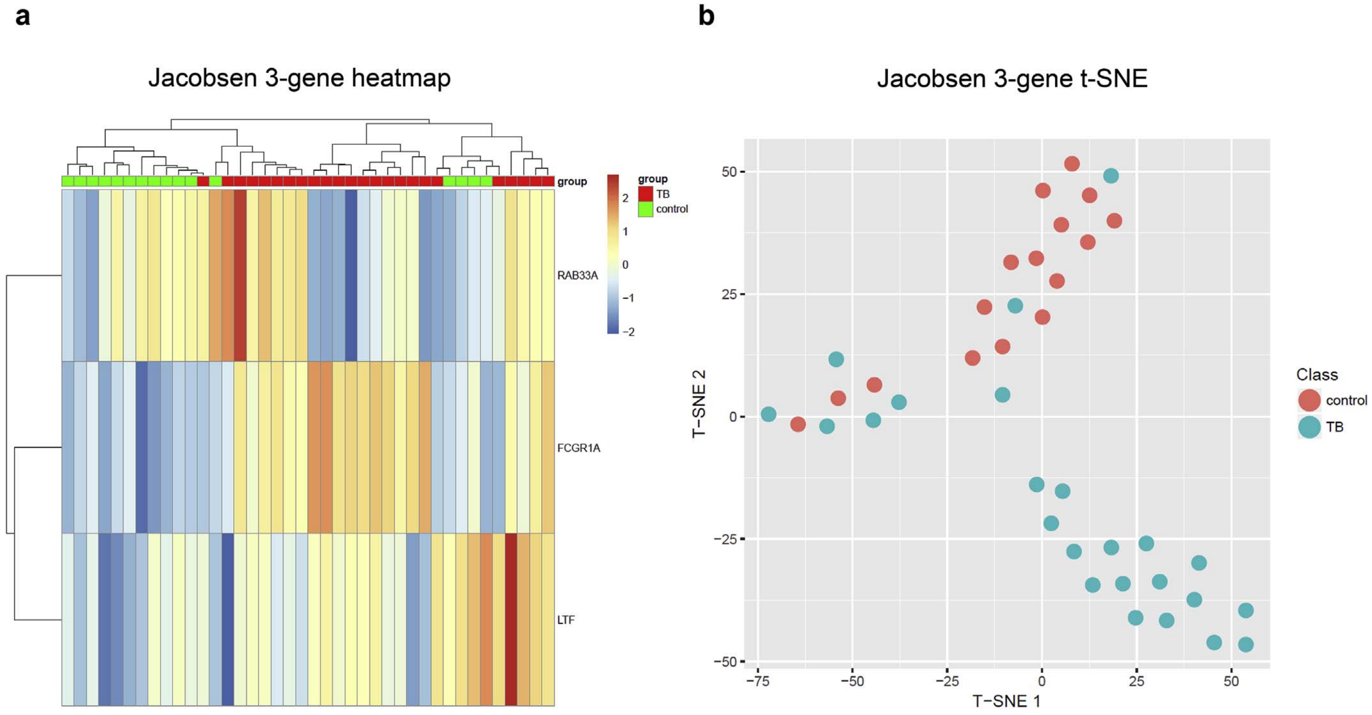

Fig. 4. Gene expression of Jacobsen et al. (2007) 3-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

transcriptional signatures can accurately classify TB and LTBI individuals within the South Indian subjects in our study. Distinct patterns of gene expression were observed in active TB patients versus LTBI individuals for each of the eight published gene sets evaluated. Whereas seven of the tested signatures were designed to characterize an active TB profile, the Zak et al. (2016) 16-gene signature was described as a predictor of risk of progression to TB disease [8]. Thus, it was surprising how accurately this signature performed in classifying TB versus LTBI individuals within our study group. The Zak 16-gene signature was obtained from a prospective cohort study, in which Mtb-infected individuals had blood samples collected for up to two years and were later classified as progressors to active TB disease or controls who remained healthy. In deriving their risk signature, they first identified candidate genes comparing gene expression of controls to that of progressors only at the most proximal time point to disease diagnosis. Thus, it is likely that their resulting signature is more reflective of subclinical or active TB disease than an early predictor of host progression risk, which would explain its strong performance in TB versus LTBI classification.

Of note, the eight published TB signatures were derived using either

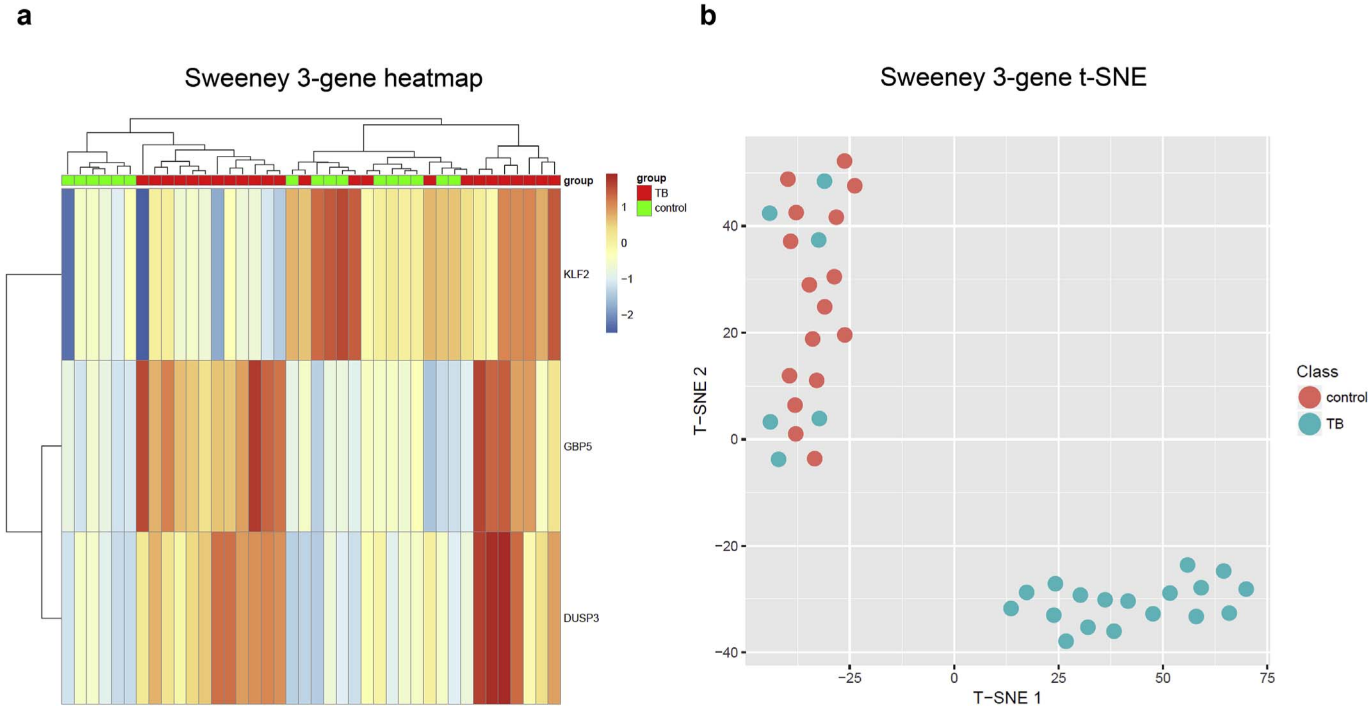

Fig. 5. Gene expression of Sweeney et al. (2016) 3-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

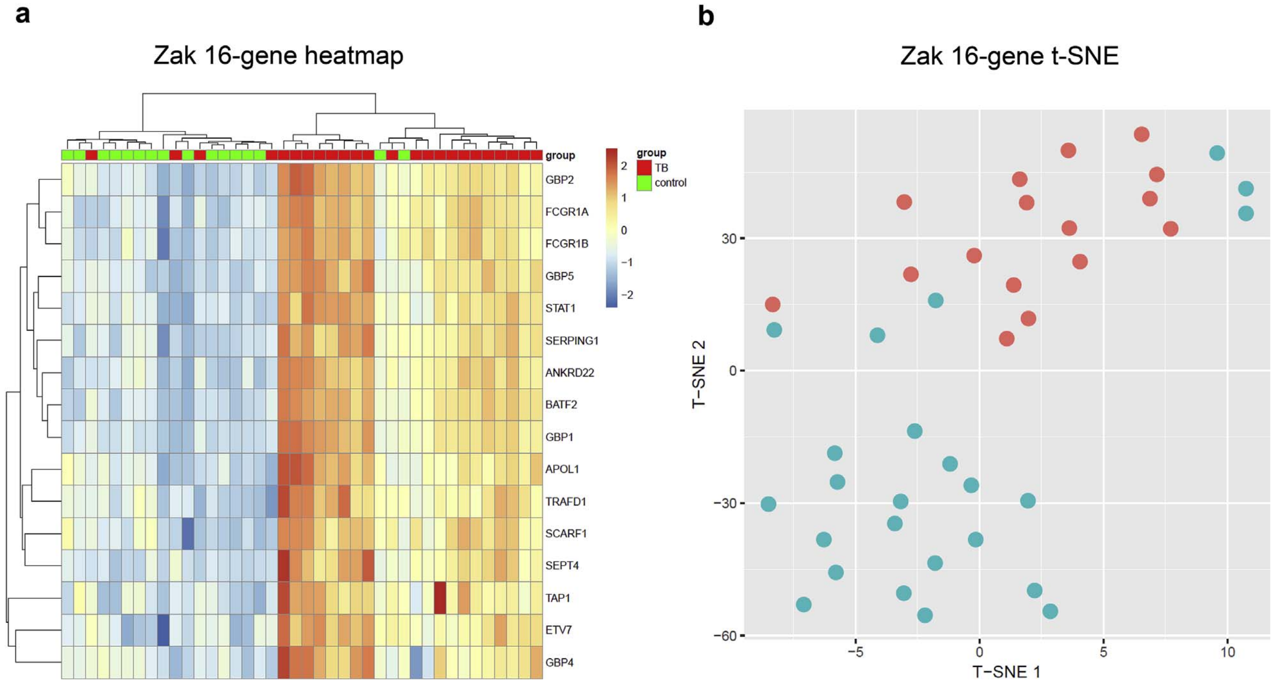

Fig. 6. Gene expression of Zak et al. (2016) 16-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

RNA-seq or microarray technology. RNA-seq technology offers a number of advantages compared to microarrays, including superior ability to detect low abundance transcripts, minimal technical issues and a better dynamic range that allows for detection of more differentially expressed genes. Head-to-head comparison of gene expression profiles from RNA-seq and Affymetrix microarray platforms show high correlation of the data obtained by the two technologies [26] and both perform similarly in gene expression-based predictive models [27]. The ability to generate highly correlative data may account for why both microarray and RNA-seq-derived TB signatures performed well in classifying TB and LTBI.

Clustering analyses based on gene expression of these published sets revealed that most subjects segregated along their TB and LTBI classifications except for a few individuals for each signature. Whereas most of these misclassified subjects were random individuals for each signature tested, one active TB case consistently clustered with LTBI subjects. In the Berry study, 25% of latently infected from the UK cohort and 10% in the South African participants with latent TB expressed the TB transcriptional signature indicating subclinical disease. In contrast to the Berry study, we did not see a TB transcriptional profile in the 16 latently-infected individuals included in our study. One reason for this difference could be that the latently infected in the Berry study were recruited from TB or TB/HIV clinics while the latent TB participants in our study were household contacts of index TB cases. Thus, the

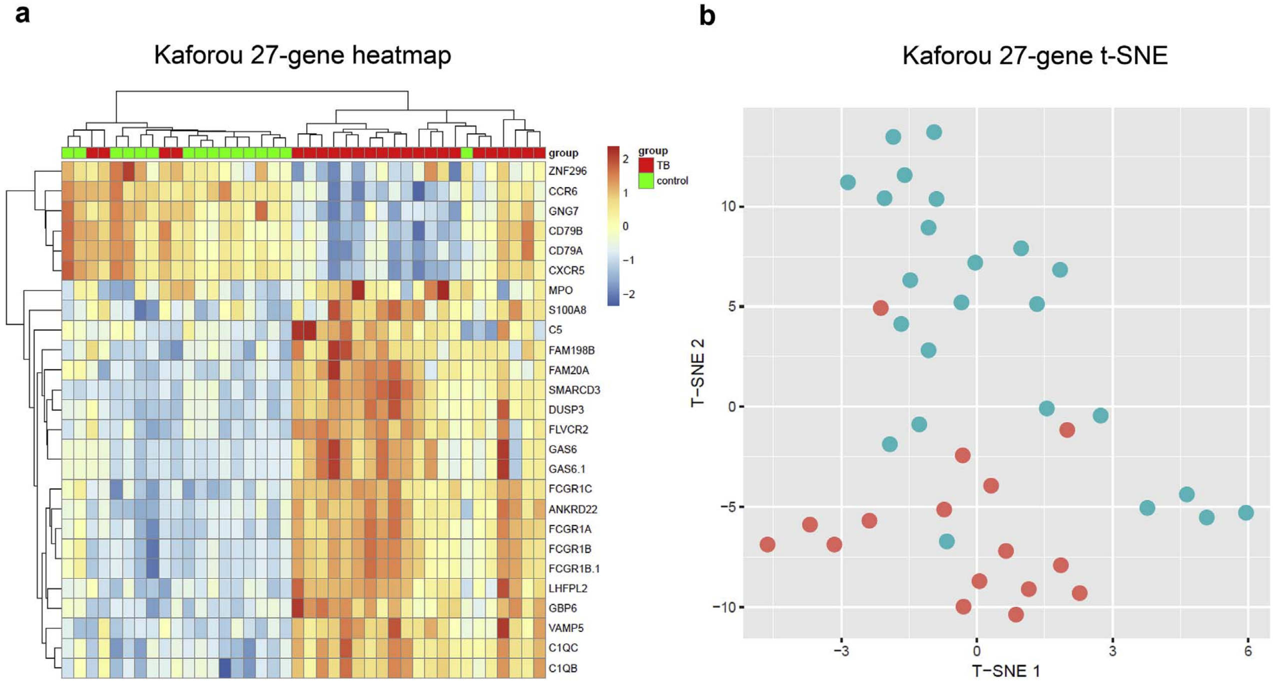

Fig. 7. Gene expression of Kaforou et al. (2013) 27-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

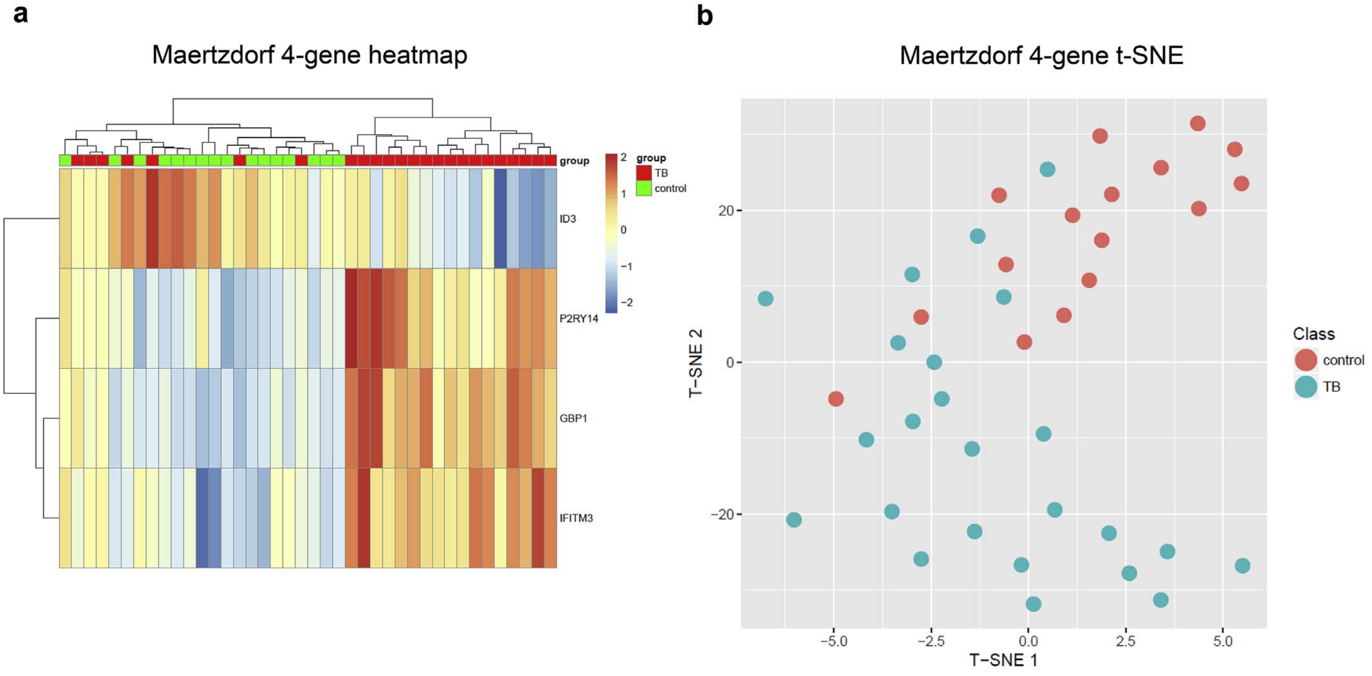

Fig. 8. Gene expression of Maertzdorf et al. (2016) 4-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

exposure to Mtb in the latently-infected household contacts is likely recent and therefore they may not have progressed to developing subclinical disease.

Two biomarker studies performed in Indian cohorts report different non-overlapping gene signatures for classifying TB. In one study, 360 top genes were selected from two microarray datasets generated in cohorts from South Africa and Gambia [28,29] and a targeted RT–PCR array was designed and then tested in an Indian cohort. This study found that a 4-gene set (*GBP1, ID3, P2RY14* and *IFITM3*) was able to distinguish TB patients from healthy individuals [10]. In the second study, the authors identified a 10-gene set to distinguish TB (*FCGR1A, HK3, RAB13, RBBP8, IFI44L, TIMM10, BCL6, SMARCD3, CYP4F3* and *SLPI*) by modeling and mining interaction networks [11]. This suggests that the candidate biomarkers discovered by the various studies can be combined to develop a highly sensitive and specific point-of-care diagnostic test for TB.

Several genes discriminative for TB disease appear to be consistently represented across multiple signatures as well as in differential gene expression analysis of TB and LTBI subjects within our cohort, even at high degrees of stringencies in testing for differential expression. Despite the small sample size, nonetheless there was a substantial proportion of DEGs identified in our dataset that were not contained

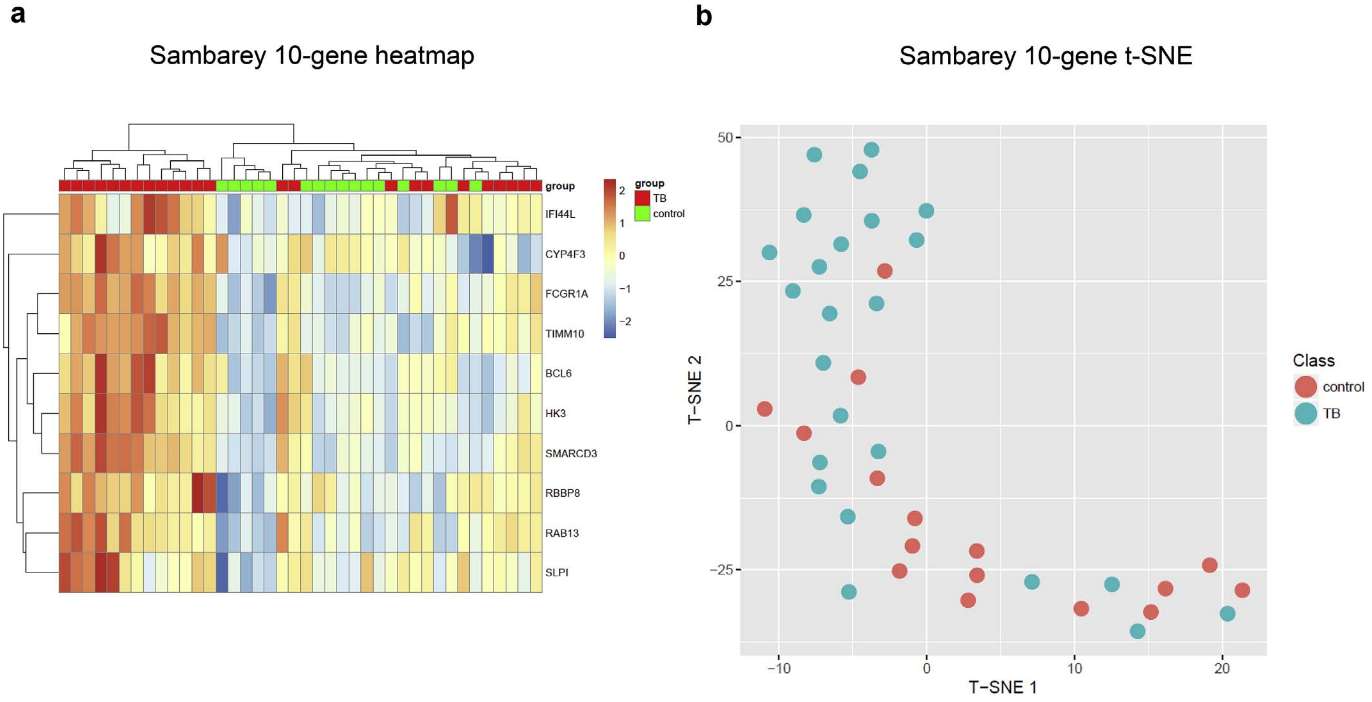

Fig. 9. Gene expression of Sambarey et al. (2017) 10-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

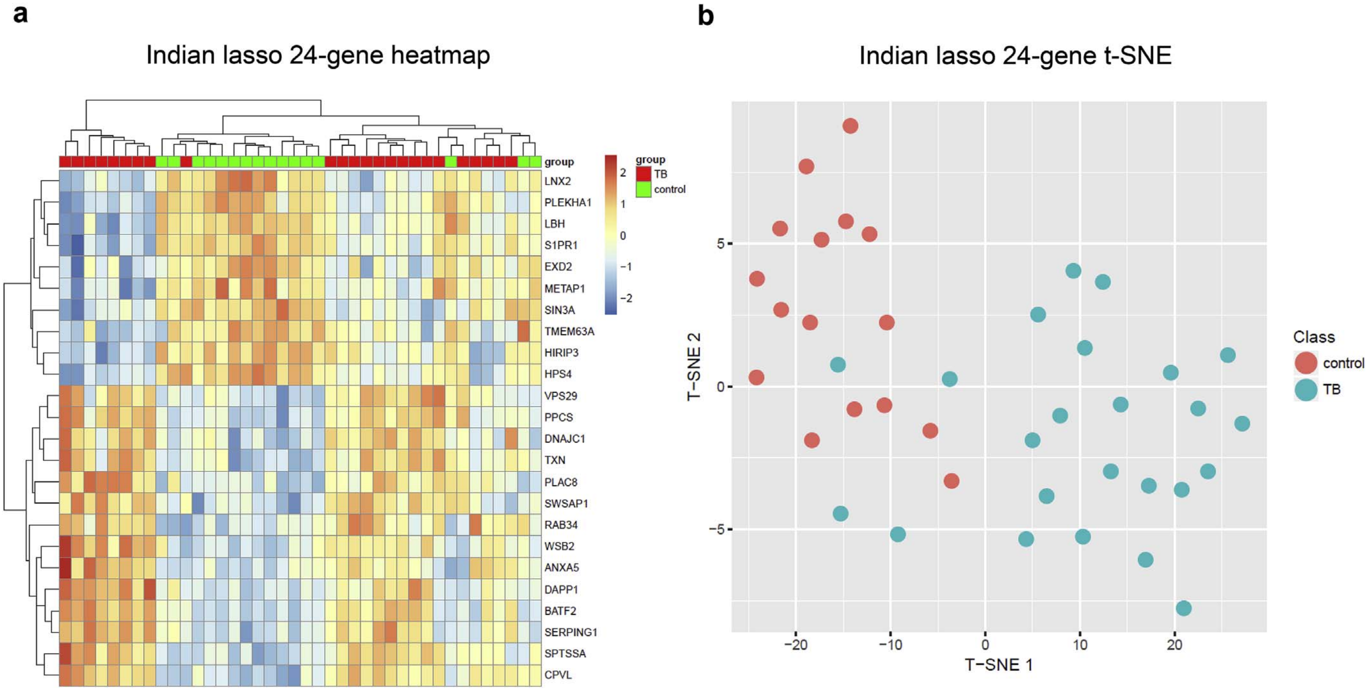

Fig. 10. Gene expression of Indian lasso method 24-gene signature in TB and LTBI subjects from a South Indian population. (a) heatmap, and (b) tSNE clustering analysis.

within the existing signatures. Future larger studies should evaluate whether the high presence of comorbidities such as smoking, alcohol usage or abuse, malnutrition, and diabetes affect DEGs between TB and latent infection. Analysis of population attributable fractions revealed that compared to the general population of the surrounding region, TB cases in our study were more likely to use alcohol if male and to be malnourished (Hochberg et al., unpublished data). Thus, future studies should determine if blood transcriptional profiles can further segregate TB disease based on these comorbidities. Our study suggests that a single diagnostic based on gene expression signatures may be accurate

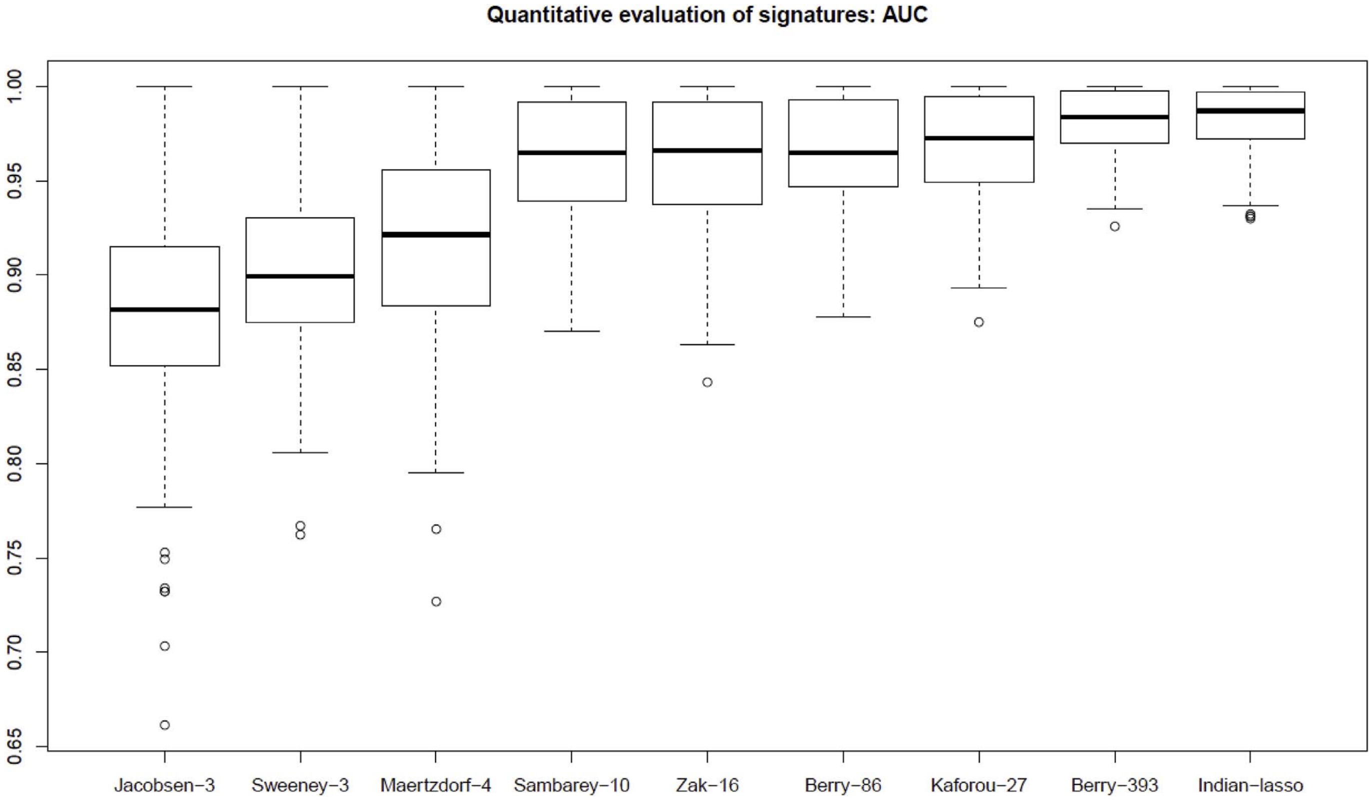

Fig. 11. Area-under-curve values obtained from receiver operating characteristic analyses in classification of TB versus LTBI. ROC curves of classifiers built from published signatures: Jacobsen et al. (2007) 3-gene, Sweeney et al. (2016) 3-gene, Maertzdorf et al. (2016) 4-gene, Sambarey et al. (2017) 10-gene, Zak et al. (2016) 16-gene, Berry et al. (2010) 86-gene, Kaforou et al. (2013) 27-gene, Berry et al. (2010) 393-gene, and 24-gene biomarker derived directly from the India dataset. Plots depict mean area under the curve (AUC) with 95% confidence intervals.

Table 1
Overall predictive performance of published signatures in classification of TB versus LTBI.

	AUC		Sensitivity		Specificity	
	Mean (95% CI)	Std. error	Mean (95% CI)	Std. error	Mean (95% CI)	Std. error
Zak-16	0.9628 (0.9572, 0.9684)	0.0028	0.9393 (0.9242, 0.9545)	0.0076	0.9286 (0.9185, 0.9388)	0.0051
Jacobsen-3	0.9007 (0.8915, 0.9100)	0.0047	0.8480 (0.8247, 0.8713)	0.0118	0.8476 (0.8319, 0.8634)	0.0079
Sweeney-3	0.9012 (0.8906, 0.9117)	0.0053	0.8607 (0.8415, 0.8800)	0.0097	0.8407 (0.8245, 0.8568)	0.0081
Berry-393	0.9879 (0.9852, 0.9905)	0.0014	0.9263 (0.9102, 0.9423)	0.0081	0.9481 (0.9394, 0.9569)	0.0044
Berry-86	0.9717 (0.9664, 0.9771)	0.0027	0.9394 (0.9259, 0.9530)	0.0068	0.9356 (0.9272, 0.9439)	0.0042
Kaforou-27	0.9613 (0.9550, 0.9676)	0.0032	0.9399 (0.9282, 0.9516)	0.0059	0.9326 (0.9240, 0.9412)	0.0043
Maertzdorf-4	0.9053 (0.8940, 0.9165)	0.0057	0.8700 (0.8513, 0.8887)	0.0094	0.8890 (0.8736, 0.9043)	0.0077
Sambarey-10	0.9578 (0.9504, 0.9652)	0.0037	0.8967 (0.8779, 0.9155)	0.0095	0.9427 (0.9330, 0.9524)	0.0049
Indian-lasso-24	0.9840 (0.9802, 0.9877)	0.0019	0.9307 (0.9160, 0.9454)	0.0074	0.9450 (0.9422, 0.9577)	0.0039

across diverse populations.

Funding

This publication is based on work supported by Award No. USB1-31150-XX-13 of the U.S. Civilian Research & Development Foundation (CRDF Global), Department of Biotechnology BT/PR8941/MED/15/108/2013 and by the National Science Foundation under Cooperative Agreement No. OISE-9531011.

Conflicts of interest

None.

Author contributions

PS, JJE, NSH, SS and CRH contributed to cohort design; SL and YZ performed the experiments and analyzed the data; PS and WEJ designed experiment and interpreted the data; NJ, SS, JP, NSH, DH, SL, GR contributed to subject recruitment and sample collection, storage and shipping, and data management; SL, YZ, PS and WEJ wrote the manuscript; PS, JJE and WEJ contributed to discussion and conclusions of the study.

Acknowledgements

We thank all of the participants who made the study possible. We also acknowledge the dedication and hard work of the field staff and study team. We thank Rachel Kubiak and Astrid Loomans for help with data management, and Laura White for help with statistics. This work was performed under the auspices of the RePORT India Program. We wish to acknowledge the support of Sudha Srinivasan and Peter Kim from DAIDS/NIAID/NIH and Jyoti M. Logani from Department of Biotechnology, Ministry of Science and Technology, Government of India.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.tube.2018.01.002.

References

[1] World Health Organization. reportGlobal tuberculosis report 2016.

[2] Tiemersma EW, van der Werf MJ, Borgdorff MW, Williams BG, Nagelkerke NJD. Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review. PLoS One 2011;6(4):e17601http://dx.doi.org/10.1371/journal.pone.0017601.

[3] Lawn SD, Mwaba P, Bates M, Piatek A, Alexander H, Marais BJ, Cuevas LE, McHugh TD, Zijenah L, Kapata N, Abubakar I, McNerney R, Hoelscher M, Memish ZA, Migliori GB, Kim P, Maeurer M, Schito M, Zumla A. Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. Lancet Infect Dis 2013;13(4):349–61. http://dx.doi.org/10.1016/S1473-3099(13)70008-2.

[4] Deffur A, Wilkinson RJ, Coussens AK. Tricks to translating TB transcriptomics. Ann Transl Med 2015;3(Suppl 1):S43. http://dx.doi.org/10.3978/j.issn.2305-5839.2015.04.12. PubMed PMID: 26046091; PMCID: PMC4437947.

[5] Jacobsen M, Repsilber D, Gutschmidt A, Neher A, Feldmann K, Mollenkopf HJ, Ziegler A, Kaufmann SH. Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis. J Mol Med (Berl) 2007;85(6):613–21. http://dx.doi.org/10.1007/s00109-007-0157-6. PubMed PMID: 17318616.

[6] Berry MP, Graham CM, McNab FW, Xu Z, Bloch SA, Oni T, Wilkinson KA, Banchereau R, Skinner J, Wilkinson RJ, Quinn C, Blankenship D, Dhawan R, Cush JJ, Mejias A, Ramilo O, Kon OM, Pascual V, Banchereau J, Chaussabel D, O'Garra A. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 2010;466(7309):973–7. http://dx.doi.org/10.1038/nature09247. PubMed PMID: 20725040; PMCID: PMC3492754.

[7] Kaforou M, Wright VJ, Oni T, French N, Anderson ST, Bangani N, Banwell CM, Brent AJ, Crampin AC, Dockrell HM, Eley B, Heyderman RS, Hibberd ML, Kern F, Langford PR, Ling L, Mendelson M, Ottenhoff TH, Zgambo F, Wilkinson RJ, Coin LJ, Levin M. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. PLoS Med 2013;10(10):e1001538. http://dx.doi.org/10.1371/journal.pmed.1001538. PubMed PMID: 24167453; PMCID: PMC3805485.

[8] Zak DE, Penn-Nicholson A, Scriba TJ, Thompson E, Suliman S, Amon LM, Mahomed H, Erasmus M, Whatney W, Hussey GD, Abrahams D, Kafaar F, Hawkridge T, Verver S, Hughes EJ, Ota M, Sutherland J, Howe R, Dockrell HM, Boom WH, Thiel B, Ottenhoff THM, Mayanja-Kizza H, Crampin AC, Downing K, Hatherill M, Valvo J, Shankar S, Parida SK, Kaufmann SHE, Walzl G, Aderem A, Hanekom WA. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. Lancet 2016. http://dx.doi.org/10.1016/s0140-6736(15)01316-1.

[9] Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Resp Med. 2016;4(3):213–24. http://dx.doi.org/10.1016/s2213-2600(16)00048-5. PubMed PMID: 26907218; PMCID: PMC4838193.

[10] Maertzdorf J, McEwen G, Weiner 3rd J, Tian S, Lader E, Schriek U, Mayanja-Kizza H, Ota M, Kenneth J, Kaufmann SH. Concise gene signature for point-of-care classification of tuberculosis. EMBO Mol Med 2016;8(2):86–95. http://dx.doi.org/10.15252/emmm.201505790. PubMed PMID: 26682570; PMCID: PMC4734838.

[11] Sambarey A, Devaprasad A, Mohan A, Ahmed A, Nayak S, Swaminathan S, D'Souza G, Jesuraj A, Dhar C, Babu S, Vyakarnam A, Chandra N. Unbiased identification of blood-based biomarkers for pulmonary tuberculosis by modeling and mining molecular interaction networks. EBioMed 2017;15:112–26. http://dx.doi.org/10.1016/j.ebiom.2016.12.009. PubMed PMID: 28065665; PMCID: PMC5233809.

[12] Prada-Medina CA, Fukutani KF, Pavan Kumar N, Gil-Santana L, Babu S, Lichtenstein F, West K, Sivakumar S, Menon PA, Viswanathan V, Andrade BB, Nakaya HI, Kornfeld H. Systems immunology of diabetes-tuberculosis comorbidity reveals signatures of disease complications. Sci Rep 2017;7(1):1999. http://dx.doi.org/10.1038/s41598-017-01767-4. PubMed PMID: 28515464; PMCID: PMC5435727.

[13] Marak B, Kaur P, Rao SR, Selvaraju S. Non-communicable disease comorbidities and risk factors among tuberculosis patients, Meghalaya, India. Indian J Tubercul 2016;63(2):123–5. http://dx.doi.org/10.1016/j.ijtb.2015.07.018. PubMed PMID: 27451823.

[14] Gajalakshmi V, Peto R. Smoking, drinking and incident tuberculosis in rural India: population-based case-control study. Int J Epidemiol 2009;38(4):1018–25. http://dx.doi.org/10.1093/ije/dyp225. PubMed PMID: 19498083.

[15] Kumar A, Jain D, Devesh G, Satyanarayana S, Kumar AM, Chadha SS, Wilson N, Nagaraja SB, Shah AN, Naik B, Yoele RD, Syed IFIA, Achanta S, Nair SA, Sharma SK, Soneja M, Krishnappa D, Prakash BC, Ravish KS, Ranganath TS, Chauhan MC, Dave PV, Narayanaswamy MV, Suryakanth MD, Bhist A, Sinha UC, Dayal R, Tekumalla RR, Nair S, Kumari AK, Subramonianpillai J, Jali MV, Mahishale VK, Hiremath MB, Khanna A, Lohiya S, Chaudhry A, Shaw N, Varadarajan S, Arthur P, Kapur A, Lonnroth K, Zachariah R, Harries AD. Screening of patients with tuberculosis for diabetes mellitus in India. Trop Med Int Health TM & IH 2013;18(5):636–45. http://dx.doi.org/10.1111/tmi.12084. PubMed PMID: 23458555.

[16] Lonnroth K, Roglic G, Harries AD. Improving tuberculosis prevention and care through addressing the global diabetes epidemic: from evidence to policy and practice. Lancet Diabet Endocrinol 2014;2(9):730–9. http://dx.doi.org/10.1016/

s2213-8587(14)70109-3. PubMed PMID: 25194886.

[17] World Health Organization IUATaLD. A WHO/the Union monograph on TB and tobacco control: joining efforts to control two related global epidemics. 2007.

[18] Lonnroth K, Williams BG, Stadlin S, Jaramillo E, Dye C. Alcohol use as a risk factor for tuberculosis - a systematic review. BMC Publ Health 2008;8:289. http://dx.doi.org/10.1186/1471-2458-8-289. PubMed PMID: 18702821; PMCID: PMC2533327.

[19] Rehm J, Samokhvalov AV, Neuman MG, Room R, Parry C, Lonnroth K, Patra J, Poznyak V, Popova S. The association between alcohol use, alcohol use disorders and tuberculosis (TB). A systematic review. BMC Publ Health 2009;9:450. http://dx.doi.org/10.1186/1471-2458-9-450. PubMed PMID: 19961618; PMCID: PMC2796667.

[20] Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010 Available online at: http://wwwbioinformaticsbabrahamacuk/projects/fastqc/.

[21] Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 2013;41(10):e108. http://dx.doi.org/10.1093/nar/gkt214. PubMed PMID: 23558742; PMCID: PMC3664803.

[22] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550. http://dx.doi.org/10.1186/s13059-014-0550-8. PubMed PMID: 25516281; PMCID: PMC4302049.

[23] van der Maaten GEH LJP. Visualizing high-dimensional data using t-SNE. J Mach Learn Res 2008;9(Nov):2579–605.

[24] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Software 2010;33(1):1–22. Epub 2010/09/03. PubMed PMID: 20808728; PMCID: Pmc2929880.

[25] Sing T, Sander O, Beerenwinkel N, Lengauer TROCR. Visualizing classifier performance in R. Bioinformatics 2005;21(20):3940–1. http://dx.doi.org/10.1093/bioinformatics/bti623.

[26] Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 2014;9(1):e78644. http://dx.doi.org/10.1371/journal.pone.0078644. PubMed PMID: 24454679; PMCID: PMC3894192.

[27] Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, Thierry-Mieg D, Wang J, Furlanello C, Devanarayan V, Cheng J, Deng Y, Hero B, Hong H, Jia M, Li L, Lin SM, Nikolsky Y, Oberthuer A, Qing T, Su Z, Volland R, Wang C, Wang MD, Ai J, Albanese D, Asgharzadeh S, Avigad S, Bao W, Bessarabova M, Brilliant MH, Brors B, Chierici M, Chu TM, Zhang J, Grundy RG, He MM, Hebbring S, Kaufman HL, Lababidi S, Lancashire LJ, Li Y, Lu XX, Luo H, Ma X, Ning B, Noguera R, Peifer M, Phan JH, Roels F, Rosswog C, Shao S, Shen J, Theissen J, Tonini GP, Vandesompele J, Wu PY, Xiao W, Xu J, Xu W, Xuan J, Yang Y, Ye Z, Dong Z, Zhang KK, Yin Y, Zhao C, Zheng Y, Wolfinger RD, Shi T, Malkas LH, Berthold F, Wang J, Tong W, Shi L, Peng Z, Fischer M. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol 2015;16:133. http://dx.doi.org/10.1186/s13059-015-0694-1. PubMed PMID: 26109056; PMCID: PMC4506430.

[28] Maertzdorf J, Ota M, Repsilber D, Mollenkopf HJ, Weiner J, Hill PC, Kaufmann SH. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. PLoS One 2011;6(10):e26938http://dx.doi.org/10.1371/journal.pone.0026938. PubMed PMID: 22046420; PMCID: PMC3203931.

[29] Maertzdorf J, Repsilber D, Parida SK, Stanley K, Roberts T, Black G, Walzl G, Kaufmann SH. Human gene expression profiles of susceptibility and resistance in tuberculosis. Gene Immun 2011;12(1):15–22. http://dx.doi.org/10.1038/gene.2010.51. PubMed PMID: 20861863.