LncRNA and predictive model for improving the diagnosis of clinically diagnosed pulmonary tuberculosis

Running title: lncRNA-based nomogram to assist tuberculosis diagnosis

Xuejiao Hu[1,2,3,†], Jingwei Zhang[1†], Shun Liao[3,4,†], Hao Bai[1], Shubham Gupta[2,3], Yi Zhou[1], Juan Zhou[1], Qian Wu[1], Zirui Meng[1], Lin Jiao[1], Lijuan Wu[1], Minjin Wang[1], Xuerong Chen[5], Yanhong Zhou[1], Xiaojun Lu[1], Zhaolei Zhang[2,3,4*] and Binwu Ying[1*]

[1] Department of Laboratory Medicine, West China Hospital, Sichuan University, Chengdu 610041, P. R China

[2] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

[3] The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada

[4] Department of Computer Science, University of Toronto, Toronto, ON, Canada

[5] Department of Respiratory and Critical Care Medicine, West China Hospital, Sichuan University, Chengdu 610041, P. R China.

* Correspondence should be addressed to Binwu Ying. Tel: [86-028-85422751]; Fax: [86-028-85422751]; Email: [docbwy@126.com]. Correspondence may also be addressed to Zhaolei Zhang. Tel: [1-4169460924]; Fax: [1-4169788287]; Email: zhaolei.zhang@utoronto.ca.

[†] The first three authors (Xuejiao Hu, Jingwei Zhang and Shun Liao) contributed equally to this manuscript.

Keywords

lncRNA, electronic health record, clinically diagnosed pulmonary tuberculosis, nomogram

LIST OF ABBREVIATIONS

TB: tuberculosis; PTB: pulmonary tuberculosis; MTB: Mycobacterium tuberculosis; IGRA: interferon-gamma release assays; lncRNA: long noncoding RNA; LTBI: latent TB infection; DE: differentially expressed; EHR: electronic health record; TBM: tuberculous meningitis; non-TB DC: non-tuberculosis disease control; PBMC: peripheral blood mononuclear cell; HTA: Human Transcriptome Array; GEO: Gene Expression Omnibus; SVM: support vector machine; VIF: variance inflation factor; DCA: decision curve analysis; RACE: rapid amplification of cDNA ends; LV: lentivirus vector; MOI: multiplicities of infection; BCG: Bacillus Calmette-Guerin; LDH: lactate dehydrogenase.

ABSTRACT

Background

Clinically diagnosed pulmonary tuberculosis (PTB) patients lack *Mycobacterium tuberculosis* (MTB) pathogenic evidence, and misdiagnosis or delayed diagnosis often occurs in clinical practice. We aimed to investigate the potential of lncRNAs and corresponding predictive models to diagnose these patients.

Methods and Findings

We totally enrolled 1372 subjects in three cohorts and conducted this study through a prospective four-step approach. In the Screening Step, we discovered five lncRNA candidates from blood samples using microarray on a small cohort. In the Selection Step, we used qRT-PCR to test five lncRNAs on the Selection Cohort (141 clinically diagnosed PTB patients, 159 non-TB disease controls, 578 healthy subjects) and determined three differentially expressed lncRNAs (*ENST00000497872*, *n333737*, *n335265*). In the Model Training Step, we developed logistic regression models to distinguish clinically diagnosed PTB cases from suspected patients of the Selection Cohort, by combining these three lncRNAs with patients' electronic health records (EHRs). The model with the best c-statistic of 0.91 was presented as a Web-based nomogram (nine variables: *ENST00000497872*, *n333737*, *n335265*, age, hemoglobin, weight loss, low-grade fever, CT calcification, TB-IGRA; available at https://xuejiao.shinyapps.io/shiny/). In the Validation Step, the nomogram was evaluated in the independent Validation Cohort, with good calibration and discrimination (c-statistic: 0.89, sensitivity: 0.80 and specificity: 0.85). A higher clinical net benefit was observed in the nomogram as compared with the EHR model using decision curve analysis. Additionally, The biological function of the most significant lncRNA was interrogated using ELISA, lactate dehydrogenase release analysis and flow cytometry. The results indicated that

ENST00000497872 may involve in regulating inflammatory cytokine production, cell death and apoptosis during MTB infection.

Conclusions

LncRNAs (*ENST00000497872*, *n333737*, *n335265*) could be potential diagnostic biomarkers for clinically diagnosed PTB. The nomogram incorporating lncRNAs and EHRs showed good performance in identifying clinically diagnosed PTB patients. Utilizing the user-friendly nomogram could facilitate the rapid identification of PTB cases among highly suspected patients with negative MTB pathogenic evidence, and thereby help to optimize TB diagnostic procedure.

Introduction

Tuberculosis (TB) persists as a serious public health problem and the world's leading cause of death from a single infectious agent [1]. In 2017, only 56% of pulmonary tuberculosis (PTB) patients reported to WHO were bacteriologically confirmed. In other words, one could argue that approximately half of all PTB cases were clinically diagnosed worldwide, and this proportion can be up to 68% in China [1]. Clinically diagnosed PTB is defined as PTB cases with symptomatic illness and negative *Mycobacterium tuberculosis* (MTB) pathogenic evidence through smear microscopy, culture or nucleic acid amplification test [1–3]. In the absence of "gold standard", patients with this type of TB are clinically diagnosed on the combination of clinical manifestations, radiological findings, TB immuno-related laboratory test, such as widely used interferon-gamma release assays (IGRA), response to empiric anti-TB treatment, and exclusion of other pulmonary diseases [2,3]. The diagnostic procedure currently in use is inadequate and time-consuming in TB case finding, and misdiagnosis or delayed diagnosis often occurs in clinically diagnosed PTB patients [3], which induces an increased risk of morbidity and mortality for patients with severe diseases [4], overtreatment and even TB transmission [5]. Thus it is urgently necessary to explore effective host immune-response signatures, rapid and accurate strategies [6,7] to diagnose PTB cases without MTB pathogenic evidence.

Long noncoding RNAs (lncRNAs), which comprise the most abundant species of noncoding RNAs longer than 200 bp, have been proved as critical regulators of inflammatory responses to infection, especially for T-cell immune response [8,9]. For example, Wang et al. demonstrated that lncRNA-CD244 mediated H3K27 trimethylation in CD8+ T cells and repressed IFN-γ/TNF-α expression during MTB-infection [8]. Increasing evidence has exposed that the expressions of lncRNAs in blood samples are closely associated with TB diseases [10–12]. Yi et al. first reported that substantial lncRNAs were aberrantly expressed

in active TB patients compared to latent TB infection (LTBI) and healthy controls [10]; subsequent studies unveiled differentially expressed (DE) lncRNA profiles in TB patients [11,12], suggesting lncRNAs as potential noninvasive biomarkers for TB detection. However, previous researches suffered from small sample size (ranging from 66 to 510) and absence of independent cohort validation, and lack of the knowledge about the diagnostic potentials and biological functions lncRNAs in clinically diagnosed PTB patients.

Recently, many efforts in the TB field have been dedicated to establishing clinical prediction rules or predictive models for TB diagnosis based on patients' electronic health record (EHR) information [13–16]. With a limited number of clinical-radiological predictors, predictive models can facilitate the diagnosis of PTB patients cost-effectively. A good example is a 6-signature model from Griesel et al. [13], which consisted of cough ≥14 days, unable to walk unaided, temperature > 39°C, chest radiograph assessment, hemoglobin and white cell count, and attained a c-statistic of 0.81 (0.80-0.82) in seriously ill HIV-infected PTB patients. Despite these advances, current EHR models remain insufficient for precise TB diagnosis. A meta-analysis from Pinto and colleagues described high pooled sensitivities (0.96 [0.93-0.98]) but low specificities (0.46 [0.35-0.50]) in clinical-radiological PTB models [16]. Compelling studies have proposed that modeling incorporating biomarkers and EHR information attain better performance for disease prediction in sepsis [17] and abdominal aortic aneurysm [18]. We previously found that combining exosomal microRNAs and EHRs achieved an AUC of up to 0.97, while EHR data alone gained an AUC of 0.67 in the diagnosis of tuberculous meningitis (TBM) [19]. Taking inspiration from previous work, it is worth developing predictive models for identifying PTB cases with negative MTB evidence, by combining specific lncRNA markers with well-defined EHR predictors.

This study aimed to investigate the diagnostic potential of lncRNAs and corresponding predictive models for PTB patients with negative MTB evidence. In addition, we explored

the regulatory functions of lncRNAs during MTB-infection to provide a biological basis for lncRNA candidates.

Methods

Study design

We performed this study through a four-stage strategy. Differentially expressed (DE) lncRNAs were profiled between clinically diagnosed PTB patients and healthy subjects using microarray; five lncRNA candidates were chosen in the Screening Step. In the Selection Step, we used qRT-PCR to detect these five lncRNAs on a much larger prospective cohort and reduced the number of the selected lncRNAs to three. In the next Model Training Step, lncRNAs and EHRs were used to develop predictive models for clinically diagnosed PTB patients and non-tuberculosis disease control (non-TB DC) patients, and the optimal model was visualized as a nomogram. Finally, we validated lncRNAs and the nomogram in an independent prospective cohort. Functional analyses were also performed to elucidate the biological significance of lncRNA biomarkers. The study strategy is shown in Fig 1.

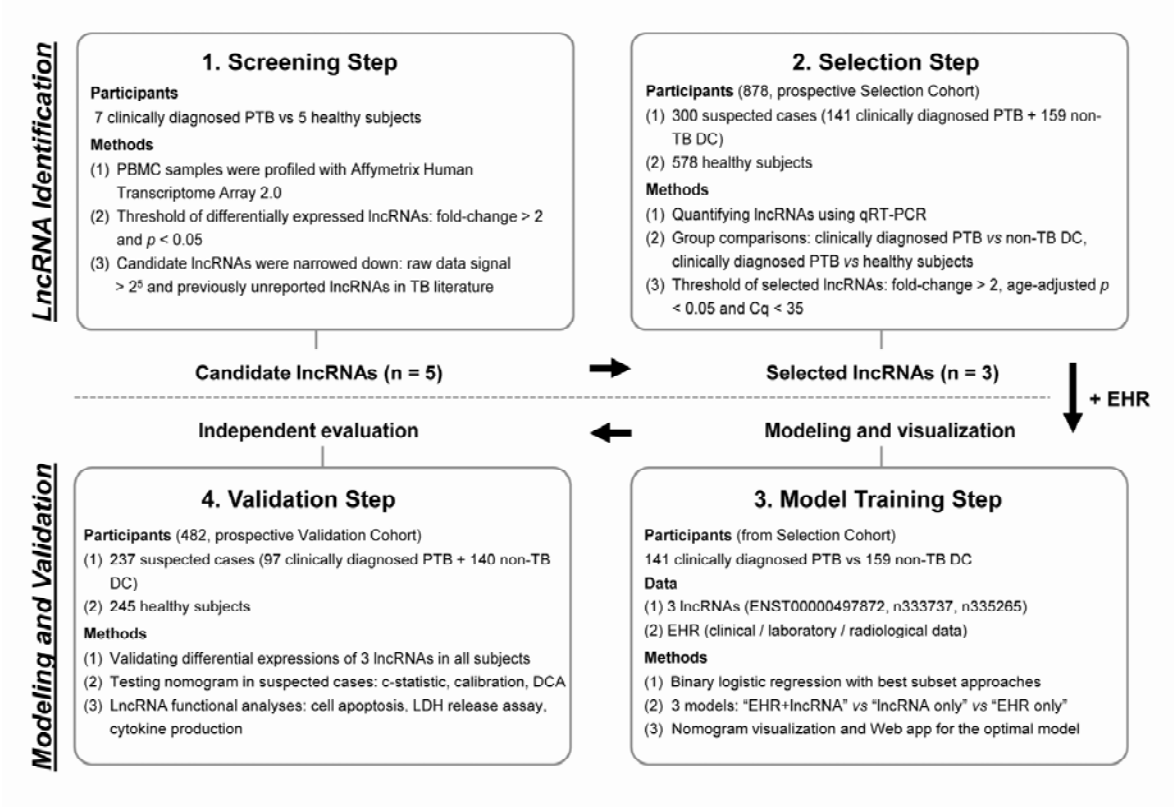

Fig 1. Overview of the strategy for investigating lncRNA and prediction model for clinically diagnosed PTB differential diagnosis.

Abbreviations: PTB, pulmonary tuberculosis; PBMC, peripheral blood mononuclear cell; non-TB DC, non-tuberculosis disease control; DE, differentially expressed; EHR, electronic health record; DCA, decision curve analysis. LDH, lactate dehydrogenase.

Subjects enrolment and diagnostic procedure

Screening Cohort We retrospectively collected age- and gender-matched 7 PTB cases and 5 healthy controls as the Screening Cohort. They were 6 males and 6 females from ages 22 to 59 years. PTB cases were clinically confirmed PTB patients with positive TB symptoms, negative MTB pathogenic examinations, and good response to anti-TB therapy. Healthy subjects had a normal physical examination and no history of TB.

Selection Cohort and Validation Cohort After screening inpatients with high clinical suspicion of PTB from West China Hospital between Dec 2014 and May 2017, a prospective Selection Cohort of 311 suspected PTB patients with negative MTB pathogenic evidence and an independent Validation Cohort of 250 suspected cases satisfied the inclusion criteria and were preliminarily enrolled. The detailed description of patients' recruitment, laboratory examinations, and treatment management are provided in S1 Appendix, Section 1–3.

A panel of experts consisting of two experienced pulmonologists reviewed all presumptive PTB patients, and final diagnoses for all cases were based on the combination of clinical assessment, radiological and laboratory results, response to the treatment. A 12-month follow-up observation was used to further confirm the classification of PTB and non-TB patients. Briefly, according to the Chinese diagnostic criteria for PTB (WS288–2008), clinically diagnosed PTB patients were diagnosed based on the following criteria: (a) clinical symptoms suggestive of active PTB patients; (b) chest imaging or bronchoscope showing lesions of active PTB; (c) repeated smear microscopy, culture and nucleic acid amplification testing negative for MTB detection, (d) appropriate response to anti-TB therapy, (e) exclusion of other lung diseases, (f) others: positive anti-TB immune response (using PPD test, IGRA or anti-TB antibody test), histopathological extra-pulmonary evidence for extra-PTB. Subjects for whom (a) MTB tests were negative, (b) confirmation of an alternative diagnosis was made, and/or (c) response to therapy was documented with a lack of anti-TB treatment, were considered as non-TB DC patients. In addition, 578 and 245 healthy subjects were simultaneously recruited in the correspond cohorts from a pool of healthy donors with a normal physical examination and no history of TB.

A total of 24 patients (11 in the Selection Cohort and 13 in the Validation Cohort) who did not consent to participate, having an uncertain diagnosis, or lost to follow-up, were further excluded. We finally enrolled a Selection Cohort of 878 participants (141 clinically

diagnosed PTB, 159 non-TB DC and 578 healthy subjects) and an independent Validation Cohort of 482 participants (97 clinically diagnosed PTB, 140 non-TB DC and 245 healthy subjects), respectively. Details of the non-TB DC are in S1 Table.

Ethical approval was obtained from the Clinical Trials and Biomedical Ethics Committee of West China [no. 2014 (198)]. Informed consents were obtained from every participant. We reported the present study in conformity to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) recommendations (S1 TRIPOD Checklist).

LncRNA detection

RNA isolation and cDNA preparation Peripheral blood mononuclear cell (PBMC) samples were freshly isolated from 3 ml blood from every participant using a Human Lymphocyte Separation Tube Kit (Dakewe Biotech Company Limited, China). Total RNA was extracted from PBMC using Trizol reagent (Invitrogen, USA). RNA concentration and purity were evaluated by using a NanoDrop ND-1000 (Thermo Fisher Scientific, USA). RNA quality was further evaluated by 1% agarose gel electrophoresis, as can be seen in S1A Fig. The PrimeScript^{TM} RT reagent Kit with gDNA Eraser (Takara, Japan) was used to remove gDNA and synthesize cDNA, according to the manufacturer's instructions.

LncRNA microarray profiling LncRNA profiles were detected using Affymetrix Human Transcriptome Array (HTA) 2.0 Chips. RNA isolation, cDNA preparation and microarray hybridization were performed by Shanghai Gminix Biological Information Company (China) based on the standard Affymetrix HTA 2.0 Chips protocol, as described by a previous study [20]. Raw data were normalized using the Robust Multi-Array Average Expression Measure algorithm. DE lncRNAs with statistical significance p-value < 0.05 and fold-change > 2 [21] were identified using the empirical Bayes moderated t-statistics and presented using

hierarchical clustering and volcano plot. The microarray data have been deposited in the Gene Expression Omnibus (GEO) under the accession GSE119143.

qRT-PCR for lncRNAs LncRNA expressions were measured using the SYBR® Green PCR Kit (Takara, Japan). Specific primers are in S2 Table. A final 10 μl reaction mixture for qRT-qPCR included 5 μl of SYBR® Premix Ex Taq™ II, 0.5 μl of 10 μM forward and reverse primer, 3 μl of ddH_2O (PCR grade), and 1 μl of template cDNA. The cycling program consisted of 95 ºC for 1 min, followed by 35 cycles at 95 ºC for 10 s, 65 - 0.5 ºC/cycle for 30 s and 72 ºC for 60 s. The samples were denatured at 95 °C for 30 s and then heated to 65 °C for 30 s at a rate of 0.2 °C/s. Fluorescence was measured to generate a high-resolution melting curve to analyze the purity of the amplification products (single peak usually means a high purity of PCR product), and for a more accurate PCR product validation, we further sequenced candidate lncRNAs in several random samples, which can be found at the following link https://github.com/xuejiaohu123/TBdiagnosisModel. The ddH_2O negative control and blank control in each reaction showed no detectable signals, ensuring the lack of contamination or nonspecific products. Each sample was run in triplicate, and a CV < 15% cutoff was established to validate the technical replicates. Raw PCR data were analyzed as Cq values, and Cq < 35 was considered acceptable (see S1B Fig). To estimate reaction efficiency of each assay, a ten-fold serial dilution of control cDNA using synthetic *GAPDH*, covering 5 orders of magnitude, was amplified and used to generate the standard curve, which can be seen in S1C Fig. LncRNA expressions were measured in a blinded fashion, normalized to the endogenous control *GAPDH*, and calculated according to the $2^{-\Delta\Delta Cq}$ method [22], where $\triangle Cq = Cq_{lncRNA} - Cq_{GAPDH}$, $\triangle\triangle Cq = \triangle Cq - \triangle average\ Cq_{(healthy\ subjects)}$.

We also assessed the short-term stability of lncRNA expressions in PBMC samples. We randomly selected blood samples from 4 individuals for incubation at 4 °C for 0 h, 1 h, 2 h, 6 h, 12 h, and 24 h, respectively. These time points were chosen because blood samples were assessed within 24 h in many clinical labs in China.

Modeling

Data used for modeling A total of 41 EHRs, including patients' demographic, clinical, laboratory, and radiological findings were collected separately by three authors who were blinded to the diagnostic categorization (see S2 Appendix, Section 1). We applied a 20% missing value as thresholds to remove incomplete features. The features with p-values < 0.05 in univariate analysis or definite clinical significance were included for modeling. A total of 14 out of 44 variables remained after filtering, including 11 EHRs and 3 lncRNAs, listed in S2 Appendix, Section 2.

Diagnostic modeling Binary logistic regression with the best subset selection procedure [23] was used to develop predictive models in distinguishing clinically diagnosed PTB cases from suspected patients of the Selection Cohort. The top 10 feature subsets were selected and compared based on the combinational criteria of discriminability (accuracy and Harrell's c-statistic), Akaike Information Criterion and goodness of fit evaluation (Hosmer-Lemeshow test, likelihood ratio test, McFadden R^2, and Nagelkerke R^2). The "EHR+lncRNA", "lncRNA only" and "EHR only" models were developed according to their respective best feature subset in the Selection Cohort. Additionally, the linear kernel support vector machine (SVM) algorithm, as described in our previous report [19], was used to confirm these models.

Nomogram presentation and evaluation We further adopted the nomogram to visualize the optimal model with the best diagnostic performances [24,25]. The variance inflation factor (VIF) quantified the multicollinearity among features (VIF > 10 indicated multicollinearity among the features in the model). Feature importance was calculated with the "varImp"

function in R package. Harrell's c-statistic was measured to quantify the discrimination ability of the nomogram (values 0.7-0.8 are considered acceptable and above 0.8 good) [13]. The nomogram was subjected to bootstrapping internal validation (500 bootstrap resamples) to calculate a bias-corrected c-statistic. The performance of the nomogram was tested in the independent Validation Cohort, with total points for each patient calculated. Nomogram calibration was assessed with the calibration curve and Hosmer-Lemeshow test (p-value > 0.05 suggested no departure from perfect fit). Decision curve analysis (DCA) [25] was performed by evaluating the clinical net benefit of the nomogram and "EHR only" model across the overall datasets. Nomogram and calibration plot were performed using the "rms" R package. DCA was performed using the "rmda" R package. The nomogram was implemented as a Web-based app using R Shiny.

Functional experiment for ENST00000497872 (i.e., lnc AL)

The most significant *ENST00000497872*, hereafter referred to as lnc AL for short, was selected for further functional investigation.

5′ and 3′ RACE analyses for lnc AL 5′ and 3′ rapid amplification of cDNA ends (RACE) analysis were performed to obtain the full length of *ENST00000497872* (called lnc AL, for short) using the SMARTer® RACE 5′/3′ Kit (Takara, Japan) according to the manufacturer's instructions. The full length of lnc AL and the blast result with NCBI database are shown in see S3 Appendix, Section 1.

Construction of stable cell lines with overexpressed and downexpressed lnc AL Based on the full length of lnc AL, recombinant lentivirus vector (LV) encoding lnc AL cDNA (named LV-lnc AL, for short), shRNA against lnc AL (named shRNA-lnc AL, for short), and negative control (LV-negative control and shRNA-negative control) were produced by GeneChem Company (Shanghai, China). Lnc AL cDNA and negative control cDNA were PCR-amplified and subcloned to the LV Ubi-MCS-SV40-EGFP-IRES-puromycin; shRNA-

lnc AL and negative control shRNA were synthesized and inserted into LV hU6-MCS-Ubiquitin-EGFP-IRES-puromycin. Details of LV-lnc AL and shRNA-lnc AL are provided in S3 Appendix, Section 2. THP-1 cells were treated with either LV-lnc AL, shRNA-lnc AL, or negative control at multiplicities of infection (MOI) of 50 (2×10^8 transducing units per mL) in the presence of Polybrene (Santa Cruz Biotechnology), according to the standard protocols. Stably transfected cells were selected with 1 μg/ml puromycin for 4 days and further confirmed by fluorescence microscopy and qRT-PCR.

MTB-infected cell model, cytokine detection, cell apoptosis and cytotoxicity analysis Bacillus Calmette-Guerin (BCG) was used to imitate active MTB infection [26]. The stably over- or down-expressing THP-1 cells (1×10^6 cells/ml) were seeded in 6-well culture plates and allowed to differentiate at 37 ºC in 5% CO_2 for 24 h in the presence of 100 nM Phorbol 12-Myristate 13-Acetate. Five groups were set: overexpressed -lnc AL group (BCG + THP-1 with LV-lnc AL), LV-negative control group (BCG + THP-1 with LV-control), downexpressed-lnc AL group (BCG + THP-1 with shRNA-lnc AL), shRNA-negative control group (BCG + THP-1 with shRNA-control) and blank control group (BCG + THP-1). Following the treatment of THP-1cells with BCG (MOI = 5) for 4 h in solution, the infected cells were cultured for 6 h, 12 h, 24 h, 48 h, and 72 h, respectively. Cells and culture supernatants were harvested to measure the expression of lnc AL and cytokine production, respectively. The expression of lnc AL was quantified using SYBR® Green PCR Kit (Takara, Japan), with *GAPDH* as the reference. ELISA reagents for the following six cytokines: TNF-α, IL-1β, IL-12 p70, IL-10, IFN-γ, and IL-6 were from R&D Systems (Shanghai, China). Apoptosis in the THP-1 cells after 24 h infection was assessed using flow cytometry (Annexin V, 633 Apoptosis Detection Kit; DOJINDO, Shanghai, China) on the BD FACSCalibur™ platform (BD Biosciences, USA). The lactate dehydrogenase (LDH) release analysis was performed using the CytoTox 96® Non-Radioactive Cytotoxicity Assay Kit

(Promega, USA) to evaluate cell death or cytotoxicity. All experiments were performed according to the manufacturer's recommendation and independently in triplicate.

Statistical Analysis

The sample size for the present study was performed using the HyLown Power and Sample Size Calculators (http://powerandsamplesize.com/Calculators/), which can simultaneously determine the error rate and statistical power. Setting the threshold at an error rate of 0.05 and a power of 80%, a sample size of at least 82 patients was required for lncRNA detection in the Selection Cohort and Validation Cohort, respectively. The present sample size incorporated the need for at least 10 clinically diagnosed PTB patient events per predictive variable for binary logistic regression analysis [13].

Univariate analysis was used to analyze categorical variables with a Chi-square test and continuous variables with the Mann-Whitney U test or t-test. All tests were 2-sided, and p-value < 0.05 was considered statistically significant. Modeling was constructed and validated by authors who were blinded to the diagnostic categorization. R code and data for modeling are available from https://github.com/xuejiaohu123/TBdiagnosisModel.

Results

Characteristics of prospectively enrolled participants

The demographic and clinical characteristics of participants in the Selection and Validation Cohorts are provided in Table 1. PTB patients had a much younger age and twice higher yield of TB-IGRA than their non-TB DC (p-value < 0.0001 for both the Selection and Validation Cohorts). There were no significant differences concerning gender, BMI, smoking status between these two groups. Healthy subjects were age-, gender-, and BMI-matched with

PTB patients, who had significantly different blood test results compared with PTB patients (Table 1).

Table 1. Demographic and clinical features of participants in the Selection and Validation Cohorts.

Clinical features	Suspected clinically diagnosed PTB patients $_{Selection}$			HS $_{Selection}$ (n=578)	p2	Suspected clinically diagnosed PTB patients $_{Validation}$			HS $_{Validation}$ (n=245)	p4
	Clinically diagnosed PTB (n=141)	Non-TB DC (n=159)	p1			Clinically diagnosed PTB (n=97)	Non-TB DC (n=140)	p3		
Gender, male	84 (59.57%)	95 (59.75%)	0.976	284 (49.13%)	0.026	58 (59.79%)	87 (62.14%)	0.715	126 (51.43%)	0.162
Age (years)	37.81 ± 17.93	56.68 ± 14.52	< 0.0001	40.59 ± 13.11	0.084	38.29 ± 17.57	57.96 ± 16.66	< 0.0001	36.82 ± 9.28	0.436
BMI (kg/m^2)	20.81 ± 2.99	20.43 ± 4.03	0.359	20.65 ± 3.19	0.57	21.59 ± 3.43	21.29 ± 3.62	0.52	21.51 ± 3.52	0.843
Smoking	61 (43.26%)	72 (45.28%)	0.725	161 (27.85%)	< 0.0001	41 (42.27%)	66 (47.14%)	0.458	84 (34.29%)	0.167
Radiologic pathology	116(82.26%)	140 (88.05%)	0.158	-		86 (88.66%)	130 (92.86%)	0.264	-	
Laboratory tests										
Positive TB-IGRA	97 (68.88%)	56 (35.22%)	< 0.0001	-		64 (66.00%)	42 (30.00%)	< 0.0001	-	
C-reactive protein	16.30 (5.32-54.05)	17.80 (6.47-60.20)	0.427	-		13.60 (3.34-43.55)	18.60 (6.56-70.63)	0.037	-	
Hematocrit	0.37 ± 0.06	0.36 ± 0.07	0.162	0.44 ± 0.04	< 0.0001	0.38 ± 0.07	0.35 ± 0.07	0.002	0.43 ± 0.04	< 0.0001
Erythrocytes	4.33 ± 0.72	4.03 ± 0.81	0.001	4.78 ± 0.46	< 0.0001	4.46 ± 0.80	3.89 ± 0.91	< 0.0001	4.80 ± 0.46	< 0.0001
Hemoglobin	122.57 ± 23.22	115.82 ± 25.20	0.017	144.46 ± 13.88	< 0.0001	125.08 ± 24.25	113.11 ± 25.52	< 0.0001	145.82 ± 13.73	< 0.0001
Platelets	238.00 (177.00-305.00)	233.00 (149.00-299.00)	0.171	193.00 (158.00-223.00)	< 0.0001	220.00 (160.50-313.00)	199.50 (137.00-290.75)	0.059	190.00 (165.00-230.00)	0.001
Leukocytes	6.03 (4.76-8.25)	6.36 (4.71-9.07)	0.488	5.92 (5.18-6.67)	0.184	6.96 (5.06-9.14)	5.93 (4.34-8.33)	0.009	5.70 (4.91-6.55)	0.217
Lymphocytes	1.15 (0.80-1.52)	1.28 (0.87-1.87)	0.056	1.86 (1.55-2.19)	< 0.0001	1.29 (0.91-1.80)	1.22 (0.86-1.62)	0.343	1.85 1.57-2.55)	< 0.0001
Neutrophils	4.02 (3.23-5.93)	4.08 (2.71-6.32)	0.956	3.47 (2.87-4.09)	< 0.0001	4.03 (2.51-5.69)	4.85 (3.21-6.70)	0.023	3.36 (2.75-3.92)	0.006
Monocytes	0.47 (0.35-0.65)	0.42 (0.26-0.61)	0.015	0.36 (0.29-0.44)	< 0.0001	0.43 (0.30-0.64)	0.46 (0.34-0.71)	0.211	0.31 (0.25-0.39)	< 0.0001
Alb	36.66 ± 6.78	36.60 ± 6.66	0.973	48.24 ± 2.67	< 0.0001	37.33 ± 7.47	36.69 ± 7.22	0.509	47.06 ± 2.25	< 0.0001
Globin	31.69 ± 7.66	30.68 ± 8.10	0.269	28.92 ± 3.29	0.041	30.41 ± 7.89	29.73 ± 7.91	0.514	27.41 ± 3.19	< 0.0001

Subscripted "Selection" or "Validation" refers to the Selection or Validation Cohort, respectively. Abbreviations: non-TB DC, non-tuberculosis disease control patients; HS, healthy subjects. p1, p-value for the comparison of clinically diagnosed PTB patients and non-TB DCs in the Selection Cohort; p2, p-value for the comparison of clinically diagnosed PTB patients and healthy subjects in the Selection Cohort; p3, p-value for the comparison of clinically diagnosed PTB patients and non-TB DCs in the Validation Cohort; p4, p-value for the comparison of clinically diagnosed PTB patients and healthy subjects in the Validation Cohort.

We noticed that clinically diagnosed PTB patients (141 + 97, 238 cases) were responsible for 29.82% (238/798) of all active PTB patients, including bacteriologically and clinically confirmed (798 cases, see S1 Appendix, Section 3). Although significantly lower than a nationwide estimate of 68% mainly based on primary public health institutions [1], this rate represents the occurrence of clinically diagnosed PTB in a referral hospital with experienced specialists in China, which warrants clinical coworkers' attention.

LncRNAs microarray profiles and candidate selection

In the Screening Step, microarray profiling generated a total of 325 differentially expressed lncRNAs (287 upregulated and 38 downregulated) for clinically diagnosed PTB patients in comparison to healthy subjects. Hierarchical clustering and the volcano plot clearly showed distinguishable lncRNA expression profiles in S2 Fig. We selected the top 5 candidate lncRNAs based on a set of combined criteria: fold-change > 2 between groups, p-value < 0.05, signal intensity > 2^5 [27], and previously unreported lncRNAs in TB literature [28]. Among these five lncRNAs, three were upregulated (*n335265*, *ENST00000518552* and *TCONS_00013664*) and two were downregulated (*n333737* and *ENST00000497872*) (S3 Table).

Differentially expressed lncRNAs in clinically diagnosed PTB

In the Selection Step, we used qRT-PCR to measure the expression level of these five candidate lncRNAs in the Selection Cohort of 141 clinically diagnosed PTB, 159 non-TB DC, and 578 healthy subjects. Two lncRNAs (*ENST00000518552* and *TCONS_00013664*) performed very low-abundance (Cq > 35) were excluded. Among the three remaining lncRNAs, *ENST00000497872* and *n333737* were downregulated, and *n335265* was upregulated in PTB patients in comparison with healthy subjects (S4 Table). Comparison between clinically diagnosed PTB cases and non-TB DC patients revealed a decreased

expression of *ENST00000497872* and *n333737* in PTB patients (S3A Fig, age-adjusted p-values both < 0.0001).

We also assessed the short-term stability of lncRNA in PBMC. S5 Table shows prolonged incubation up to 24 h had minimal effect on the expression of *ENST00000497872*, *n333737* and *n335265*, which was in accordance with the previous report that lncRNA existed stably in blood [29].

Diagnostic modeling and nomogram visualization

In the Model Training Step, three different logistic regression models, "EHR+lncRNA", "EHR only" and "lncRNA only" were built in the Selection Cohort. Features in these models are listed in the S2 Appendix, Section 3. The VIF between the features ranged from 1.02 to 1.29, indicating no collinearity within models. The "EHR+lncRNA" model yielded the highest bias-corrected c-statistic of 0.91 in distinguishing clinically diagnosed PTB from suspected patients, compared with the "EHR only" and "lncRNA only" model (0.85 and 0.81, respectively). The "EHR+lncRNA" model also presented the best performance in sensitivity, specificity, accuracy, positive predictive value, and negative predictive value (see Table 2).

Table 2. Performances of diagnostic models in differentiating clinically diagnosed PTB from suspected patients.

Model performance	Selection Cohort			Validation Cohort		
	EHR+lncRNA (Nomogram)	EHR only	lncRNA only	EHR+lncRNA (Nomogram)	EHR only	lncRNA only
Corrected c-statistic	0.91 (0.87-0.94)	0.85 (0.81-0.89)	0.81 (0.76-0.85)	0.89 (0.84-0.92)	0.85 (0.80-0.89)	0.79 (0.73-0.84)
Sensitivity	0.84 (0.80-0.88)	0.79 (0.74-0.84)	0.72 (0.67-0.77)	0.80 (0.75-0.85)	0.74 (0.68-0.80)	0.65 (0.59-0.71)
Specificity	0.85 (0.81-0.89)	0.78 (0.73-0.83)	0.73 (0.68-0.78)	0.85 (0.79-0.89)	0.83 (0.78-0.88)	0.74 (0.68-0.80)
Accuracy	0.85 (0.80-0.89)	0.79 (0.74-0.83)	0.72 (0.67-0.77)	0.83 (0.77-0.87)	0.79 (0.73-0.84)	0.71 (0.64-0.76)
Positive predictive value	0.83 (0.79-0.87)	0.74 (0.70-0.80)	0.68 (0.65-0.75)	0.77 (0.72-0.82)	0.75 (0.69-0.81)	0.60 (0.54-0.66)
Negative predictive value	0.86 (0.82-0.89)	0.82 (0.78-0.86)	0.76 (0.71-0.81)	0.86 (0.82-0.90)	0.81 (0.76-0.86)	0.78 (0.73-0.83)

Note: Corrected c-statistic is a bias-corrected Harrell's c-statistic through 500 bootstrapping internal validation method. The original c-statistic for nomogram, "EHR only" model and lncRNA model was 0.92, 0.87, and 0.82 in the Selection Cohort, respectively; 0.90, 0.86, and 0.80 in the Validation Cohort, respectively.

The optimal "EHR+lncRNA" model was displayed as a nomogram (Fig 2A), in which points are assigned based on the feature rank order of the effect estimates. Top five important features were *ENST00000497872*, age, *n333737*, CT calcification, and TB-IGRA (S6 Table). Calibration curve in the Selection Cohort (Fig 2B) indicated a good agreement between nomogram prediction and actual PTB cases and was confirmed by the nonsignificant Hosmer-Lemeshow test (p-value = 0.957). To derive a prediction, we draw a vertical line towards the "Point" axis to the corresponding point for each feature and sum all of these points to a total point score. A vertical line towards the "Total Points" axis is drawn to determine the PTB probability. We generated this nomogram as a free online app (available at https://xuejiao.shinyapps.io/shiny/) for better clinical use. This simple app allows the user to insert the values of specific predictors, and the outcome of the risk prediction is spontaneously given as a whole number percentage.

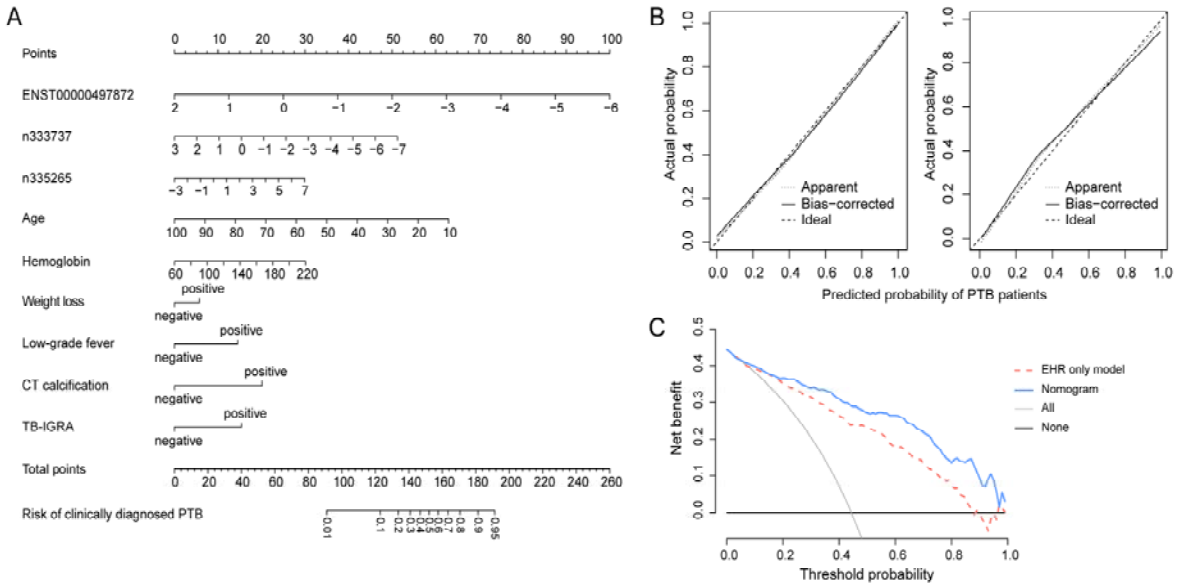

Fig 2. Nomogram for the prediction of clinically diagnosed PTB based on the optimal models. (A), Nomogram to predict the risk of clinically diagnosed PTB patients.

(B), Calibration plot in the Selection Cohort (left in B) and Validation Cohort (right in B). The 45-degree dash line (Ideal) represents a perfect prediction. The dotted line (Apparent) and full line (Bias-corrected) represent the uncorrected and corrected prediction of the nomogram, respectively.

(C), Decision curve analysis for the nomogram and "EHR only" model. Red line: the nomogram. Blue dash: the "EHR only" model. The horizontal black line along the x-axis assumes that no patients have PTB, whereas the gray line represents the hypothesis that all patients have PTB.

Validation for lncRNAs and the nomogram

In the Validation Step, we tested the 3 lncRNAs in an independent Validation Cohort of 97 clinically diagnosed PTB cases, 140 non-TB DC and 245 healthy subjects, and observed similar dysregulation pattern of lncRNAs as that in the Selection Cohort (S4 Table, S3B Fig), which indicated 3 lncRNAs as potential biomarkers for PTB patients with negative MTB pathogen evidence.

We also applied three models to the Validation Cohort, and their performances are shown in Table 2. The nomogram achieved a superior discrimination (bias-corrected c-statistic of 0.89 [0.84-0.92]), good calibration (Fig 2B, and p-value = 0.668 for Hosmer-Lemeshow test) for clinically diagnosed PTB prediction. DCA in Fig 2C further showed that the nomogram outperformed the conventional "EHR only" model with a higher clinical net benefit where the threshold of choice for PTB risk ranged from 0.2 to 1.

In addition, SVM model composed of "EHR+lncRNA" illustrated a better performance in the Selection and Validation Cohorts when compared with the SVM model solely based on EHRs or lncRNAs (S4 Fig), which also supported the feasibility of integrating lncRNAs and EHRs.

LncRNA response to anti-TB treatment

LncRNAs were further assessed to predict anti-TB treatment response. Paired samples before and after 2-month intensive therapy [30] were collected from 22 clinically diagnosed PTB patients, and the expressions of *ENST00000497872*, *n333737* and *n335265* were measured by qRT-PCR. All the patients had a good response to therapy based on the observed clinical and radiological relief. Accordingly, *ENST00000497872* and *n333737* significantly increased in the clinically diagnosed PTB patients (p-values = 0.005 and 0.0005, respectively, Fig 3), which implied restored lncRNA expressions after therapy.

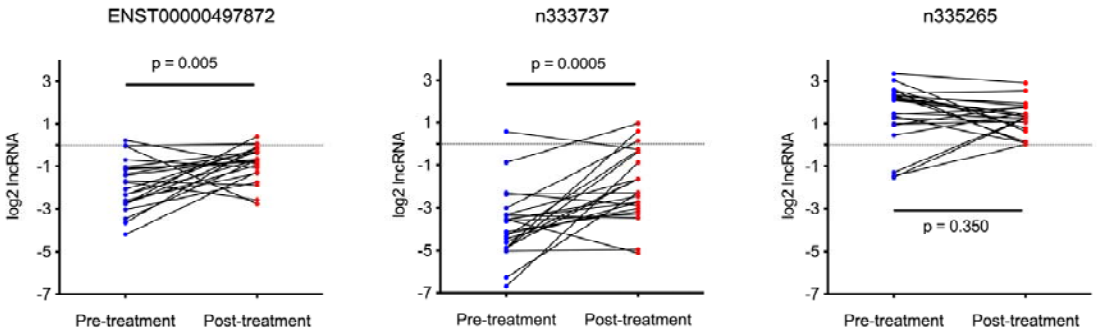

Fig 3. Alteration of lncRNAs before and after 2-month intensive therapy. The blue dots and red dots indicate lncRNA expressions before and after the 2-month intensive anti-TB treatment, respectively. Altered lncRNA expressions were calculated using $\log_2$ lncRNA (post-treatment expression / pre-treatment expression). The Wilcoxon matched-paired rank test was used for comparisons among paired samples.

Functional studies of ENST00000497872

We investigated whether *ENST00000497872* (i.e., lnc AL) could affect the host immune response. As shown in Fig 4 and S5 Fig, following 24 h and 48 h post BCG-infection, lnc AL overexpression (i.e., LV- lnc AL group) led to the decreased production of proinflammatory cytokines TNF-α and IL-1β and the increased level of INF-γ (Fig 4A). In parallel, knockdown of lnc AL (i.e., ShRNA- lnc AL group) resulted in a significant increase of TNF-

α and IL-1β and reduction of INF-γ. An increasing trend of cell apoptosis (Fig 4B and 4C) and cell death (Fig 4D) was observed when lnc AL was knocked down. These results implicate an inflammatory regulation of lnc AL during MTB-infection.

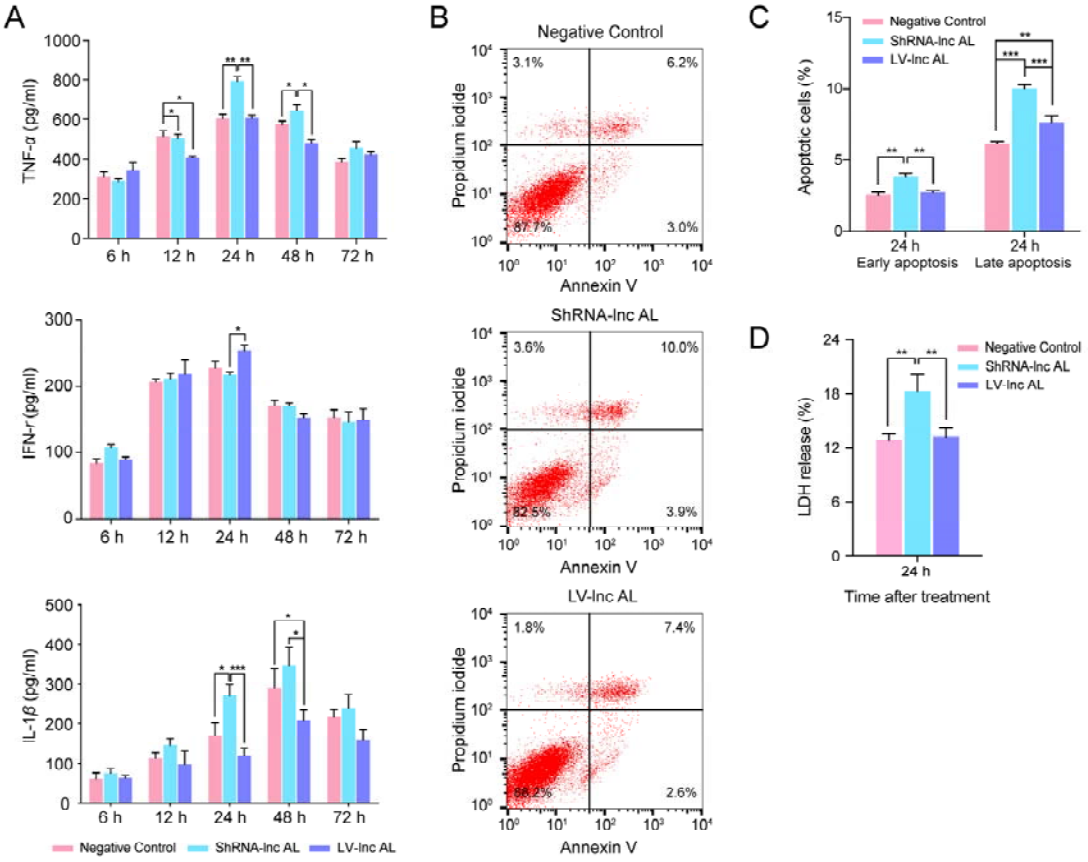

Fig 4. Regulation of lncRNA on inflammatory cytokine, cell apoptosis and cytotoxicity in BCG-infected THP-1 cells.

(A), Alteration of cytokine expression. (B), Cell apoptotic assay with flow cytometry. (C), Quantitative results of apoptotic assay. (D), LDH release analysis for cell cytotoxicity. The average values of LV-control and shRNA-control were calculated as the values of the negative control. Blank control is not shown in the figure. The results of three cytokines (IL-12 p70, IL-10 and IL-6) are not significant and not shown in the Fig 4A. One-way ANOVA and Bonferroni's post-test comparison among groups are adopted. *p-value < 0.05, **p-value < 0.01, and ***p-value < 0.001.

Discussion

The present work focused on the challenging issue of the accurate diagnosis for PTB patients with negative MTB pathogenic evidence. We identified three differentially expressed lncRNAs in clinically diagnosed PTB patients. More importantly, we firstly developed and validated a novel nomogram incorporating lncRNA signatures and conventional EHR features, which can effectively discriminate PTB patients from patients with suspected diseases.

Implementation of three lncRNAs (*ENST00000497872*, *n333737* and *n335265*) to a conventional EHR model enabled us to better identify PTB from suspects, with the c-statistic of the model increased from 0.85 to 0.89. The expression of *ENST00000497872* and *n333737* increased significantly after 2-month intensive therapy, indicating that changes in lncRNA levels may reflect patients' responses to anti-TB treatment, which needs further validation. *ENST00000497872* (chr14:105703964-105704602) is located close to *Immunoglobulin Heavy Constant Alpha 1* gene (chr14: 105703995-105708665), and *n333737* (chr14:21712368-21712835) overlaps with *T cell receptor alpha variable 2* gene (chr14:21712321-21712843). We speculated that these two lncRNAs might participate in regulating host immune response. Indeed, subsequent functional experiment demonstrated their biological significance of lncRNA marker regarding its involvement of regulating inflammatory cytokine production and cell apoptosis during MTB-infection; however, further studies are needed to investigate the underlying molecular mechanisms. Consistent with published lncRNA data [8–12, 31], these findings provided additional affirmative evidence that lncRNAs could participate in TB immunoregulation and serve as promising biomarkers for TB diagnosis.

In addition to lncRNAs, six EHR predictors identified in the nomogram were age, CT calcification, TB-IGRA, low-grade fever, hemoglobin and weight loss. From our study, age was an important negative indicator, which seems in discord with the general concept that

advanced age correlates with a higher TB susceptibility [32]. The different enrollment of control subjects may explain this inconsistency. Previous studies included healthy and/or vulnerable subjects as controls, while we prospectively enrolled inpatients with a wide range of pulmonary diseases and older ages as disease controls. Other five well-defined EHR features, despite their nonspecificity, still were essential in differentiating between PTB and non-TB pulmonary disorders, as proposed by prior findings [14–16].

We carefully developed and validated the models in large prospective cohorts with standardized statistical methods. Feature subset selection, bootstrapping internal validation and independent validation were used to guarantee the model's reliability. We proposed that c-statistic as an overall measure of model performance and DCA quantifying the incremental clinical value of models should be considerd simultaneously to identify useful tools. "EHR+lncRNA" model had excellent discrimination (through c-statistic and diagnostic parameters), reliable calibration (via calibration curve and Hosmer-Lemeshow test), and potential clinical utility for decision-making (using DCA), which exceeded the conventional EHR model for clinically diagnosed PTB prediction. Within non-sputum based features, "EHR+lncRNA" model avoided some common problems such as poor sputum quality or problematic sampling [33], making it more reliable and clinically useful. Our study serves as the first proof-of-concept that integrating lncRNA signatures and EHR data could be a more promising diagnostic approach for PTB with negative pathogenic evidence.

Nomogram has been shown to remarkably promote early diagnosis of intestinal tuberculosis [34] and prognosis prediction in PTB [35] and TBM [24]. "EHR+lncRNA" model herein was visualized as a nomogram and further implemented in an app. The online nomogram uses readily obtainable predictors and automatically outputs a personalized quantitative risk estimate for PTB. Utilizing this user-friendly tool may facilitate the rapid identification of PTB cases among highly suspected patients with negative MTB pathogenic

evidence, and thereby help to optimize TB diagnostic procedure, especially in the resource-constrained areas with high TB prevalence.

Our study has several limitations. We have only explored a small number of lncRNAs in large cohorts; more potential candidates should be interrogated in the future. Modeling was conducted based on data from a single large hospital, and multi-center validation studies are needed before recommending the nomogram. Because Xpert MTB/RIF is still not routinely available in most clinical laboratories of China, and previous Xpert studies reported moderate sensitivities ranged from 28% to 73% [36–38] in smear-negative PTB patients, we did not consider Xpert in our research, which may limit the generalization of our findings.

In conclusion, our study showed that three lncRNAs (*ENST00000497872*, *n333737*, and *n335265*) were potential biomarkers for clinically diagnosed PTB patients. We also developed and validated a novel nomogram incorporating lncRNAs and EHRs, which may be a useful predictive tool in identifying PTB patients with negative MTB pathogenic evidence.

Conflict of interest

All the authors have reviewed and approved the manuscript and declared no competing interests. This work was funded by the University - City Science and Technology Cooperation Project of Sichuan University & Panzhihua City [2018CDPZH-13], the National Natural Science Foundation of China [81672095], and Natural Science and Engineering Council of Canada [RGPIN-2017-06743].

Acknowledgments

We thank Dr. Yi Zhang from Gminix (Shanghai, China) for assistance with bioinformatic analyses.

Author Contributions

Conceptualization: Xuejiao Hu, Xuerong Chen, Zhaolei Zhang, Binwu Ying.

Data curation: Xuejiao Hu, Jingwei Zhang, Shun Liao, Hao Bai, Shubham Gupta.

Formal analysis: Xuejiao Hu, Shun Liao, Shubham Gupta.

Funding acquisition: Zhaolei Zhang, Binwu Ying.

Investigation: Xuejiao Hu, Hao Bai, Yi Zhou, Juan Zhou, Qian Wu, Zirui Meng, Lin Jiao.

Methodology: Xuejiao Hu, Jingwei Zhang, Hao Bai, Yi Zhou, Juan Zhou, Qian Wu, Lijuan Wu, Minjin Wang.

Project administration: Xuerong Chen, Yanhong Zhou, Xiaojun Lu.

Resources: Xuejiao Hu, Hao Bai, Xiaojun Lu, Binwu Ying.

Software: Xuejiao Hu, Shun Liao.

Supervision: Xuerong Chen, Yanhong Zhou, Xiaojun Lu, Zhaolei Zhang.

Validation: Hao Bai, Shubham Gupta.

Visualization: Xuejiao Hu, Shun Liao, Shubham Gupta.

Writing – original draft: Xuejiao Hu, Jingwei Zhang, Shun Liao, Binwu Ying.

Writing – review & editing: Xuejiao Hu, Jingwei Zhang, Shun Liao, Hao Bai, Shubham Gupta, Yi Zhou, Juan Zhou, Qian Wu, Zirui Meng, Lin Jiao, Lijuan Wu, Minjin Wang, Xuerong Chen, Yanhong Zhou, Xiaojun Lu, Zhaolei Zhang, Binwu Ying.

References

1. World Health Organization. Global tuberculosis report; 2018. Available from: https://www.who.int/tb/publications/global_report/en/. Accessed 18 September 2018.
2. The National Health and Family Planning Commission (NHFPC). Pulmonary tuberculosis diagnostic criteria WS288-2017. (in Chinese) Available from: http://www.sific.com.cn/Scripts/ueditor/asp/upload/file/ 20171129/ 15119447887623625.pdf. Accessed 9 November 2017.

3. Gao M. Interpretation of clinical diagnosed pulmonary tuberculosis case in new national diagnostic standard on pulmonary tuberculosis. Chinese Journal of Antituberculosis. 2018; 40(3):243-6. (in Chinese) Available from: http://www.zgflzz.cn/EN/article/advancedSearchResult.do.
4. Getahun H, Harrington M, O'Brien R, Nunn P. Diagnosis of smear-negative pulmonary tuberculosis in people with HIV infection or AIDS in resource-constrained settings: informing urgent policy changes. The Lancet. 2007; 369(9578):2042-9.
5. Tostmann A, Kik SV, Kalisvaart NA, Sebek MM, Verver S, Boeree MJ, et al. Tuberculosis transmission by patients with smear-negative pulmonary tuberculosis in a large cohort in the Netherlands. Clinical Infectious Diseases. 2008; 47(9):1135-42.
6. Haas CT, Roe JK, Pollara G, Mehta M, Noursadeghi M. Diagnostic ‘omics’ for active tuberculosis. BMC medicine. 2016; 14(1):37.
7. Walzl G, McNerney R, du Plessis N, Bates M, McHugh TD, Chegou NN, et al. Tuberculosis: advances and challenges in development of new diagnostics and biomarkers. The Lancet Infectious Diseases. 2018; 18:e199-e210.
8. Wang Y, Zhong H, Xie X, Chen CY, Huang D, Shen L, et al. Long noncoding RNA derived from CD244 signaling epigenetically controls CD8+ T-cell immune responses in tuberculosis infection. Proceedings of the National Academy of Sciences. 2015; 112:E3883-E92.
9. Zhang Q, Chao T-C, Patil V-S, Qin Y, Tiwari S-K, Chiou J, et al. The long noncoding RNA ROCKI regulates inflammatory gene expression. The EMBO Journal. 2019; 38(8).
10. Yi Z, Li J, Gao K, Fu Y. Identifcation of differentially expressed long non-coding RNAs in CD4+ T cells response to latent tuberculosis infection. Journal of Infection. 2014; 69(6):558-68.
11. He J, Ou Q, Liu C, Shi L, Zhao C, Xu Y, et al. Differential expression of long non-coding

RNAs in patients with tuberculosis infection. Tuberculosis. 2017; 107:73-9.

12. Yan H, Xu R, Zhang X, Wang Q, Pang J, Zhang X, et al. Identifying differentially expressed long non-coding RNAs in PBMCs in response to the infection of multidrug-resistant tuberculosis. Infection and Drug Resistance. 2018; 11:945-59.
13. Griesel R, Stewart A, van der Plas H, Sikhondze W, Rangaka MX, Nicol MP, et al. Optimizing tuberculosis diagnosis in human immunodeficiency virus–infected inpatients meeting the criteria of seriously Ill in the world gealth organization algorithm. Clinical Infectious Diseases. 2017; 66(9):1419-26.
14. Soto A, Solari L, Agapito J, Acuna-Villaorduna C, Lambert M-L, Gotuzzo E, et al. Development of a clinical scoring system for the diagnosis of smear-negative pulmonary tuberculosis. Brazilian Journal of Infectious Diseases. 2008; 12(2):128-32.
15. Siddiqi K, Lambert M-L, Walley J. Clinical diagnosis of smear-negative pulmonary tuberculosis in low-income countries: the current evidence. The Lancet Infectious Diseases. 2003; 3(5):288-96.
16. Pinto LM, Pai M, Dheda K, Schwartzman K, Menzies D, Steingart KR. Scoring systems using chest radiographic features for the diagnosis of pulmonary tuberculosis in adults: a systematic review. European Respiratory Journal. 2013; 42(2):480-94.
17. Taneja I, Reddy B, Damhorst G, Zhao SD, Hassan U, Price Z, et al. Combining biomarkers with EMR data to identify patients in different phases of sepsis. Scientific Reports. 2017; 7(1):10800.
18. Li J, Pan C, Zhang S, Spin JM, Deng A, Leung LL, et al. Decoding the genomics of abdominal aortic aneurysm. Cell. 2018; 174(6):1361-72. e10.
19. Hu X, Liao S, Bai H, Wu L, Wang M, Wu Q, et al. Integrating exosomal microRNAs and electronic health data improved tuberculosis diagnosis. EBioMedicine. 2019; 40:564-73.
20. Wang P, Xue Y, Han Y, Lin L, Wu C, Xu S, et al. The STAT3-binding long noncoding

RNA lnc-DC controls human dendritic cell differentiation. Science. 2014; 344(6181):310-3.

21. Tang X, Li Y, Liang S, Jiang W, Liu F, Ge W, et al. Development and validation of a gene expression-based signature to predict distant metastasis in locoregionally advanced nasopharyngeal carcinoma: a retrospective, multicentre, cohort study. The Lancet Oncology. 2018; 19(3):382-93.
22. Mavridis K, Stravodimos K, Scorilas A. Downregulation and Prognostic Performance of MicroRNA 224 Expression in Prostate Cancer. Clinical Chemistry. 2013; 59(1):261-9.
23. Hosmer DW, Jovanovic B, Lemeshow S. Best subsets logistic regression. Biometrics. 1989; 45(4):1265-70.
24. Thao LTP, Heemskerk AD, Geskus RB, Mai NTH, Ha DTM, Chau TTH, et al. Prognostic models for 9-month mortality in tuberculous meningitis. Clinical Infectious Diseases. 2017; 66(4):523-32.
25. Allotey J, Fernandez-Felix BM, Zamora J, Moss N, Bagary M, Kelso A, et al. Predicting seizures in pregnant women with epilepsy: Development and external validation of a prognostic model. PLoS Med. 2019; 16(5):e1002802.
26. Huang J, Jiao J, Xu W, Zhao H, Zhang C, Shi Y, et al. MiR-155 is upregulated in patients with active tuberculosis and inhibits apoptosis of monocytes by targeting FOXO3. Molecular medicine reports. 2015; 12(5):7102-8.
27. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proceedings of the National Academy of Sciences. 2008; 105(30):10513-8.
28. MacLean E, Broger T, Yerliyaka S, Fernandez-Carballo BL, Pai M, Denkinger CM. A systematic review of biomarkers to detect active tuberculosis. Nature Microbiology. 2019; 4(5):748-58.

29. Zhang K, Shi H, Xi H, Wu X, Cui J, Gao Y, et al. Genome-Wide lncRNA Microarray Profiling Identifies Novel Circulating lncRNAs for Detection of Gastric Cancer. Theranostics. 2017; 7(1):213-27.
30. Jacobson KR. TUberculosis. Annals of Internal Medicine. 2017;166(3):ITC17-ITC32.
31. Chen YG, Satpathy AT, Chang HY. Gene regulation in the immune system by long noncoding RNAs. Nature Immunology. 2017; 18(9):962.
32. Thomas TY, Rajagopalan S. Tuberculosis and aging: a global health problem. Clinical Infectious Diseases. 2001; 33(7):1034-9.
33. Steingart KR, Ng V, Henry M, Hopewell PC, Ramsay A, Cunningham J, et al. Sputum processing methods to improve the sensitivity of smear microscopy for tuberculosis: a systematic review. The Lancet Infectious Diseases. 2006; 6(10):664-74.
34. He Y, Zhu Z, Chen Y, Chen F, Wang Y, Ouyang C, et al. Development and Validation of a Novel Diagnostic Nomogram to Differentiate Between Intestinal Tuberculosis and Crohn's Disease: A 6-year Prospective Multicenter Study. American Journal of Gastroenterology. 2019; 114(3):490-9.
35. Costa-Veiga A, Briz T, Nunes C. Unsuccessful treatment in pulmonary tuberculosis: factors and a consequent predictive model. The European Journal of Public Health. 2018; 28(2):352-8.
36. Sohn H, Aero AD, Menzies D, Behr M, Schwartzman K, Alvarez GG, et al. Xpert MTB/RIF testing in a low tuberculosis incidence, high-resource setting: limitations in accuracy and clinical impact. Clinical Infectious Diseases. 2014; 58(7):970-6.
37. Steingart KR, Sohn H, Schiller I, Kloda LA, Boehme CC, Pai M, et al. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. Cochrane Database of Systematic Reviews. 2014; (1):CD009593.
38. Lombardi G, Di Gregori V, Girometti N, Tadolini M, Bisognin F, Dal Monte P.

Diagnosis of smear-negative tuberculosis is greatly improved by Xpert MTB/RIF. PloS ONE. 2017; 12(4):e0176186.

Supporting information captions

S1 Appendix. Patients' recruitment, laboratory examinations, treatment management and classification.

S2 Appendix. EHR and lncRNA data collection, data-preprocessing, and features before and involving in different models.

S3 Appendix. Functional analysis for *ENST00000497872*.

S1 Fig. RNA electrophoresis, amplification curve of qRT-PCR and standard curve of control cDNA.

S2 Fig. Hierarchical clustering and volcano plot for differentially expressed lncRNA profiles in the Screening Cohort.

S3 Fig. LncRNA expressions between clinically diagnosed PTB patients and non-TB disease controls in the Selection and Validation Cohorts.

S4 Fig. AUC values of support vector machine models for clinically diagnosed PTB cases versus non-TB disease controls.

S5 Fig. qPCR analysis of *ENST00000497872* expression in BCG-infected THP-1 cells.

S1 Table. Disease controls in the present study.

S2 Table. Specific qRT-PCR primers for lncRNAs.

S3 Table. The expression of five candidate lncRNAs in the Screening Cohort.

S4 Table. The comparison of lncRNA expression between clinically diagnosed PTB patients and healthy subjects in the Selection and Validation Cohorts.

S5 Table. Short-term stability evaluation of lncRNAs in PBMC samples

S6 Table. The details of "EHR+lnRNA" logistic regression model to differentiate clinically diagnosed PTB among 300 highly suspected patient.