Title

Blood transcriptomic stratification of short-term risk in contacts of tuberculosis.

Authors

Jennifer Roe[1], Cristina Venturini[1], Rishi K Gupta[2], Celine Gurry[1], Benjamin M Chain[1], Yuxin Sun[3], Jo Southern[2], Charlotte Jackson[2], Marc C Lipman[4,5], Robert F Miller[2], Adrian R Martineau[6], Ibrahim Abubakar[2]*, Mahdad Noursadeghi[1,7]*

Affiliations

[1]Division of Infection & Immunity, University College London, United Kingdom. [2]Institute for Global Health, University College London, United Kingdom. [3]Department of Computer Science, University College London, United Kingdom. [4]UCL Respiratory Medicine, University College London. [5]Department of Respiratory Medicine, Royal Free London NHS Trust, London, United Kingdom. [6]Blizard Institute, Queen Mary University of London, United Kingdom. [7]National Institute for Health Research University College London Hospitals Biomedical Research Centre, United Kingdom.

*These authors made an equal contribution.

Corresponding author

Mahdad Noursadeghi, Division of Infection & Immunity, Cruciform Building, University College London, London WC1E 6BT, United Kingdom. Email: m.noursadeghi@ucl.ac.uk.

Summary

Discovery and validation of a three-gene blood transcription signature for incipient tuberculosis and its application for stratification of risk of disease in HIV negative contacts of active tuberculosis.

Abstract

Background

The highest risk of tuberculosis (TB) arises in the first few months after exposure. We reasoned that this risk reflects incipient disease among TB contacts. Blood transcriptional biomarkers of TB may predate clinical diagnosis, suggesting they offer improved sensitivity to detect subclinical incipient disease. Therefore, we sought to test the hypothesis that refined blood transcriptional biomarkers of active TB will improve stratification of short-term disease risk in TB contacts.

Methods

We combined analysis of previously published blood transcriptomic data with new data from a prospective HIV negative United Kingdom cohort of 333 TB contacts. We used stability selection as an alternative computational approach to identify an optimal signature for short-term risk of active TB, and evaluated its predictive value in independent cohorts.

Results

In a previously published HIV negative South African case-control study of patients with asymptomatic *Mycobacterium tuberculosis* infection, a novel three-gene transcriptional signature comprising *BATF2, GBP5 and SCARF1* achieved a positive predictive value (PPV) of 23% for progression to active TB within 90 days. In a new UK cohort of 333 HIV negative TB contacts with a median follow up of 346 days, this signature achieved a PPV of 50% (95% confidence interval: 15.7-84.3) and NPV of 99.3% (97.5-99.9). By comparison, peripheral blood interferon gamma release assays in the same cohort achieved a PPV of 5.6% (2.1-11.8).

Conclusion

This blood transcriptional signature provides unprecedented opportunities to target therapy among TB contacts with greatest risk of incident disease.

Keywords

Blood transcriptome; biomarker; tuberculosis; diagnosis.

Introduction

The causative agent of tuberculosis, *Mycobacterium tuberculosis* (Mtb) is an obligate pathogen [1]. In the absence of an effective vaccine, the earliest possible identification of disease offers the best strategy to reduce onward transmission. Early treatment also offers the opportunity to adopt shorter and simpler treatment regimens, to limit pathology and reduce TB-associated morbidity. This rationale, and the fact that risk of TB is highest within the first few months after significant exposure [2,3] provide the basis for screening close contacts of patients with active TB, in order to identify prevalent disease and offer preventative treatment to individuals with asymptomatic infection.

Detection of immune memory for Mtb, using interferon gamma release assays (IGRA) or the tuberculin skin test (TST), is widely used to identify infected individuals. These tests have poor sensitivity, estimated as 50-85% for identifying contacts who progress to TB. In addition, they have poor positive predictive value (PPV), estimated to be <5% for two-year cumulative TB incidence or 1-1.5 per 100 person years [4–6]. In the absence of clinical disease, the ability to stratify differential risk of progression to TB remains extremely limited, leading to unnecessary treatment of people at low risk. The fact that overall risk is low means that significant numbers of individuals refuse the offer of treatment, thereby undermining the overall effectiveness of contact tracing to prevent incident TB.

Blood transcriptomic biomarkers have emerged as a sensitive approach for identification of active TB [7–11]. Importantly, these may predate conventional clinical diagnosis [12,13]. In a South African cohort of HIV negative individuals with latent TB infection [12], the sensitivity of a 16-gene blood transcriptional signature that discriminated individuals who progressed to a diagnosis of active TB from those who did not, improved as the time interval between sampling and diagnosis reduced. These data suggest that blood transcriptional signatures can identify pre-symptomatic incipient disease. The fact that blood transcriptional changes may not be suitable for stratification of long-term risk was highlighted in a second multi-cohort African study of house-hold contacts that specifically excluded patients who progressed to active TB within three months of enrolment [13]. In that study, a four-gene blood transcriptional signature discriminated between progressors and non-progressors with only modest receiver operating characteristic area under the curve (AUC) of 0.69 and positive predictive value (PPV) of 3%, equivalent to that of IGRAs or TST.

These studies focussed on deriving blood transcriptional signatures for prospective risk of incident TB for up to two years. However, they may not represent the most sensitive biomarkers of preclinical incipient disease underlying the short-term risk of TB amongst contacts. We recently reported that *BATF2* gene expression provided a single blood transcript that accurately discriminated active from latent TB infection [11]. *BATF2* was identified by comparing transcriptional profiles from patients with active and treated TB. Of note, *BATF2* is a component of the 16-gene signature reported by Zak et al [12]. In the present study we tested the ability of *BATF2* on its own, to identify incipient TB disease in the Zak cohort. In addition, we sought to improve on the predictive value of *BATF2*, using stability selection as an alternative computational approach to identifying discriminating features in high dimensional data. We then validated our findings in two independent data sets. First, the previously published South African cohort of HIV negative individuals with latent TB infection [12] and then a new UK cohort of TB contacts.

Methods

Analysis of previously published data

Raw sequencing data from Zak et al [12] was pseudoaligned to the human transcriptome (Ensembl Human GRCh38) using Kallisto [14]. The abundance for protein coding RNA was expressed as $\log_2$-transformed transcripts per million (TPM) [15]. Microarray data from Roe et al [11] were used as normalised $\log_2$-transformed data. RNAseq and microarray data were standardised by subtracting the mean and dividing by the standard deviation of each data set. Gene expression data from each study were annotated with Human Genome Organisation Nomenclature Committee (HGNC) gene symbols.

Blood transcriptional profiling of a UK cohort of TB contacts.

Close contacts of patients with TB were invited to participate (Supplementary methods). The study was approved by the UK National Research Ethics Service (reference: 14/EM/1208). All participants provided written informed consent. At enrolment, IGRAs were done using the QuantiFERON-TB Plus assay (Qiagen, Germany), and peripheral blood RNA was collected into Tempus^{TM} tubes for transcriptional profiling by RNA sequencing (Supplementary methods). These data are available in ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under the accession number E-MTAB-6845. At the end of the study, participants who progressed to active TB were identified by linkage with the national electronic TB register as previously described [17]. Local case notes were reviewed in order to identify individuals who had received preventative treatment.

Stability selection to refine the optimal gene signature for incipient TB.

We combined stability selection with Support Vector Machines (SVM) [18] (Supplementary methods), to identify a ranking of the individual transcripts independent of other genes, which discriminated patients with active TB pre-treatment from those who had been successfully treated [11]. Once we had selected a consistent sub-set of discriminating genes using stability selection, we used these genes to train a traditional L2 Norm learning SVM using kernlab with a linear kernel in the R statistical computing platform [19] as previously described [11]. We used this SVM to generate decision values for each data set tested. ROC curves were plotted in GraphPad Prism version 7 (GraphPad Software, La Jolla, California, USA) and used to calculate the ROC AUC. The Youden Index for each ROC curve was derived from the sum of sensitivity and specificity–1 [20], and the predictive value of each transcriptional signature was estimated using Bayesian conditional probabilities [21]. 95% confidence intervals are provided for each measure of test performance.

Results

Positive predictive value of BATF2 blood transcript levels for short-term risk of incident TB.

We first sought to test the hypothesis that elevated blood levels of *BATF2* gene expression identified individuals with subclinical incipient disease leading to incident TB in the short-term. We compared *BATF2* blood transcript levels in all samples from individuals with latent TB in the Zak cohort [12], who did not progress to active TB (N=48), with samples from individuals who progressed to active TB within 90 days (N=12), 91-360 days (N=29), and after an interval of greater 360 days (N=25). In this analysis, we ensured that samples were not incorrectly allocated to the non-progressor group because of inadequate follow up, by restricting the non-progressor group to include only samples from individuals who had more than 12 months follow up after sample was collected. Consistent with our hypothesis, *BATF2* blood transcript levels were highest in the group of individuals who progressed to active TB within 90 days (Figure 1A). Accordingly, discrimination of individuals who progress to TB from non-progressors using *BATF2* levels, achieved the highest ROC AUC of 0.93 (0.86-1) in the group who progressed to TB within 90 days (Figure 1B). The threshold for *BATF2* levels which discriminated this group of progressors from non-progressors with the greatest accuracy was identified by the ROC curve Youden index. This threshold achieved a sensitivity of 0.83 (0.52-0.98) and specificity of 0.92 (0.83-0.99) (Figure 1C), giving a positive likelihood ratio of 13.3 (4.3-41.1). At the same threshold, the sensitivity of elevated *BATF2* levels to identify individuals who progressed to TB after 90 days was reduced to 0.52 (0.36-0.74), in the 91-360 days interval, and to 0.24 (0.12-0.49) after 360 days. The cumulative TB risk in this cohort approximated to 1.5% [12]. Using this pre-test probability, we estimated the PPV of elevated *BATF2* levels for diagnosis of TB within 90 days to be 13% (Figure 1C). These were comparable to those of disease within 90 days using the 16-gene signature described by Zak et al, which achieved a ROC AUC of 0.94, (0.89-1) (Supplementary Figure 1).

A refined three-gene signature to predict active TB within 90 days.

In the clinical cohort described by Zak et al, their 16-gene signature offered no advantage to measuring *BATF2* transcripts alone to discriminate progressors within 90 days from non-progressors. Nonetheless, we hypothesised that the addition of selected genes may further improve the performance of *BATF2* alone as a biomarker of incipient TB. We reasoned that the most discriminating transcriptional biomarker may be different among subsets of cases. Therefore, a combination of the genes most frequently ranked top in multiple subsamples of the data may give the optimal gene signature for incipient TB. This approach to feature selection in high dimensional data has been called stability selection [18]. We used stability selection to rank the transcripts that best discriminated subsets of our previously published active and treated TB cases [11]. Using this ranking, we trained an SVM model to discriminate active and treated TB with a cumulative number of genes and tested how accurately each SVM model correctly classified progressor and non-progressor patients in the Zak cohort (Supplementary Figure 2B). The most accurate classification was achieved by the top three genes comprising *BATF2*, *GBP5* and *SCARF1* (Supplementary Figure 2C). We represented the three-gene score for each individual in the Zak cohort, as the distance from the separating hyperplane that discriminates between two classes in the SVM model (Figure 2A). The three-gene SVM model discriminated between non-progressors and those who progressed to TB within 90 days with a ROC AUC of 0.96, (0.92-1) (Figure 2B). At the Youden Index, this ROC curve generated a sensitivity of 0.83 (0.52-0.98) and specificity of 0.96 (0.83-0.99). This was equivalent to a positive likelihood ratio of 20 (5-79.5) and a PPV for active TB within 90 days of 23%, given a prior probability of 1.5% (Figure 2C).

Predictive value of the three-gene signature for active TB in a new UK cohort of TB contacts.

The predictive value of a TB biosignature based on *BATF2*, *GBP5* and *SCARF1* described above was obtained from estimates of sensitivity, specificity derived from case-control data, and estimates of prior probability. A prospective, independent observational cohort was required to confirm these findings. In addition, although our three-gene signature was discovered in data derived from a UK cohort of patients with active TB and validated in a South African cohort of patients with latent TB, additional validation in a further independent UK cohort at risk of active TB was necessary to extend the evidence for its generalisability. Therefore, we obtained blood transcriptomic data from a new observational HIV negative cohort of 333 close contacts of cases of active TB, representing a group at highest risk of developing disease in the short-term (Table 1). Median follow-up of the cohort was 346 days (IQR: 250-450). A total of 6 participants in the cohort progressed to a diagnosis of TB disease 3-342 days after recruitment to the study (Table 2).

We used the novel three-gene model to calculate a decision score as described above, for each patient in the UK TB contacts cohort. First, we sought to define the distribution of three-gene scores in 192 IGRA negative contacts as a control population among TB contacts with low risk of developing disease. This group was younger, had fewer non-UK born individuals and fewer household contacts of TB, compared to the IGRA positive individuals, reflecting known risk factors for acquisition of Mtb infection (Table 1). The three-gene scores among the IGRA negative group showed a parametric distribution. Therefore, we used standard score of two (Z_2) to represent the 97.7th percentile, and three (Z_3) to represent the 99.9th percentile as thresholds for an elevated three gene score (Figure 3A). We then compared the distribution of three-gene scores among all the IGRA positive contacts, and tested the hypothesis that the three-gene score stratifies differential risk of incident TB in this cohort. Among individuals in the cohort who did not receive preventative therapy, we compared the incident TB rate per 100 person years in all IGRA positive and IGRA negative contacts with three-gene scores greater than the Z_2 and Z_3 thresholds (Figure 3B).

Using the Z_2 threshold, TB incidence rate among individuals with a low three-gene score was 0.76 per 100 person years (0.19-3.05). Among 19 individuals with a high three-gene score, the TB incidence rate was 27.7 per 100 person years (10.4-73.8). The incidence rate ratio (incidence of TB for those with a high three-gene score compared to those with a low three-gene score) was 36.3 (6.6-198.1). Using the Z_3 threshold, TB incidence rate among individuals with a low three-gene score was 0.74 per 100 person years (0.18-3.0). Among 8 individuals with a high three-gene score, the TB incidence rate was 72.0 per 100 person years (27.0-191.8). The incidence rate ratio (incidence of TB for those with a high three-gene score compared to those with a low three-gene score) was 97.4 (17.8-532.0). By comparison, incidence rates were 5.8 per 100 person years (2.6-13.0) and 0 per 100 person years among IGRA-positive and -negative contacts, respectively. Overall, at the Z_2 threshold, the three-gene score achieved a PPV of 21.1% (6.1-46.6), and positive likelihood ratio of 13 (6.2-27.6). At the Z_3 threshold, the three-gene score achieved a PPV of 50% (15.7-84.3), and positive likelihood ratio of 48.8 (15.8-150.5). Both these thresholds for the three-gene score achieved NPV of 99.3% (97.5-99.9 for the Z_3 threshold and 97.4-99.9 for the Z_2 threshold), and negative likelihood ratio of 0.35 (0.11-1.1) for cumulative incident TB within two years. The IGRA result achieved a PPV of 5.6% (2.1-11.8) and an NPV of 100% (98.1-100).

Discussion

The present study was based on the premise that blood transcriptional biomarkers of active TB predate clinical presentation of disease and consequently serve as biomarkers of incipient TB. We confirmed this hypothesis in data from a South African case-control study, using a single transcriptional biomarker for active TB, BATF2. The addition of two further genes generated a three-gene signature derived from independent data, which further enhanced the discrimination between short-term progressors and non-progressors in the case-control study. Finally, we validated our findings in a new UK cohort study.

We provide the first proof of concept data for a blood transcriptional signature that predicts the short-term risk of disease in TB contacts with substantially greater PPV than is achieved in current practice using IGRAs. The three-gene signature in the present study was derived from comparison of patients with pulmonary TB before and after treatment, but predicted cases of extrapulmonary disease in the UK cohort, consistent with the hypothesis that blood transcriptional signatures of TB are not specific to different sites of disease. Our blood transcriptional signature predicted disease progression in four of six close contacts. The patient who progressed within three days, was symptomatic at enrolment, but the others had no symptoms of active disease, supporting the underlying hypothesis that this signature predates clinical presentation of disease. The potential clinical impact of this approach is to enable more precise targeting of preventative antimicrobial TB treatment. On the basis of the differential PPVs among contacts of active TB, risk stratification using the blood transcriptional signature may reduce the number needed to treat to <5, compared to >20 using IGRAs. In addition, the better PPV of the blood transcriptional signature may be expected to incentivise increased treatment acceptance and completion rates. Taken together, these effects have the potential to transform the efficiency and therefore scalability of contact tracing as part of a TB control program.

In the South African case-control study, the sensitivity of the three-gene signature to identify individuals who progressed to TB reduced as interval time to TB increased. Consistent with this, in the UK cohort study, the three-gene signature failed to identify two of six contacts that progressed to active TB, with the longest disease-free intervals, both greater than six months. These data highlight the reduced sensitivity for long-term risk of incident disease. Therefore, interval follow-up measurements for IGRA positive contacts may be needed to prevent cases beyond the first six months. Thereafter the numbers needed to screen and treat with preventative therapy for the very small residual long-term risk of disease, may not justify the economic cost and risk of drug toxicity. To illustrate this, in a large observational cohort of 4,861 TB contacts with median follow-up 2.9 years, 52% and 73% of incident TB cases occurred within six and 12 months of contact screening, respectively [6]. Therefore, a strategy of testing recent TB contacts with blood transcriptional biomarkers at baseline and after a six-month interval may identify up to three quarters of incident TB cases among contacts.

The major limitation of our study is the low frequency of progressive disease in the UK cohort. Despite the fact that TB contacts have the highest short-term risk of disease, the absolute two-year cumulative incidence of disease is low, necessitating very large scale cohorts to definitively assess the correlates of progression [6]. Consequently, in the present study, the confidence intervals of the PPV for the three-gene signature are wide, albeit significantly better than IGRA which achieves a PPV of <5% in large scale studies. The impact of HIV co-infection is also untested. In addition, it is clear that a high three-gene score is evident in individuals who do not develop active TB. This observation may reflect spontaneous resolution of subclinical TB among some individuals, inadequate follow up of these patients, or a lack of specificity for TB. Therefore, further assessments of potential confounding of this transcriptional signature by HIV co-infection, other co-morbidities and longitudinal studies of the expression of this signature with and without TB treatment are required. Notwithstanding these limitations, we propose that the three-gene blood transcriptional signature offers exciting new opportunities to transform risk stratification for progression to TB disease among contacts of active TB. This application of the blood transcriptional signature is consistent with proposals from the World Health Organisation for a non-sputum test with sensitivity and specificity of >75%, that predicts risk of disease in patients with incipient TB [22]. Our findings pave the way for extended validation of the signature, “head to head” comparison of the performance of the different published blood transcriptional signatures of TB and development of the technology to allow scale-up of near-patient testing.

Author contributions

JR, ML, RFM, ARM, IA and MN conceived of the study. RG, JS and CJ undertook sample and data collection, JR and CG undertook sample processing. JR, CV, RG, BMC, YS and MN undertook data analysis. JR, RG, IA and MN wrote the manuscript with input from all the authors. *These authors made an equal contribution.

Funding support

This work was supported by an MRC Fellowship to JR (MR/L001756/1), an NIHR fellowship to RG (DRF-2018-11-ST2-004), a Wellcome Trust Investigator Award to MN (207511/Z/17/Z), the British Lung Foundation (TB05/11), the UCL Technology Fund, the Rosetrees Trust and the NIHR Biomedical Research Centre Funding to University College Hospitals NHS Foundation Trust and University College London. IA was funded by NIHR (SRF-2011-04-001; NF-SI-0616-10037), MRC, UK Department of Health and the Wellcome Trust.

Conflict of interest statement

JR, ARM and MN have a patent application pending in relation to blood transcriptomic biomarkers of tuberculosis. RM reports personal fees from Gilead, outside the submitted work. All other authors declare no other conflict of interest.

References

1. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol **2018**; 16:202.

2. Sloot R, Schim van der Loeff MF, Kouw PM, Borgdorff MW. Risk of tuberculosis after recent exposure. A 10-year follow-up study of contacts in Amsterdam. Am J Respir Crit Care Med **2014**; 190:1044–1052.

3. Behr MA, Edelstein PH, Ramakrishnan L. Revisiting the timetable of tuberculosis. BMJ **2018**; 362:k2738.

4. Zellweger J-P, Sotgiu G, Block M, et al. Risk Assessment of Tuberculosis in Contacts by IFN-γ Release Assays. A Tuberculosis Network European Trials Group Study. Am J Respir Crit Care Med **2015**; 191:1176–1184.

5. Haldar P, Thuraisingam H, Patel H, et al. Single-step QuantiFERON screening of adult contacts: a prospective cohort study of tuberculosis risk. Thorax **2013**; 68:240–246.

6. Abubakar I, Drobniewski F, Southern J, et al. Prognostic value of interferon-γ release assays and tuberculin skin test in predicting the development of active tuberculosis (UK PREDICT TB): a prospective cohort study. Lancet Infect Dis **2018**;

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

8. Anderson ST, Kaforou M, Brent AJ, et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. N Engl J Med **2014**; 370:1712–1723.

9. Maertzdorf J, McEwen G, Weiner J, et al. Concise gene signature for point-of-care classification of tuberculosis. EMBO Mol Med **2015**;

10. Sweeney TE, Khatri P. Blood transcriptional signatures for tuberculosis diagnosis: a glass half-empty perspective - Authors' reply. Lancet Respir Med **2016**; 4:e29.

11. Roe JK, Thomas N, Gil E, et al. Blood transcriptomic diagnosis of pulmonary and extrapulmonary tuberculosis. JCI Insight **2016**; 1. Available at: https://insight.jci.org/articles/view/87238. Accessed 6 October 2016.

12. Zak DE, Penn-Nicholson A, Scriba TJ, et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. Lancet Lond Engl **2016**;

13. Suliman S, Thompson E, Sutherland J, et al. Four-gene Pan-African Blood Signature Predicts Progression to Tuberculosis. Am J Respir Crit Care Med **2018**;

14. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol **2016**; 34:525–527.

15. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci Theor Den Biowissenschaften **2012**; 131:281–285.

16. Treating latent TB | Information for the public | Tuberculosis | Guidance | NICE. Available at: https://www.nice.org.uk/guidance/ng33/ifp/chapter/treating-latent-tb.

17. Aldridge RW, Shaji K, Hayward AC, Abubakar I. Accuracy of Probabilistic Linkage Using the Enhanced Matching System for Public Health and Epidemiological Studies. PloS One **2015**; 10:e0136179.

18. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Ser B Stat Methodol **2010**; 72:417–473.

19. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2015. Available at: http://www.R-project.org/.

20. Youden WJ. Index for rating diagnostic tests. Cancer **1950**; 3:32–35.

21. López Puga J, Krzywinski M, Altman N. Points of significance: Bayes' theorem. Nat Methods **2015**; 12:277–278.

22. World Health Organization. Consensus meeting report: development of a target product profile (TPP) and a framework for evaluation for a test for predicting progression from tuberculosis infection to active disease. 2017. Available at: https://apps.who.int/iris/handle/10665/259176.

Tables

Table 1. Summary characteristics of UK TB contacts cohort

		All (n=333)		IGRA +ve (n=141)		IGRA –ve (n=192)		P value
Age	Median (IQR), years	33 (26-47)		40.5 (29-51)		32 (25-43)		<0.05
		n	%	n	%	n	%	
Gender	Male	173	52.0	77	54.6	96	50.0	ns
	Female	155	46.5	62	44.0	93	48.4	ns
	NA	5	1.5	2	1.4	3	1.6	ns
Ethnicity	White	73	21.9	24	17.0	49	25.5	ns
	Indian, Pakistani or Bangladeshi	120	36.0	51	36.2	69	35.9	ns
	Black African	56	16.8	32	22.7	24	12.5	<0.05
	Mixed	64	19.2	29	20.6	35	18.2	ns
	Other	10	3.0	2	1.4	8	4.2	ns
	NA	10	3.0	3	2.1	7	3.6	ns
Country of Birth	Non-UK	259	77.8	126	89.4	133	69.3	<0.05
	UK	68	20.4	12	8.5	56	29.2	<0.05
	NA	6	1.8	3	2.1	3	1.6	ns
TB contact type	Household	218	65.5	103	73.0	115	59.9	<0.05
	Non-Household	77	23.1	25	17.7	52	27.1	<0.05
	NA	38	11.4	13	9.2	25	13.0	ns
Social risk factor	Yes	35	10.5	13	9.2	22	11.5	ns
	No	298	89.5	128	90.8	170	88.5	ns
	NA	0	0	0	0	0	0	ns
Diabetes	Yes	27	8.1	16	11.3	11	5.7	ns
	No	296	88.9	122	86.5	174	90.6	ns
	NA	10	3	3	2.1	7	3.6	ns
Smoking	Current	76	22.8	32	22.7	44	22.9	ns
	Previous	43	12.9	18	12.8	25	13	ns
	No	21.	63.1	89	63.1	121	63	ns
	NA	4	1.2	2	1.4	2	1	ns
Preventative therapy	Yes	34	10.2	34	24.1	0	0	n/a
	No	289	86.8	97	68.8	192	100	n/a
	NA	10	3.0	10	7.1	0	0	n/a

Abbreviations: IQR- interquartile range; NA- not available; n/a- not applicable. For statistical tests, age was compared by a Mann Whitney U test and categorical variables were compared using Chi squared tests. Social risk factors included history of homelessness, imprisonment or harmful drug use.

Table 2. Selected characteristics of patients who progressed to a diagnosis of TB in the UK cohort study

Disease free interval (days)	Age (years)	Gender	Ethnicity	IGRA	Site of disease	Culture confirmed	Symptoms at screening	Screening chest radiograph
3	32	Female	Indian	Positive	Uterus	Yes	Yes	Normal
90	16	Male	Black African	Positive	Pleural	No	No	Normal
137	21	Male	Mixed	Positive	Lymph node	No	No	Normal
210	21	Female	Mixed	Positive	Lymph node	No	No	Paratracheal adenopathy
272	52	Male	Indian	Positive	Pleural	Yes	No	Normal
342	27	Female	Black African	Positive	Pleural	No	No	Normal

Figure Legends

Figure 1

Identification of incipient TB by measurement of blood BATF2 transcript levels.

(A) *BATF2* transcript levels (TPM) are shown for blood samples from all patients in the Zak cohort (with at least 12 months follow up after the time of sampling) who did not progress to TB (NP), and for samples from patients who progressed to a diagnosis of TB within the time intervals indicated. **(B)** Receiver operating characteristic (ROC) curves for discriminating between NP and progressors in each time interval shown using the *BATF2* transcript level. **(C)** Positive predictive value for a diagnosis of TB (PPV_{TB}) for patients who progress to TB within 90 days, using sensitivity and specificity values derived from the optimal Youden Index (dotted line in (A)) of the ROC curve in (B) and a range of pre-test (PT) probabilities. Arrows highlight the PPV_{TB} of 13% for PT probability of 1.5%.

Figure 2

Identification of incipient TB by a novel 3-gene model incorporating blood transcript levels of BATF2, GBP5 and SCARF1.

(A) Three-gene scores derived from the SVM model to discriminate between active and treated TB, are shown for blood samples from all patients in the Zak cohort (with at least 12 months follow up) who did not progress to TB (NP), and for samples from patients who progressed to a diagnosis of TB within the time intervals indicated. **(B)** ROC curves for discriminating between NP and progressors in each time interval shown using the three-gene scores. **(C)** Positive predictive value for a diagnosis of TB (PPV_{TB}) for patients who progress to TB within 90 days, using sensitivity and specificity values derived from the optimal Youden Index of the ROC curve in (B) and a range of pre-test (PT) probabilities. Arrows highlight the PPV_{TB} of 23% for PT probability of 1.5%.

Figure 3

Blood transcriptomic 3-gene score at recruitment in contacts of active TB.

(A) Frequency distribution of 3-gene scores in IGRA negative contacts of active TB showing threshold (dashed lines) for identification of a high 3-gene score based on the mean+2 SD (Z_2) or +3 SD (Z_3) of the scores among IGRA negative cases. **(B)** Individual 3-gene scores for untreated IGRA positive and negative contacts who developed active TB or remained healthy on follow up.

Figure 1

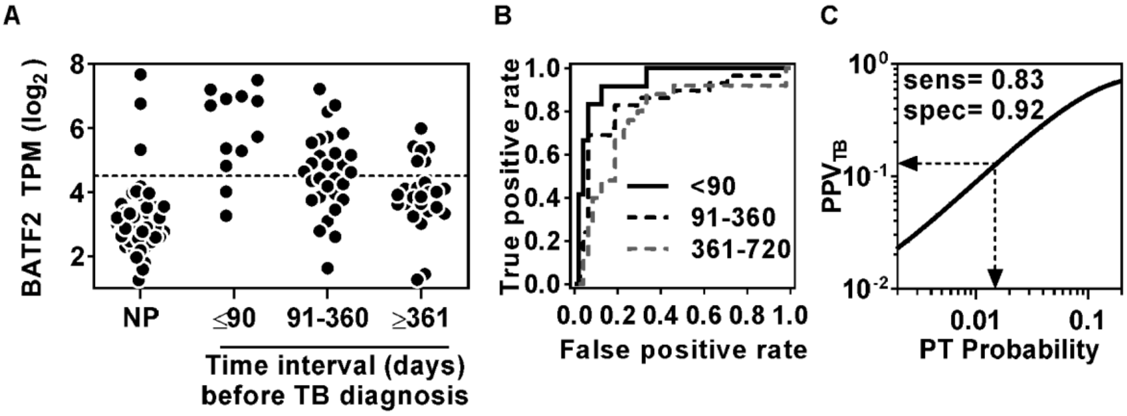

A
B
C
BATF2 TPM ($\log_2$)
8
6
4
2
NP
≤90
91-360
≥361
Time interval (days) before TB diagnosis
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
<90
91-360
361-720
0.0 0.2 0.4 0.6 0.8 1.0
False positive rate
PPV_{TB}
10^0
10^{-1}
10^{-2}
sens= 0.83
spec= 0.92
0.01
0.1
PT Probability

Figure 2

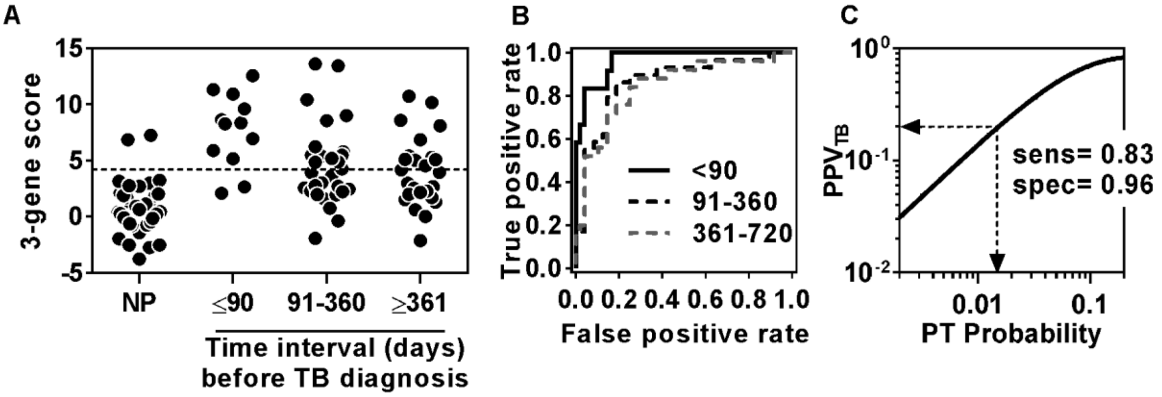
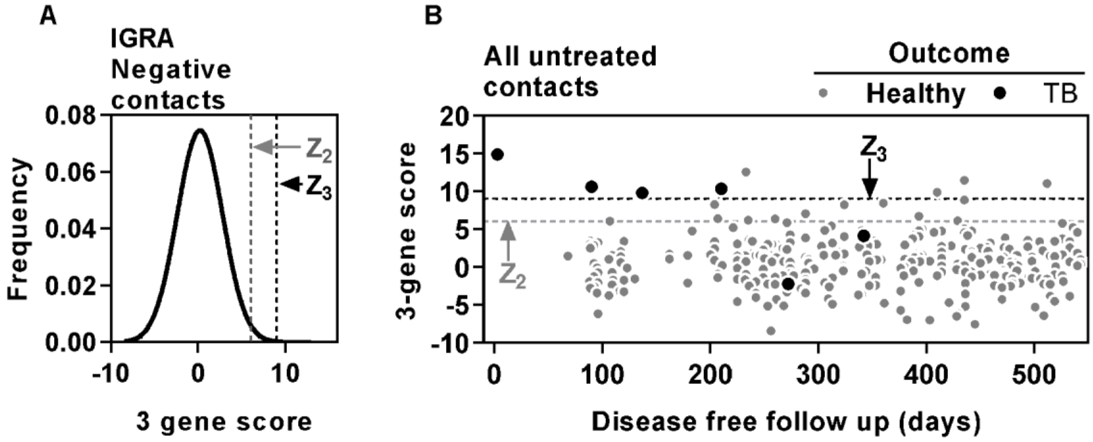

A
15
10
5
0
-5
3-gene score
NP
≤90
91-360
≥361
Time interval (days)
before TB diagnosis
B
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0 0.2 0.4 0.6 0.8 1.0
False positive rate
<90
91-360
361-720
C
10^0
10^{-1}
10^{-2}
PPV_{TB}
sens= 0.83
spec= 0.96
0.01
0.1
PT Probability

Figure 3


A
IGRA
Negative
contacts
Frequency
0.08
0.06
0.04
0.02
0.00
-10
0
10
Z_2
Z_3
3 gene score
B
All untreated
contacts
Outcome
● Healthy
● TB
3-gene score
20
15
10
5
0
-5
-10
Z_3
Z_2
0
100
200
300
400
500
Disease free follow up (days)