scientific reports

Check for updates

OPEN

Blood RNA signatures predict recent tuberculosis exposure in mice, macaques and humans

Russell C. Ault[1,2,3,4], Colwyn A. Headley[1,2,3], Alexander E. Hare[3,4], Bridget J. Carruthers[2], Asuncion Mejias[5] & Joanne Turner[1,2]✉

Tuberculosis (TB) is the leading cause of death due to a single infectious disease. Knowing when a person was infected with *Mycobacterium tuberculosis* (*M.tb*) is critical as recent infection is the strongest clinical risk factor for progression to TB disease in immunocompetent individuals. However, time since *M.tb* infection is challenging to determine in routine clinical practice. To define a biomarker for recent TB exposure, we determined whether gene expression patterns in blood RNA correlated with time since *M.tb* infection or exposure. First, we found RNA signatures that accurately discriminated early and late time periods after experimental infection in mice and cynomolgus macaques. Next, we found a 6-gene blood RNA signature that identified recently exposed individuals in two independent human cohorts, including adult household contacts of TB cases and adolescents who recently acquired *M.tb* infection. Our work supports the need for future longitudinal studies of recent TB contacts to determine whether biomarkers of recent infection can provide prognostic information of TB disease risk in individuals and help map recent transmission in communities.

Tuberculosis (TB) is the leading killer due to a single infectious disease, causing over 1 million deaths per year[1]. Despite renewed efforts to combat the TB epidemic, the current decline in TB incidence of 1.5% per year has fallen far short of the needed 4–5% annual decline to meet the 2020 goals for the World Health Organization's (WHO) End TB Strategy[2]. While approximately ¼ (1.7 billion) of the world's population has been infected with its causative agent *Mycobacterium tuberculosis* (*M.tb*), only 5 to 10% of infected individuals will develop active TB disease during their lifespan, with the remainder controlling the infection in a state known as latent TB infection (LTBI)[3,4]. Recent global workshops have reemphasized targeting transmission of TB as critical to accelerating efforts to reduce the burden of TB disease throughout the world[5,6]. Two critical areas for understanding and preventing TB transmission are knowing where and when transmission occurs, and preventing infected individuals from progressing to active TB disease and thereafter transmitting the bacteria via the airborne route[7,8].

Historically, successful control of TB in nations has followed from a reduction in transmission to very low levels[7,9]. Studies of close contacts, and in particular household contacts, of active TB cases are a critical tool for identifying new active TB cases from recent transmission and targeting therapy for preventing both subsequent disease and transmission. However, in high incidence countries where most of the burden of disease resides, more than 80% of TB transmission occurs outside of the home[10,11]. Genotyping *M.tb* isolates from active TB cases coupled with comparative genomic analysis has permitted population-level identification of hotspots of localized transmission, but these data are mostly available retrospectively and thus do not allow real-time monitoring of TB transmission in a community, particularly in areas of high incidence[12]. It thus remains unknown whether with current methods TB transmission can be appreciably disrupted in high incidence settings. This is in contrast to low incidence settings where both contact studies and targeting specific higher incidence communities have been effective[13].

Recent infection is the single strongest clinical risk factor for developing active TB disease in immunocompetent persons, who comprise the vast majority of LTBI and active TB cases[14–19]. However, time since exposure or infection is very difficult to ascertain in the clinical setting, and its estimate is often unreliable[20]. Moreover, there are no known validated biomarkers of recent exposure or infection beyond conversion on a tuberculin skin test (TST) or IFN-γ release assay (IGRA), which requires longitudinal sampling. At the same time, treating all

[1]Texas Biomedical Research Institute, San Antonio, TX, USA. [2]Department of Microbial Infection and Immunity, Ohio State University, Columbus, OH, USA. [3]Biomedical Sciences Graduate Program, Ohio State University, Columbus, OH, USA. [4]Medical Scientist Training Program, Ohio State University, Columbus, OH, USA. [5]Center for Vaccines and Immunity, Abigail Wexner Research Institute at Nationwide Children's Hospital, Columbus, OH, USA. ✉email: joanneturner@txbiomed.org

LTBI + individuals in areas of high TB incidence to prevent the development of active TB is not feasible and would entail unnecessary risk to the vast majority of LTBI + individuals who will never develop disease. Mass LTBI treatment also poses a theoretical risk of promoting drug resistant *M.tb* strains. Prospective gene expression-based (RNA) signatures of risk of developing active TB disease have been recently identified for LTBI + adolescents and adult healthy household contacts (HHCs)[21–23]. While the positive predictive value of these RNA signatures of risk of active TB is higher than TST/IGRAs, they are still significantly less than ideal: to prevent one case of active TB, ~ 37-64 LTBI + people not at risk need to be treated (*vs.* ~ 85 for TST/IGRA)[21–24]. It is currently unknown whether these RNA signatures correlate with time since infection. Importantly, their positive predictive value for TB progression and the number needed to treat could be dramatically improved if combined with accurate knowledge of time since infection in the same individual.

Building on this prior work, we assess RNA expression as a potential biomarker of recent exposure or infection with *M.tb*. Using our murine data, and recently published studies in cynomolgus macaques and humans[21,22,25,26], we show for the first time that RNA expression predicts recent infection/exposure in all three species. Moreover, in both macaques and humans, these RNA signatures of recent infection/exposure are independent of the recently identified signatures of individual prospective TB disease risk. Our work supports the need for future longitudinal studies of recent TB contacts to determine whether biomarkers of recent infection can provide prognostic information of TB disease risk in individuals and help map recent transmission in communities.

Results

Blood genome-wide RNA expression accurately discriminates early vs. late *M.tb* infection time periods in C57BL/6 mice. While several published studies have made genome-scale measurements of the in vivo host response to *M.tb* at several time points in mice[27–29], none have addressed the question of whether these parameters can predict infection time point. To determine whether it is possible to predict time since *M.tb* infection in mice via a blood RNA signature, we measured genome-wide RNA expression in whole blood in C57BL/6 mice following low dose aerosol *M.tb* infection. Mouse cohorts were sacrificed every month post-infection for 5 months (n = 4 per time point) along with age-matched uninfected C57BL/6 mice (n = 1–2 per time point). While *M.tb* colony forming units (CFUs) were not measured, it is well characterized that in this mouse strain lung bacterial burden increases exponentially from the day of *M.*tb infection until the peak of the adaptive immune response in the lungs at 1 month post-infection, thereafter remaining stable for approximately 300 days[30–32]. Thus, lung CFUs do not predict time since infection in this model after one month post-infection.

Principle component analysis (PCA) of our whole dataset revealed that the blood transcriptional state of *M.tb* infection during the first five months was distinct from that of uninfected mice, with uninfected and infected mice being entirely separable along the 1st principle component (21.8% of data variance; Fig. 1A). When we performed PCA on only *M.tb* infected mice, we found that early (30–60 days) and late (90–150 days) time periods were transcriptionally distinct, being separable along the 1st and 2nd principle components (18.8% and 16.5% of data variance, respectively; Fig. 1B). Only 1 mouse from the 60 day time point clustered with the late time period along the 2nd principle component.

To find a predictive RNA signature of time since *M.tb* infection, we used the Random Forest Classifier algorithm, without hyperparameter tuning due to low sample size, to predict early (30–60 days) versus late (90–150 days) infection time period. This analysis treated time as a binary variable. Using out-of-bag predictions (approximately threefold cross-validation) to obtain an unbiased estimate of predictive performance, we found that we could predict early versus late infection time period with 0.99 area under the curve (AUC) (95% CI 0.96–1.00, $P = 1.6 \times 10^{-5}$; 87.5% sensitivity, 91.7% specificity for early infection; Fig. 1C). To assess whether each month post-infection could be predicted accurately, thus treating time as a continuous variable, we performed Random Forest Regression with threefold cross-validation and confirmed that days 30 and 60 were predicted to be earlier time points than days 90–150 (Fig. 1D). Days 90–150 were not resolved. Low group size precludes confident quantification of the degree to which days 30 and 60 can be separated. Probes used in these models as well as their feature importance for the regression model are shown in Table S1. We further assessed our classifier model's predictions on uninfected mice, which showed their similarity to the late (90–150 days) time points (Figure S1). Taken together, these data indicate that we can broadly discriminate early and late *M.tb* infection in this cohort of C57BL/6 mice based on the whole blood transcriptomic response. Treating time as a binary variable (Fig. 1C) or as a continuous variable (Fig. 1D) gave similar results.

Blood RNA signature discriminates early versus late *M.tb* infection time periods in cynomolgus macaques. While inbred mice are a suitable model for studying molecular components of the immune response to *M.tb*, they do not replicate the variable clinical outcomes of *M.tb* infection in humans. Cynomolgus macaques, an outbred non-human primate model for TB, do exhibit heterogeneity in clinical outcomes, with approximately half of macaques progressing to symptomatic active TB disease that can be verified radiologically and bacteriologically within the first 6 months of infection, and the remainder controlling the infection in a latent state[33,34]. The lung pathology of *M.tb* infection in cynomolgus macaques also better replicates several features of human lung pathology than mice[34].

To determine whether our findings in the murine model translated to the more human-like cynomolgus macaque model of *M.tb* infection, we mined publicly available data from a longitudinal study of *M.tb* infection in macaques[25]. In that study, cynomolgus macaques were infected with a low dose of *M.tb* in the lung, and their blood was sampled at 11 time points post-infection and 2 time points pre-infection for genome-wide RNA expression analysis. Importantly, while the study's authors provided a broad, unsupervised analysis of their data according to time periods of infection, they did not assess our hypothesis that blood genome-wide RNA expression predicts time period or time point post-infection[25]. To test our hypothesis and allow comparison with our

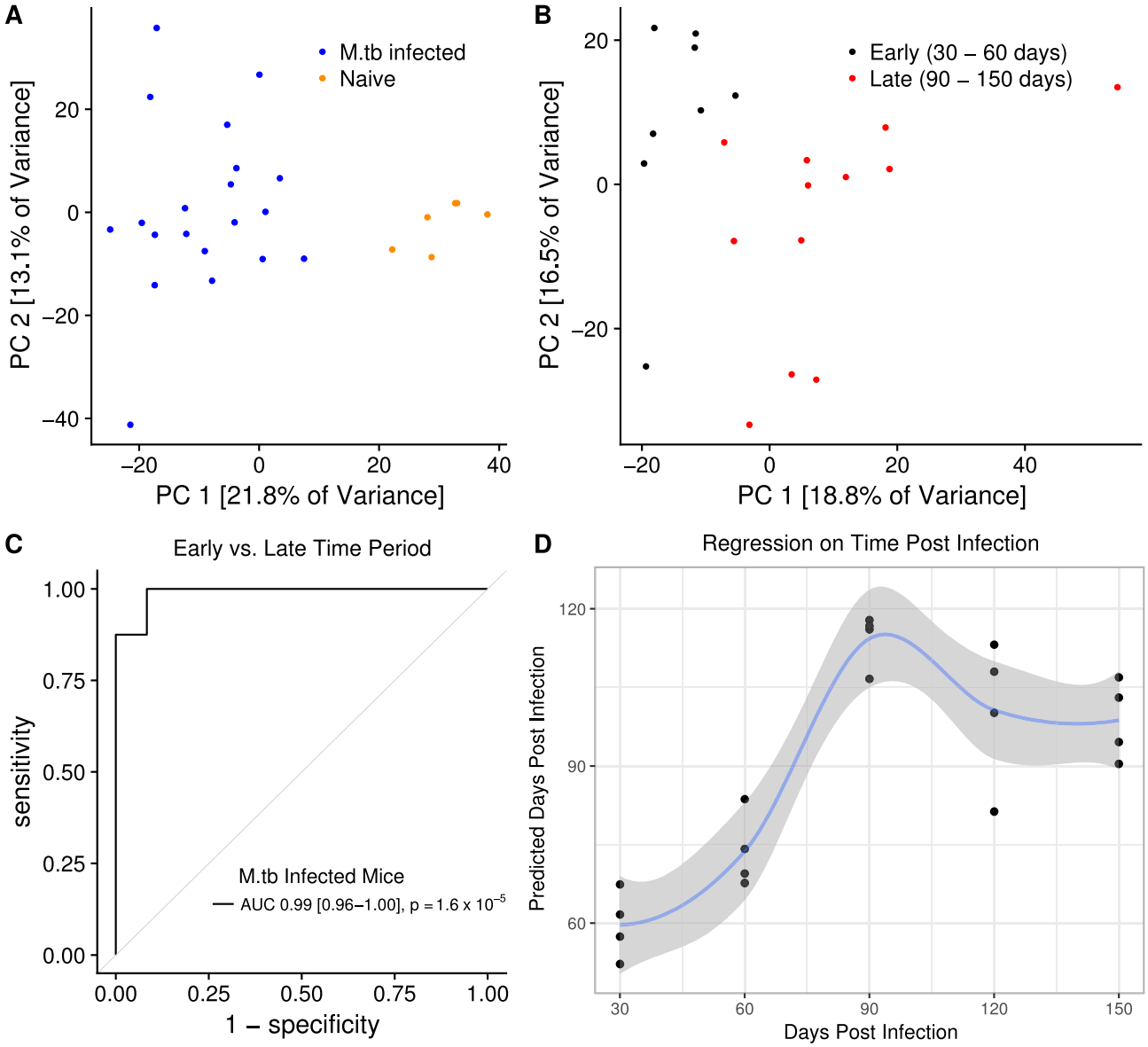

Figure 1. Blood genome-wide RNA expression discriminates early versus late *M.tb* infection time periods in C57BL/6 mice. Principle component analysis of genome-wide RNA expression measured via microarray in (**A**) all mice (n = 6 uninfected mice, n = 20 *M.tb* infected mice) stratified by infection status and (**B**) only *M.tb* infected mice stratified by time period post-infection. (**C**) ROC curve for out-of-bag performance of Random Forest Classifier predicting time period post-infection (1–2 months versus 3–5 months; *P* from Wilcoxon test, 95% confidence interval shown). (**D**) Random Forest Regression out-of-bag predictions of monthly time point post-infection. Fit curve calculated via the Loess method with 95% CI shown.

mouse data and recently available human data, we restricted our analysis to 8 time points from 20 days through 180 days (6 months) post-infection. To permit comparison of different computational models and allow a final unbiased estimate of predictive performance, we randomly divided the 38 macaques from this study into a training set and a test set, keeping the ratio of macaques with latent and active TB balanced in both groups (Figure S2).

Using ninefold cross-validation on the training set, we found that Regularized Logistic Regression, a linear method with regularization terms to reduce the number of probes in the signature, was not inferior to several nonlinear classification methods in predicting early (20–56 days) versus late (90–180 days) infection time period (Figure S3). We thus chose Regularized Logistic Regression to find a predictive RNA signature of time since *M.tb* infection in cynomolgus macaques. We found that this model predicted early (20–56 days) versus late (90–180 days) infection time period with an AUC of 0.78 in the training set (95% CI 0.72–0.85, $P = 5.6 \times 10^{-13}$; ninefold cross-validation; Fig. 2A), and an AUC of 0.81 in the test set (95% CI 0.71–0.91, $P = 1.6 \times 10^{-7}$; Fig. 2A).

Importantly, our model was trained and tested on macaques irrespective of their present or future TB disease status. If our model partially predicted disease status rather than only time period post-infection, the proportion of samples from macaques with active disease would differ between predicted and actual early time period samples. However, we found that there was no change in the proportion of samples from macaques with active

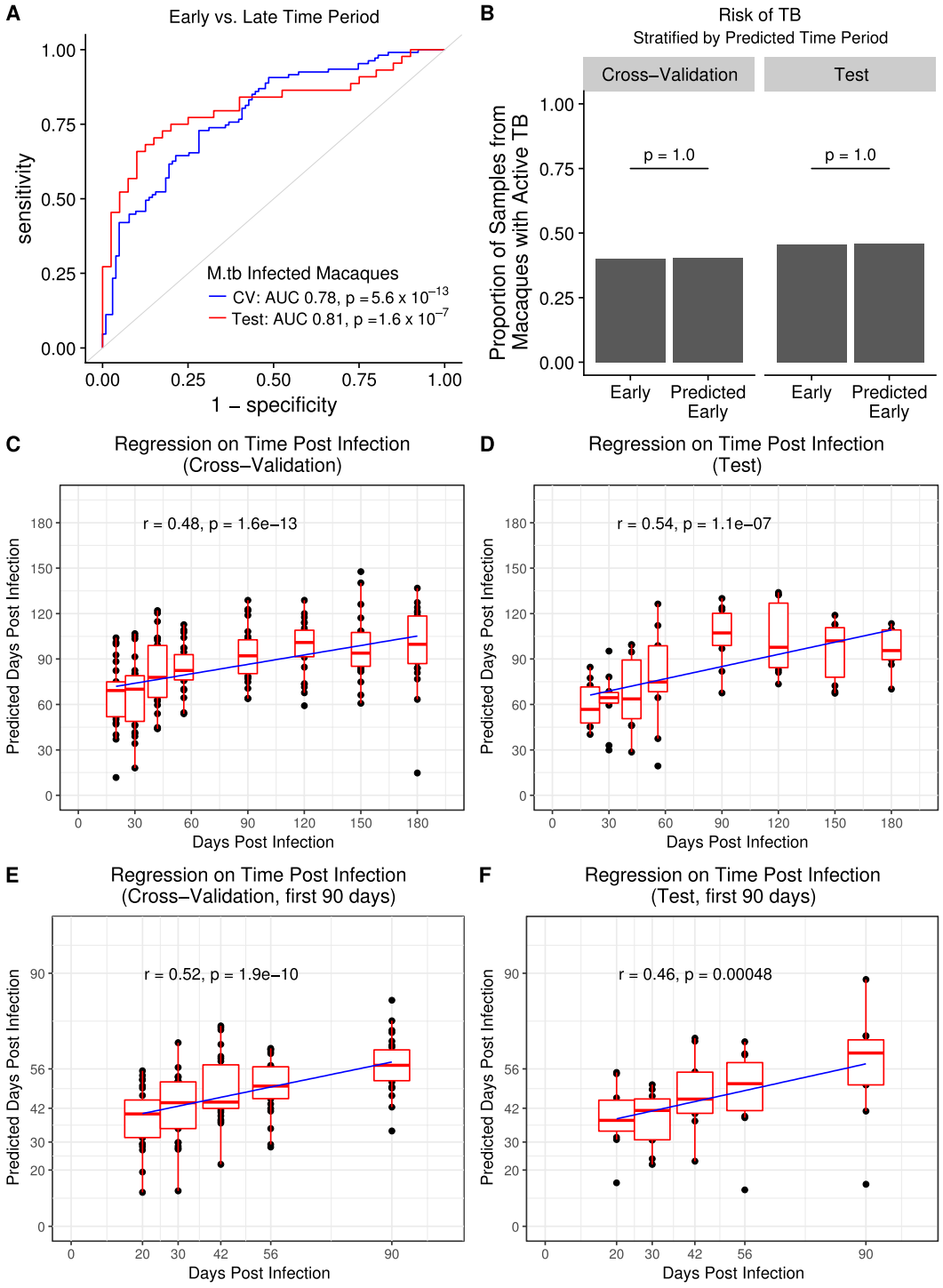

Figure 2. Blood RNA signature discriminates early versus late *M.tb* infection time periods in cynomolgus macaques. (**A**) ROC curves for Regularized Logistic Regression prediction of time period post-infection (20–56 days versus 90–180 days) from RNA expression in cynomolgus macaques on ninefold cross-validation in the training set (blue curve; n = 107 early time period samples, n = 103 late time period samples) and final model prediction on test set (red curve; n = 44 early time period, n = 40 late time period) (*P* from Wilcoxon test). (**B**) Comparison between early (20–56 days) (n = 107 train, n = 44 test) versus predicted early (n = 104 train, n = 50 test) time period samples in proportion of samples from macaques that develop active TB (*P* from Fischer's Exact test). Regularized Linear Regression predictions of time point post-infection for (**C**) ninefold cross-validation in the training set (n = 210) and for (**D**) final model prediction on the test set (n = 84). (E–F) Predictions from models trained and evaluated only on samples from the first 90 days post-infection (n = 134 train, n = 55 test). Boxplots represent medians with interquartile ranges for the predictions at each time point (best fit line shown, *P* from Pearson test).

disease in the predicted early time periods relative to the actual early time periods, in both the training and test sets ($P = 1.0$, $P = 1.0$ respectively; Fig. 2B). This was also true focusing on late time period predictions ($P = 1.0$ training, $P = 1.0$ test; data not shown).

Next, to assess whether each month post-infection could be predicted in cynomolgus macaques, treating time as a continuous variable, we performed Regularized Linear Regression with ninefold cross-validation on the training set and confirmed that days 20–56 were predicted as earlier time points than days 90–180, in both the training set and in the test set (Fig. 2C–D). As in our murine model analysis days 90–180 were not resolved. Quantitatively, the median absolute error (MAE) of the model was 38.5 days (Pearson's $r = 0.48$, $P = 1.6 \times 10^{-13}$) on the training set and 35.7 days ($r = 0.54$, $P = 1.1 \times 10^{-7}$) on the test set. Probes selected and used by the final trained regression model to predict in the test set are shown in Table S2. To assess whether we could predict specific time point of infection within the first 3 months, as suggested by our murine data, we trained a model on only time points from 20–90 days (Fig. 2E–F). The MAE of this model was 15.8 days on the training set and 14.3 days on the test set ($r = 0.52$, $P = 1.9 \times 10^{-10}$ and $r = 0.46$, $P = 4.7 \times 10^{-4}$, respectively).

We further assessed our early versus late classifier model's predictions on the pre-infection and 3–10 day time points (Figure S4). This showed that the pre-infection and 3–10 day time points were more similar to the late (90–180 days) time period than the distinct early (20–56 days) time period (Figure S4).

Taken together, these data indicate that we can broadly discriminate early and late *M.tb* infection in this cohort of cynomolgus macaques based on the whole blood transcriptomic response, and that we can moderately discriminate between the first two months of infection. These predictions do not depend on disease status, and the accuracy of the predictions is quantitatively lower in cynomolgus macaques than in C57BL/6 mice, as reflected by the AUC analyses. Treating time as a binary variable (Fig. 2A) or as a continuous variable (Fig. 2C–F) gave similar results.

Blood RNA expression of 250 genes predicts time since active TB exposure in humans. We next sought to determine whether our findings in mice and cynomolgus macaques could translate to humans. Whereas the day of infection is known in animal models, the precise time of exposure resulting in infection is difficult to determine in humans, even in careful clinical studies. One surrogate for time of infection in humans is time of IGRA or TST conversion in people who were known to be IGRA/TST negative previously. This would synchronize a human study cohort to the time of an initial systemic T cell response to *M.tb*. To test this hypothesis we accessed public data from South African adolescents who acquired latent *M.tb* infection during longitudinal blood sampling every 6 months[26]. We found that Regularized Logistic Regression was unable to predict the first time point of known IGRA conversion from 6 months post-first known IGRA conversion (0.54 AUC, 95% CI 0.27–0.82, $P = 0.64$ on test set; Fig. 3A). Notably, the biological event of actual IGRA conversion in this cohort could have occurred anytime between the first time point of IGRA positivity and the preceding 6 months. Given our findings in mice and macaques that the RNA signature of time since *M.tb* infection occurs within a brief window of 2–3 months, we interpret these findings to mean that sampling blood every 6 months in humans is unlikely to constitute a cohort where actual time of IGRA conversion is synchronized sufficiently to discover an RNA signature of time since IGRA conversion.

Another study design that could identify RNA correlates of recent infection in humans is a household contact study wherein healthy contacts of active TB cases are enrolled within a certain time from the date of diagnosis of the active TB case and sampled longitudinally. Important limitations of this design that could reduce the power to detect RNA correlates of recent infection are that the precise time of infection is not known and individuals who are IGRA + at enrollment may have been infected either from the present exposure or in the more distant past. Cognizant of these limitations, we accessed publicly available data from the Grand Challenges 6-74 (GC6-74) study of healthy household contacts (HHCs) of patients with active pulmonary TB[22]. HHCs in this cohort were enrolled within 2 months of the diagnosis of the active TB index case and had blood samples drawn at baseline, 6 months and/or 18 months post-enrollment[22]. Because our mouse and macaque analysis suggests that blood transcriptional changes are most prominent in an early 3 month window post-infection, we focused our first analysis on the baseline and 6 month time points. This included data from Gambian and Ethiopian cohorts but excluded data from the South African cohort because 6 month time points were not available for South Africa[22]. We used the same training/test split as the authors in the Gambian cohort but randomly split the Ethiopian cohort 50/50 between our training and test sets. Importantly, with this training/test split and our data pre-processing, we could predict risk of TB with 0.72 AUC (95% CI 0.60–0.83, $P = 1.6 \times 10^{-4}$; data not shown) in the training set by tenfold cross-validation and 0.70 AUC (95% CI 0.53–0.88, $P = 0.0071$; data not shown) in the test set using Regularized Logistic Regression. From the GC6-74 and the Adolescent Cohort Study (ACS) we used the RISK4 genes (*BLK*, *CD1C*, *GAS6* and *SEPT4*), the post-hoc selected *C1QC*, *TRAV27*, *ANKRD22*, *OSBPL10* genes and the 16 correlate of risk (COR) predictive genes together for this analysis[21,22]. When we used these same genes to train a model to predict time since TB exposure, we obtained no predictive performance, whether the model was trained with these gene sets separately or together ($P > 0.05$ for all test set predictions; Fig. 3B). This suggests that genes selected for optimal prediction of prospective TB risk do not change across these two time points post-exposure.

To find a predictive RNA signature of time since TB exposure in these data, and as the study authors performed for TB risk prediction, we used the Wilcoxon test on the training set to select transcripts that differed in expression between baseline and 6 month time points[22]. Using Regularized Logistic Regression we found that these genes predicted baseline versus 6 month time points with 0.90 AUC (95% CI 0.84–0.96, $P = 1.9 \times 10^{-13}$; tenfold cross-validation; Fig. 3C) in the training set and 0.69 AUC (95% CI 0.56–0.81, $P = 0.0039$; Fig. 3C) in the test set. We further used the final genes selected by the model (250 genes, Table S3) on the training set to train a model to predict risk of TB. As expected, these genes exhibited no direct predictive performance for risk

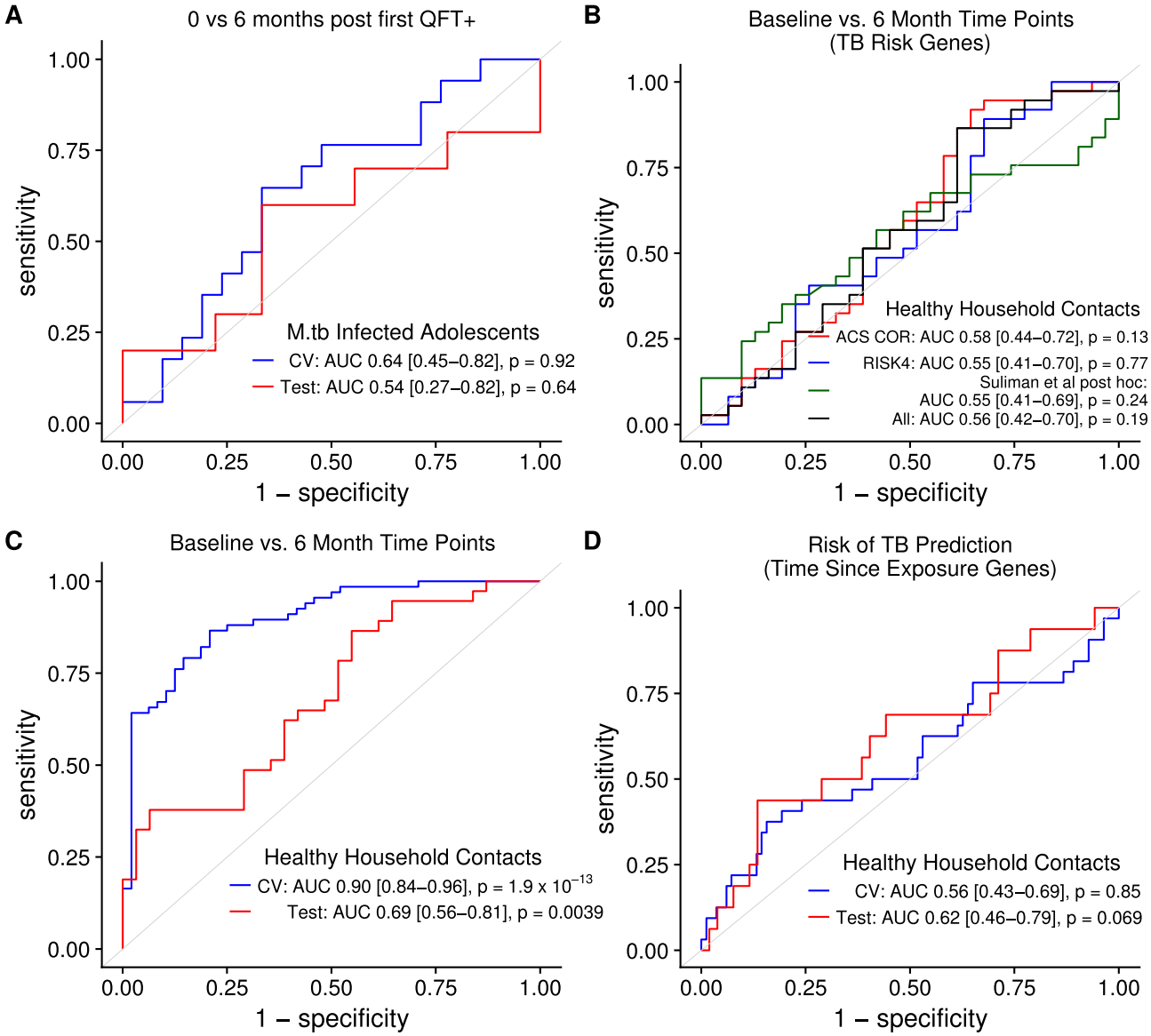

Figure 3. Blood RNA expression of 250 genes predicts time since active TB exposure in humans. (**A**) ROC curves for prediction of time since first known IGRA + (0 vs. 6 months) in South African adolescents who acquire *M.tb* infection for tenfold cross-validation in the training set (blue curve; n = 17 0 month samples, n = 21 6 month samples) and final model prediction on the test set (red curve; n = 10 0 month, n = 9 6 month) using Regularized Logistic Regression. (**B**) ROC curves for Regularized Logistic Regression prediction of time since active TB exposure (baseline vs. 6 months post-enrollment) in GC6-74 Gambia and Ethiopia test set (n = 37 baseline samples, n = 31 6 months samples) using expression of genes from published signatures that predict prospective risk of active TB. (**C**) ROC curves for Regularized Logistic Regression prediction of time since active TB exposure for tenfold cross-validation on the Gambia and Ethiopia training set (blue curve; n = 67 baseline, n = 48 6 months) and for final model prediction (contains 250 genes) on the Gambia and Ethiopia test set (red curve; n = 37 baseline, n = 31 6 months). (**D**) ROC curves for prediction of prospective risk of TB for tenfold cross-validation on the Gambia and Ethiopia training set (blue curve; n = 67 baseline, n = 48 6 months) and for final model prediction on the test set (red curve; n = 37 baseline, n = 31 6 months) using the 250-gene set that predicted time since active TB exposure. *P* values for all ROC curves are from Wilcoxon test, and 95% confidence intervals are shown.

of TB on the training or test sets ($P = 0.85$, $P = 0.07$, respectively; Fig. 3D). In summary, our findings with the household contact study design in humans parallel the results in macaques in that we can predict broad time period post-exposure via the whole blood transcriptomic response. Moreover, this transcriptomic signature of time period post-exposure to an active TB case is independent of the transcriptomic signature of risk of TB recently identified in the GC6-74 and ACS studies[21,22].

Time since TB exposure in humans is associated with alteration in CD4+T cell proportion and immune activation pathways. Cell-type deconvolution algorithms have recently been used with genome-wide RNA expression data to help identify changes in immune cell proportions in the blood that are associated with TB disease, prospective TB disease risk and treatment success[26,35]. To identify immune cell populations that are associated with time since TB exposure in the GC6-74 study, we used the leukocyte expression signature matrix 'immunoStates' and linear regression to infer leukocyte proportions for each subject's sample[35]. We found that the proportion of CD4 + α/β T cells was increased at 6 months versus baseline time point in the Gambian and Ethiopian cohorts ($P = 0.0079$; linear mixed model; Fig. 4A), but was not significantly changed at 18 months ($P = 0.20$ vs. baseline; linear mixed model, included South African cohort; Fig. 4B). We saw no significant differences in NK cell proportion over time in the Gambian and Ethiopian cohorts (Fig. 4C–D). Likewise, no other cell types estimated by the 'immunoStates' signature matrix showed significant differences over time in these cohorts ($P > 0.05$, linear mixed model, data not shown). This result with CD4 + α/β T cells and NK cells is consistent with the conclusion that the RNA signature of time since TB exposure is independent from the RNA signature of prospective TB risk, since both T cells and NK cells are known to decrease in circulation in active TB disease[26,36].

Our RNA signature of baseline versus 6 month time points post-exposure included 250 genes selected by Regularized Logistic Regression (Table S3). We utilized Ingenuity Pathway Analysis (IPA) to identify pathways associated with these genes. The majority of enriched canonical pathways (-log(p value) > 2) were associated with immune cell signaling, including B cells (B cell receptor and PI3K signaling), T cells (T cell receptor, PKCθ, regulation of IL-2 expression, 4-1BB and CD28 signaling), cytokines (IL-6, IL-15, IL-12, TNF, IL-8, IL-10 and IL-17A), innate immune cells (dendritic cell maturation and LPS-stimulated MAPK signaling) and humoral immunity (Fc Epsilon RI Signaling) (Fig. 4E, Table S4). Other enriched canonical pathways were related to cellular injury and toxicity (apoptosis), metabolism, nervous system signaling, PPAR signaling, cell cycle regulation and intracellular & second messenger signaling (Table S4). Considering the overall direction of change in the immune pathways between 6 month versus baseline time points, the upregulation of several pro-inflammatory signaling pathways (IL-6, IL-8, FLT3 signaling, PI3K signaling in B Lymphocytes and Dendritic Cell maturation) and decrease in anti-inflammatory signaling (PPAR signaling) suggests that an increase in peripheral blood immune activation occurs at the 6 month time point after exposure (Fig. 4E).

To compare transcriptional modules altered in humans to those altered in mice and macaques, we used the recently defined disco score to identify concordantly and discordantly altered modules between these species[37]. Several modules related to T cells and NK cells were enriched (adjusted $P < 0.05$) in each pairwise comparison between two species (human vs. mouse, human vs. monkey, and monkey vs. mouse) (Fig. 5). Several B cell-related modules were uniquely concordantly regulated between macaques and mice (Fig. 5).

Application of reduced 6-gene expression signature of time since active TB exposure to adolescent *M.tb* infection acquisition cohort confirms its identification of recent infection in humans. Implementation of our newly discovered RNA signature of time since active TB exposure using qRT-PCR would require a more parsimonious gene set than the 250 genes heretofore described. To find a reduced gene signature we ran a forward search using the MetaIntegrator R package[38]. This method identified 6 genes, *RP11-552F3.12, PYURF, TRIM7, TUBGCP4, ZNF608* and *BEAN1*, that recapitulated the performance of the 250 gene signature on baseline versus 6 month time point discrimination with 0.86 AUC (95% CI 0.80–0.93, $P = 1.7 \times 10^{-11}$; Fig. 6A) in our GC6-74 training set and 0.68 AUC (95% CI 0.55–0.81, $P = 0.0055$; Fig. 6A) in the test set. Independent validation of this signature requires a cohort wherein recent *M.tb* infection is documented and time points are available to test whether the signature allows discrimination between recent and more remote infection. While the cohort of South African adolescents who acquired latent *M.tb* infection did not permit discovery of an RNA signature of recent *M.tb* infection, we reasoned that the whole cohort would be powered for validation of our signature discovered in the GC6-74 household contact study design[26]. Three genes, *TRIM7*, *ZNF608*, *TUBGCP4*, from our 6-gene signature were represented by detected probes in the microarray used in this study. These 3 genes discriminated the first time point of known IGRA conversion from all pre-conversion time points (6 months and 12 months prior to known conversion) with 0.72 AUC (95% CI 0.58–0.87, $P = 0.0030$; Fig. 6B). These 3 genes likewise discriminated the first time point of known IGRA conversion from all sampled time points (6, 12 months prior to conversion and 6, 12 months after known conversion) with 0.68 AUC (95% CI 0.56–0.81, $P = 0.0039$; Fig. 6B). Figure S5 shows the trajectory of the 3 gene score over time, being highest at the first time point of known IGRA conversion.

Given that time since active TB exposure is the single strongest clinical risk factor for developing TB disease in immunocompetent persons, the finding that time since exposure and risk of TB, as predicted by the blood transcriptomic response, are independent in the GC6-74 study of healthy household contacts suggests that these signatures could be combined to possibly better predict risk of TB when the time of exposure is unknown[17,19,20]. While the GC6-74 study was not powered for this particular secondary analysis, we assessed whether the highest 6-gene score during longitudinal sampling allowed discrimination of subjects who did or did not progress to active TB disease during study follow-up. The highest 6-gene score did not discriminate progressors from non-progressors in the test set, whether the subjects were from South Africa (AUC 0.51, 95% CI 0.40–0.63, $P = 0.60$; Fig. 6C) or from Gambia or Ethiopia (AUC 0.63, 95% CI 0.46–0.80, $P = 0.095$; Fig. 6C). The same result was observed in the ACS cohort of IGRA + adolescents with unknown exposure history (AUC 0.55, 95% CI 0.43–0.66, $P = 0.78$; Fig. 6C).

We additionally tested whether the 6-gene signature discriminated early from late time periods post-infection as defined in our analysis of animal models of *M.tb* infection. The 5 genes represented by detected probes of homologous genes in the mouse microarray data discriminated the early (30–60 days) versus late (90–150 days)

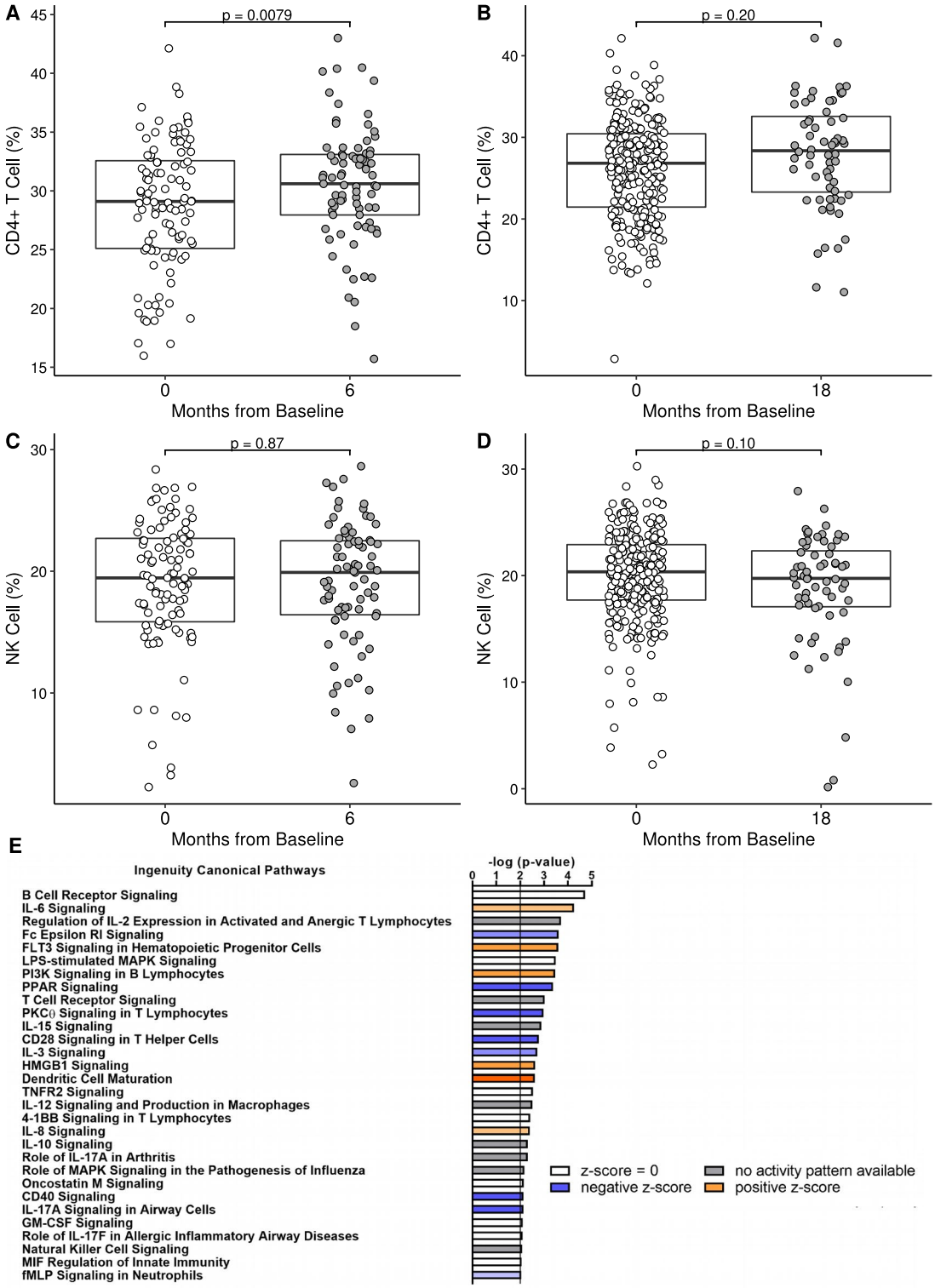

Figure 4. Time since TB exposure in humans is associated with alteration in CD4 + T cell proportion and immune activation pathways. Changes in CD4 + T cell percentages (**A**,**B**) and NK cell percentages (**C**,**D**) in GC6-74 healthy household contacts cohort at baseline (n = 104 in **A**,**C**; n = 272 in **B**,**D**), 6 month (**A**,**C**; n = 79) and 18 month (**B**,**D**; n = 64) time points after active TB exposure were determined by cell-type deconvolution (*P* from linear mixed model). Boxplots represent medians with interquartile ranges. (**E**) Top immunity related enriched canonical pathways in the 250-gene RNA signature of time since exposure to active TB index case (6 months vs. baseline) by IPA (*P* from Fisher's Exact test).

Figure 5. Enriched transcriptional modules are concordantly or discordantly regulated during recent *M.tb* exposure or infection between mice, macaques or humans by disco analysis. *P* from CERNO statistical test.

infection time period with 0.80 AUC (95% CI 0.60 – 1.00, $P = 0.012$; Fig. 6D). Only 1 gene was represented by a detected probe in the macaque microarray data, and it alone did not discriminate the early versus late infection time period (0.52 AUC, 95% CI 0.45 – 0.58, $P = 0.70$; Fig. 6D). Of note, this study utilized a human microarray platform for the macaque samples, which may have contributed to reduced measurability of the macaque homologues to these human genes[25]. Because of the different measurement platforms between the animal and human data (microarray vs RNAseq) we could not directly apply our human 250 gene signature to the animal data.

Discussion

Early clinical studies in the pre-antibiotic era in the relatively isolated Faroe Islands shed light on the clinical features of primary infection with *M.tb* in humans, which often include fever, elevated erythrocyte sedimentation rate, X-ray abnormalities and, less often, erythema nodosum[39,40]. With time of exposure to an active TB case pinpointed within a two week period, and sometimes to a single day, Poulsen determined that these clinical features accompany and/or follow TST conversion, which occurs within 6 weeks of exposure[14,39,40]. While these clinical features of initial *M.tb* infection are transient and not specific to *M.tb* infection, a method to determine that a person is currently in the first 1–2 years post initial infection would have great prognostic value for near-future

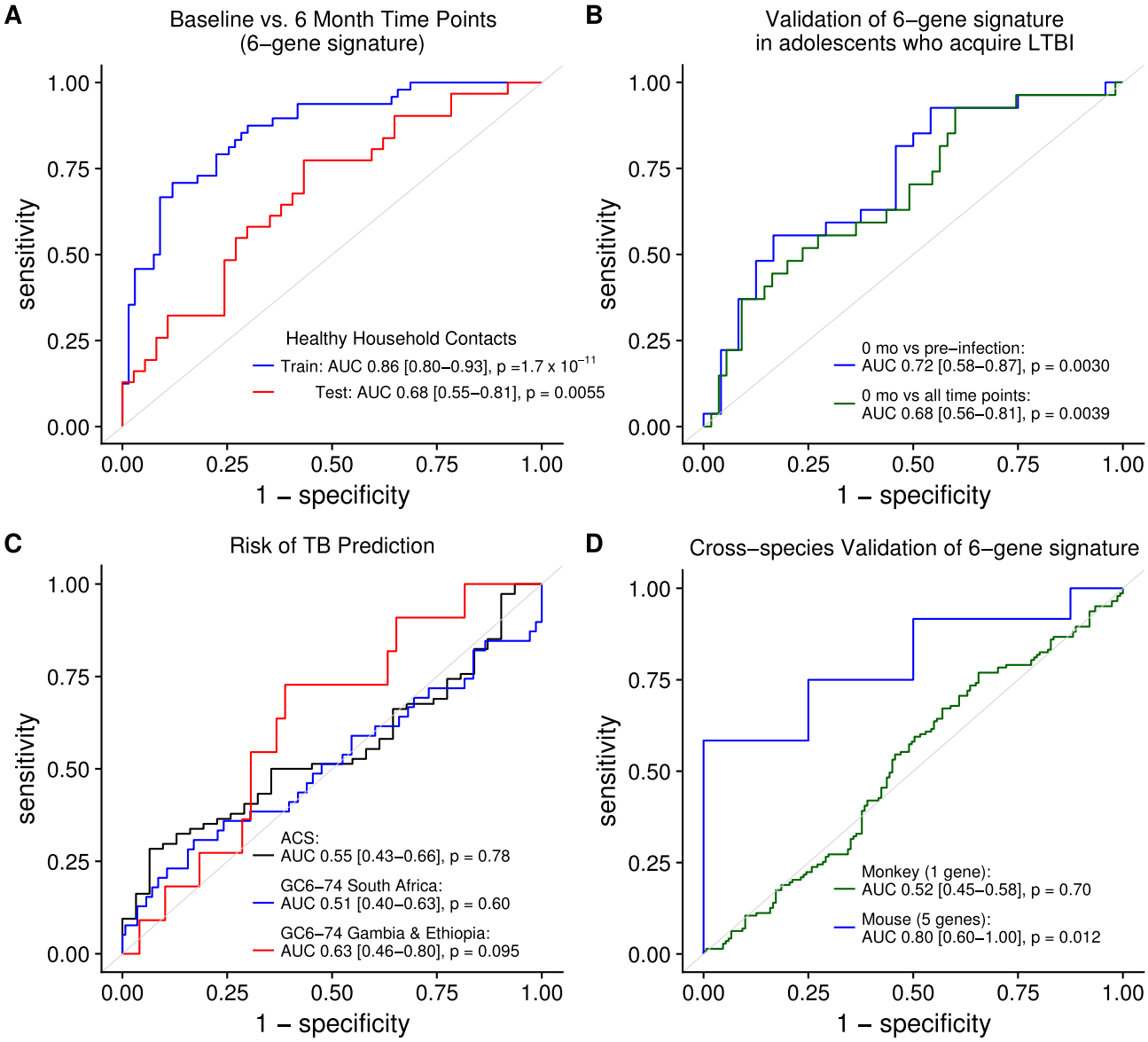

Figure 6. Application of reduced 6-gene signature of time since active TB exposure to adolescent *M.tb* infection acquisition cohort confirms its identification of recent infection in humans. (**A**) ROC curves for 6-gene score prediction of time since active TB exposure in the Gambia and Ethiopia training set (blue curve; n = 67 baseline samples, n = 48 6 months samples) and for the Gambia and Ethiopia test set (red curve; n = 37 baseline, n = 31 6 months). (**B**) ROC curves for discrimination between time of first known IGRA + and all pre-conversion time points (blue curve; n = 27 0 month, n = 24 pre-conversion) and between time of first known IGRA + and all other time points (green curve; n = 27 0 month, n = 24 pre-conversion and n = 31 6 or 12 months after known conversion) in South African adolescents who acquire *M.tb* infection using 3-gene score from genes detected in microarray data. (**C**) ROC curves for prediction of prospective risk of TB using highest 6-gene score observed per individual in the ACS cohort (n = 74 nonprogressors, n = 31 progressors), GC6-74 Gambia and Ethiopia test set (n = 49 nonprogressors, n = 11 progressors) and GC6-74 South Africa cohort (n = 141 nonprogressors, n = 39 progressors). (**D**) ROC curves for discrimination of early and late time periods post-infection in mice (blue curve; n = 8 early mice, n = 12 late mice) and macaques (green curve; n = 151 early samples, n = 143 late samples) using genes from the 6-gene signature that were detected in the respective microarrays. *P* values for all ROC curves are from Wilcoxon test, and 95% confidence intervals are shown.

TB disease and could allow real-time geospatial mapping of recent TB transmission in communities[14,15,17,39,41–43]. In our proof of concept analysis, we sought to determine whether it is possible to develop an RNA-based blood test to detect recent exposure or infection with *M.tb*. For TB disease risk prediction, we hypothesized that such a test could complement the recently developed RNA signatures of TB disease risk that are based on detecting incipient TB, which is the asymptomatic phase of early TB disease during which pathology progresses gradually before full-blown clinical TB[21,22,44,45]. Because RNA signatures of incipient TB are only sensitive in the 1–3 months preceding symptomatic TB diagnosis, their current proposed use involves serially screening infected individuals

on a regular basis[45]. We envision that a blood test for recent infection could be used concurrently with signatures of incipient TB at initial testing to aid in the decision to start treatment or to rule out further screening in individuals who are initially negative for incipient TB. Using our mouse data and published macaque data, we have demonstrated a highly accurate RNA signature of recent infection with *M.tb* (1–2 vs. 3–5/6 months post-infection). Using the GC6-74 cohort of HHCs of patients with active pulmonary TB, we discovered 250-gene and 6-gene human RNA signatures of recent exposure (0–2 vs. 6–8 months post-diagnosis of index case) that validated within a held-out test set[22]. Using an independent cohort of adolescents who acquired *M.tb* infection during 6-month longitudinal sampling, we demonstrated that the 6-gene signature could discriminate the first known time point of IGRA conversion from pre-conversion time points and from 6–12 months later with modest accuracy (0.68 AUC)[26]. However, this 6-gene signature was unable to provide prognostic information of TB risk in the GC6-74 cohort or the ACS cohort. The incomplete time point sampling of most individuals, and 6-month sampling likely reduced the power to find an association between our 6-gene signature score and TB risk in these two studies. Nevertheless, we believe the sampling constraints and target populations of these studies, adults who are HHCs and adolescents with LTBI of unknown exposure history, both in highly endemic areas, are mostly in line with what may be feasible for applying transcriptional signatures of TB risk for targeted treatment to reduce TB incidence[45]. Given that early blood transcriptional changes occurred within a short 3 month window in our mouse and macaque analyses, and the human data analyzed are not inconsistent with this brief timeline, we believe that blood RNA signatures for recent *M.tb* infection are too brief in duration to yield a useful biomarker to improve prediction of TB risk for targeted preventive therapy. Nevertheless, the finding that all three species exhibit an early blood transcriptional response to *M.tb* infection is highly novel and has implications for further research into common and different early immune responses to *M.tb* among these species.

Because the vast majority of TB disease burden can be accounted for epidemiologically by recent infection (past 1–2 years), we hypothesize that, on average, the factors influencing progression of disease have resolved by 2 years post initial infection in humans[3,14]. Therefore, we hypothesize that a biological correlate of recent infection that has the longest duration during that time when the outcome of early disease progression has not been resolved would have the highest chance of being useful as a complement to tests for incipient TB in predicting TB risk. Importantly, most or all biological correlates of recent infection will not play a causal role in determining who will or will not progress to TB disease because most infected individuals never develop TB disease. This is supported by the fact that our RNA signature of recent infection was independent of known RNA signatures of TB risk. Our RNA signature of recent infection is likely characteristic of a very early phase of infection when many factors defining the outcome of infection have yet to resolve. A biosignature of recent *M.tb* infection that is useful in predicting individual TB disease risk will derive its utility not from showing who among recently infected individuals will progress to TB disease but rather by discriminating between recently infected individuals who are still at some risk of TB disease and those infected remotely for whom the factors driving disease progression have resolved. The longer the duration of a biological correlate of recent infection, the higher sensitivity it will have for identifying recently infected individuals.

Our estimated cell type and pathway analyses suggest that both cellular and molecular signatures of immune activation associated with recent exposure and could be interrogated by other modalities such as epigenetics. Immune cell differences between recently acquired and remotely acquired infection have been reported by others in single cohorts without longitudinal sampling[20,46]. The high enrichment of B cell signaling in our signature is interesting, and a recent case control study in a single cohort showed that several IgG and IgA antibodies to *M.tb* antigens strongly discriminated (AUC > 0.90) active TB contacts who converted on TST from non-converters both at first known conversion and 3 months prior[47].

Our analyses and these considerations suggest that sampling IGRA-, untreated HHCs every month (or more frequently) for one to two years, starting as soon as possible after the diagnosis of their respective index case and determining IGRA conversion events, would allow for the discovery of biosignatures of recent *M.tb* infection that could be useful for helping predict TB disease risk. Follow-up in such a cohort for TB progression would allow better assessment of how signatures of recent infection and signatures of incipient TB could be combined to improve TB risk prediction. The addition of chest X-rays with deep machine learning analysis could be useful to discover heretofore unknown, specific radiogenomic features of recent infection or incipient TB[48,49]. After IGRA conversion, staggered sampling at different times could reduce the study's burden on individual subjects and allow more precise estimation of the duration of any biomarker. Most follow-up in such a study would have to be performed on those who refuse preventive treatment, as treatment would need to be offered because recent infection is precisely documented. Another potential benefit of such a study is that validated biomarkers that associate strongly with TST/IGRA conversion but precede conversion, such as currently unvalidated IgG and IgA markers, could be used to identify *M.tb* infection before TST/IGRA conversion and thus reduce the burden of follow-up of recent contacts in TB control programs and potentially help reduce LTBI treatment time[47].

If deployed in population screening efforts, a test for recent *M.tb* infection could also allow real-time geospatial mapping of recent TB transmission in communities. This could greatly help the application of current control methods to reduce TB transmission and disease in high incidence settings. While it is possible that our current 6-gene signature of recent *M.tb* infection could be evaluated in the future for this purpose, we think it would be more prudent to first find biomarkers of recent *M.tb* infection that have a longer duration and are useful for individual TB risk prediction. Being predictive of individual TB risk is more difficult to achieve than correlation with recent infection, as we have shown. Both applications are related, because the better a test for recent infection associates with disease progression in an otherwise unselected population of previously infected individuals, the more sensitive it will be for contact tracing and transmission studies. Nevertheless, biomarkers of varying duration could be jointly useful for the application of mapping recent transmission.

Our results in mice, macaques and humans, together with recent literature, suggest that future longitudinal studies of HHCs may be successful at identifying more accurate biomarkers of time since *M.tb* infection in

humans. Our study represents one of only a handful of studies since Poulsen's early work showing that there are biological events in the early human response to *M.tb* infection that can be reproducibly measured[39]. Future biomarker studies may enable the study of early events of infection in humans both routinely and ethically and permit the identification of immunological or other biological events that determine whether an exposed person will develop TB disease or control the infection[5,50–52]. This could greatly aid vaccine development for TB as no correlates of protection for TB are yet known[53]. We also expect that more accurate biomarkers of time since *M.tb* infection will be excellent tools to help better understand the human phenotypes of IGRA reversion and persistent resistance to IGRA conversion[51,54].

Our current analysis has some limitations. Because most transmission occurs outside the household contact setting, many individuals in the GC6-74 study were TST + at enrollment (~ 51.4% in Ethiopia, ~ 36.3% in The Gambia), and follow-up TST in this study were incomplete, it is highly likely that many, and possibly the majority, of contacts in this study were not infected or re-infected from their index TB case[10,11,55]. However, the 6-gene RNA signature discovered in this cohort was validated in adolescents where recent *M.tb* infection was documented via IGRA conversion in 100% of study participants[26]. Finally, our current analysis excluded HIV co-infection.

Methods

Study design. The objective of this study was to identify blood RNA correlates of time since *M.tb* infection or exposure. We first infected mice with *M.tb* via the aerosol route and measured genome-wide RNA expression at pre-specified time points. Unsupervised analysis revealed potential discrimination between mice sacrificed at early time points (1–2 months) vs. late time points (3–5 months). Cross-validation without hyperparameter tuning identified an unbiased RNA signature that accurately predicted early vs. late time period post-infection. We then retrospectively mined publicly available data from a prospective *M.tb* infected cynomolgus macaque cohort and a prospective healthy household contact human cohort to identify RNA signatures that predicted these same time periods post-infection. The human RNA signature was validated in an independent cohort, adolescents who were recently infected with *M.tb* during longitudinal sampling.

Mice. Specific pathogen-free, 6–12 week old, female C57BL/6 wild-type mice (The Jackson Laboratory, Bar Harbor, ME) were maintained in ventilated cages inside a biosafety level 3 (BSL3) facility and provided with sterile food and water ad libitum. All protocols were approved by The Ohio State University's Institutional Laboratory Animal Care and Use Committee. Mice experiments were performed in accordance with the U.S. National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Mouse aerosol infection and blood collection. *M.tb* Erdman (ATCC no. 35801) was obtained from the American Type Culture Collection. Stocks were grown according to published methods[56]. Mice were infected with *M.tb* Erdman using an inhalation exposure system (Glas-Col) calibrated to deliver 50 to 100 CFUs to the lungs of each mouse, as previously described[56,57]. At specific time points post-*M.tb* infection, infected and age-matched uninfected mice were sacrificed and blood collected (400 µL) from the heart into 1.2 mL Tempus reagent and stored at − 80 °C. No formal randomization was employed for choosing cages of mice to be sacrificed at each time point. For the *M.tb* infected mice, sample size per time point was determined by using the number we routinely use for well-powered molecular and immunological studies in inbred mice. No blinding was performed for the mouse study.

RNA processing and microarray hybridization. Whole blood RNA was processed, quantified using a NanoDrop 1000 Spectrophotometer (NanoDrop Technologies) and RNA integrity (RIN) determined by a 2100 Bioanalyzer (Agilent). Samples with RIN ≥ 6.5 were submitted for hybridization onto Illumina Mouse WG 6-V2 BeadChips and scanned on an Illumina BeadStation system. Microarray data are available in the Gene Expression Omnibus (GEO) database under accession number GSE124688.

Microarray data pre-processing. For our murine data, Illumina BeadStudio/GenomeStudio software was used to subtract background and scale average signal intensity for each sample to the global median average intensity across all samples. Probes with a detection P value ≤ 0.01 in at least 10% of mice were filtered for analysis. Thereafter R scripts were used to quantile normalize the data, set all values < 10 to 10 and $\log_2$ transform the data. Probes were filtered by two-fold change in expression from the median in at least 10% of samples. For the macaque data (GSE84152), microarray data pre-processing was performed as previously described[25]. The data from the human adolescent cohort of IGRA converters (GSE116014) was pre-processed identically as the macaque data, except that data were quantile normalized and no batch correction was performed. When these adolescent data were used to validate the 6-gene signature, the data were downloaded at the gene-level using the R MetaIntegrator package, before additional pre-processing[38].

RNA-seq data pre-processing. Human data from the Grand Challenges 6–74 (GC6-74) cohort were downloaded at the gene count level from GEO (GSE94438). Genes with read count ≤ 5 in 50% of samples were excluded. Data were quantile normalized and $\log_2$ transformed. To facilitate comparisons with a common RNA-seq alignment pipeline, gene counts were obtained from the ARCHS[4] resource when comparing data from the Adolescent Cohort Study (GSE79362) and GC6-74 cohorts using the 6-gene signature[58]. These data were otherwise processed identically.

Machine learning predictions. For predicting time since infection in mice, we used the Random Forest algorithm in R with default parameter values[59]. Out-of-bag predictions were used to estimate model accuracy, which corresponds approximately to threefold cross-validation.

To predict time since infection in macaques, we randomly partitioned the macaques into training (70%) and test (30%) sets. We compared several different machine learning algorithms using the R caret package[60]. These included: Random Forest (R ranger package[61]), Gradient Boosted Machines (R gbm package[62]), Support Vector Machines using Polynomial (R kernlab package[63]) or RBF kernels (R kernlab package[63]) and Regularized Logistic Regression (R glmnet package[64]). Ninefold cross validation was used in the training set to optimize model hyperparameters and assess predictive performance, with all samples related to individual macaques being partitioned into the same held-out fold to ensure unbiased cross-validation. The caret package implementation did not permit tenfold cross validation for this dataset, as in humans, but the results should be equivalent. Only Regularized Logistic Regression was used for predictions in the test set and Regularized Linear Regression for predicting each time point post-infection after Regularized Logistic Regression was shown to be superior in predicting time period post-infection.

To predict time since TB exposure, time since IGRA conversion or prospective risk of TB in humans, we used tenfold cross validation on the training set (either GC6-74 or Adolescent IGRA converter cohort), with each subject's samples partitioned into the same held-out fold, to optimize Regularized Logistic Regression model hyperparameters before predicting on the test set. Prior to performing this procedure for time since TB exposure on the GC6-74 training set, we performed feature selection on genes by a Wilcoxon test ($P < 0.05$). Where longitudinal data were available for individual macaques or persons, each time point was considered as an independent sample.

Forward search to discover parsimonious 6-gene signature. A forward search was performed in the GC6-74 Gambia and Ethiopia training set on genes selected by a Wilcoxon test ($P < 0.05$) using the R MetaIntegrator package as previously described[38,65]. The stopping threshold for increase in AUC with the addition of each gene was varied until a signature comprising less than 10 genes and including both upregulated and downregulated genes at 6 months post-enrollment (vs. baseline) was obtained. The final signature's score is calculated on normalized $\log_2$ expression values as a difference between upregulated and downregulated genes: (*RP11-552F3.12* + *PYURF* + *TRIM7* + *TUBGCP4*)–(*ZNF608* + *BEAN1*). When applying this score to microarray data, multiple detected probes that mapped to these genes, using the R biomaRt package, were averaged[66]. Genes without corresponding detected probes were omitted from the calculation.

Cell type deconvolution, pathway and transcriptional module analysis. Cell type proportions in blood were estimated from RNA-seq data as previously described using the R MetaIntegrator package[26,35,38]. Gene-level expression for this deconvolution was obtained from the ARCHS[4] resource[58]. For pathway analysis, the 250 genes comprising the signature of time since exposure to an active TB case (6 months vs. baseline) were analyzed using canonical pathway analysis with QIAGEN's Ingenuity Pathway Analysis platform (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity). To compare transcriptional modules that were concordantly or discordantly regulated between mice, macaques and humans at early and late post-exposure time points, we used the R disco and tmod packages with transcriptional modules from Li et al.[67–69]. Genes used in this analysis included all detected probes (mice and macaques) and genes (humans). Differential expression and ortholog assignment were performed as previously described[37]. The 6-month time point in the GC6-74 cohort was taken as the early time point in humans based on the results in Figs. 5 and S3, as this time point had the highest 6-gene score (data not shown).

Statistical analysis. All statistical analyses were performed in R (version 3.4.3). Prediction performance was evaluated using receiver operator characteristic (ROC) curves. Statistical significance of the area under the curve (AUC) was assessed using the one-sided Wilcoxon test via the R verification package[70]. ROC graphs and confidence intervals were obtained via the R pROC package[71]. Pearson test was used for correlation analysis. Fischer's exact test (two-sided) was used to determine statistical significance of comparisons between proportions in evaluating the independence of the time since infection signatures from risk of TB disease in macaques. We used linear mixed models to assess the significance of cell type proportion changes with time since TB exposure via the R lme4 package[72]. Subject and site were included as random effects and time since exposure and site as fixed effects. These two-sided P values were obtained via the Satterthwaite approximation. The IPA canonical pathway P values were calculated by a one-sided Fisher's Exact Test, with $P < 0.01$ considered as significant. The transcriptional module P values were calculated using the CERNO statistical test, with $P < 0.05$ considered as significant after Benjamini–Hochberg correction[37]. For all other statistical tests, $P < 0.05$ was considered as significant.

Data availability

Mouse microarray data are available in the Gene Expression Omnibus database under Accession Number GSE124688. Published data used in this study are available in the Gene Expression Omnibus database under Accession Numbers GSE79362, GSE84152, GSE94438 and GSE116014.

Code availability

Source code for all analyses is publicly available in a GitHub repository: https://github.com/remi10001/TB.

Received: 11 February 2020; Accepted: 18 September 2020
Published online: 09 October 2020

References

1. World Health Organization. *Global Tuberculosis Report 2018*. https://www.apps.who.int/medicinedocs/en/m/abstract/Js23553en/.
2. World Health Organization. *Global strategy and targets for tuberculosis prevention, care and control after 2015*. https://www.who.int/tb/post2015_strategy/en/ (2014).
3. Houben, R. M. G. J. & Dodd, P. J. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLOS Med.* **13**, e1002152 (2016).
4. Vynnycky, E. & Fine, P. E. M. Lifetime risks, incubation period, and serial interval of tuberculosis. *Am. J. Epidemiol.* **152**, 247–263 (2000).
5. Keshavjee, S. *et al.* Moving toward tuberculosis elimination. Critical issues for research in diagnostics and therapeutics for tuberculosis infection. *Am. J. Respir. Crit. Care Med.* **199**, 564–571 (2018).
6. Shah, N. S., Kim, P., Kana, B. D. & Rustomjee, R. Getting to zero new tuberculosis infections: insights from the National Institutes of Health/US Centers for Disease Control and Prevention/Bill & Melinda Gates Foundation workshop on research needs for halting tuberculosis transmission. *J. Infect. Dis.* **216**, S627–S628 (2017).
7. Churchyard, G. *et al.* What we know about tuberculosis transmission: an overview. *J. Infect. Dis.* **216**, S629–S635 (2017).
8. Dowdy, D. W. *et al.* Designing and evaluating interventions to halt the transmission of tuberculosis. *J. Infect. Dis.* **216**, S654–S661 (2017).
9. Wiker, H. G., Mustafa, T., Bjune, G. A. & Harboe, M. Evidence for waning of latency in a cohort study of tuberculosis. *BMC Infect. Dis.* **10**, 37–46 (2010).
10. Verver, S. *et al.* Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *The Lancet* **363**, 212–214 (2004).
11. Yates, T. A. *et al.* The transmission of *Mycobacterium tuberculosis* in high burden settings. *Lancet Infect. Dis.* **16**, 227–238 (2016).
12. Zelner, J. L. *et al.* Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. *J. Infect. Dis.* **213**, 287–294 (2016).
13. Cegielski, J. P. *et al.* Eliminating tuberculosis one neighborhood at a time. *Am. J. Public Health* **103**, 1292–1300 (2013).
14. Behr, M. A., Edelstein, P. H. & Ramakrishnan, L. Revisiting the timetable of tuberculosis. *BMJ* **362**, k2738 (2018).
15. Fox, G. J., Barry, S. E., Britton, W. J. & Marks, G. B. Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur. Respir. J.* **41**, 140–156 (2013).
16. Kasaie, P., Andrews, J. R., Kelton, W. D. & Dowdy, D. W. Timing of tuberculosis transmission and the impact of household contact tracing. An agent-based simulation model. *Am. J. Respir. Crit. Care Med.* **189**, 845–852 (2014).
17. Reichler, M. R. *et al.* Risk and timing of tuberculosis among close contacts of persons with infectious tuberculosis. *J. Infect. Dis.* **218**, 1000–1008 (2018).
18. Sloot, R., Schim van der Loeff, M. F., Kouw, P. M. & Borgdorff, M. W. Risk of tuberculosis after recent exposure. A 10-year follow-up study of contacts in Amsterdam. *Am. J. Respir. Crit. Care Med.* **190**, 1044–1052 (2014).
19. Sutherland, I. Recent studies in the epidemiology of tuberculosis, based on the risk of being infected with tubercle bacilli. *Adv. Tuberc. Res.* **19**, 1–63 (1976).
20. Halliday, A. *et al.* Stratification of latent *Mycobacterium tuberculosis* infection by cellular immune profiling. *J. Infect. Dis.* **215**, 1480–1487 (2017).
21. Zak, D. E. *et al.* A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet* **387**, 2312–2322 (2016).
22. Suliman, S. *et al.* Four-Gene Pan-African Blood Signature Predicts Progression to Tuberculosis. *Am. J. Respir. Crit. Care Med.* **197**, 1198–1208 (2018).
23. Leong, S. *et al.* Cross-validation of existing signatures and derivation of a novel 29-gene transcriptomic signature predictive of progression to TB in a Brazilian cohort of household contacts of pulmonary TB. *Tuberculosis* **120**, 101898 (2020).
24. World Health Organization. *Development of a Target Product Profile (TPP) and a framework for evaluation for a test for predicting progression from tuberculosis infection to active disease*. https://apps.who.int/iris/bitstream/handle/10665/259176/WHO-HTM-TB-2017.18-eng.pdf;jsessionid=EBD2B5F9B500750ECB57D8E796BFD533?sequence=1 (2017).
25. Gideon, H. P., Skinner, J. A., Baldwin, N., Flynn, J. L. & Lin, P. L. Early whole blood transcriptional signatures are associated with severity of lung inflammation in cynomolgus macaques with *Mycobacterium tuberculosis* infection. *J. Immunol.* **197**, 4817–4828 (2016).
26. Chowdhury, R. R. *et al.* A multi-cohort study of the immune factors associated with *M. tuberculosis* infection outcomes. *Nature* **560**, 644–648 (2018).
27. Gonzalez-Juarrero, M. *et al.* Immune response to *Mycobacterium tuberculosis* and identification of molecular markers of disease. *Am. J. Respir. Cell Mol. Biol.* **40**, 398–409 (2009).
28. Mollenkopf, H.-J., Hahnke, K. & Kaufmann, S. H. E. Transcriptional responses in mouse lungs induced by vaccination with *Mycobacterium bovis* BCG and infection with *Mycobacterium tuberculosis*. *Microbes Infect.* **8**, 136–144 (2006).
29. Shi, L. *et al.* Infection with *Mycobacterium tuberculosis* induces the Warburg effect in mouse lungs. *Sci. Rep.* **5**, 18176 (2015).
30. Beamer, G. L. & Turner, J. Murine models of susceptibility to tuberculosis. *Arch. Immunol. Ther. Exp. (Warsz.)* **53**, 469–483 (2005).
31. Medina & North. Resistance ranking of some common inbred mouse strains to *Mycobacterium tuberculosis* and relationship to major histocompatibility complex haplotype and Nramp1 genotype. *Immunology* **93**, 270–274 (1998).
32. Gill, W. P. *et al.* A replication clock for *Mycobacterium tuberculosis*. *Nat. Med.* **15**, 211–214 (2009).
33. Capuano, S. V. *et al.* Experimental *Mycobacterium tuberculosis* infection of cynomolgus macaques closely resembles the various manifestations of human *M. tuberculosis* infection. *Infect. Immun.* **71**, 5831–5844 (2003).
34. Lin, P. L. *et al.* Quantitative comparison of active and latent tuberculosis in the cynomolgus macaque model. *Infect. Immun.* **77**, 4631–4642 (2009).
35. Vallania, F. *et al.* Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nat. Commun.* **9**, 4735 (2018).
36. Scriba, T. J. *et al.* Sequential inflammatory processes define human progression from *M. tuberculosis* infection to tuberculosis disease. *PLOS Pathog.* **13**, e1006687 (2017).
37. Domaszewska, T. *et al.* Concordant and discordant gene expression patterns in mouse strains identify best-fit animal model for human tuberculosis. *Sci. Rep.* **7**, 1–13 (2017).
38. Haynes, W. A. *et al.* Empowering multi-cohort gene expression analysis to increase reproducibility. *Pac. Symp. Biocomput.* **22**, 144–153 (2016).
39. Poulsen, A. Some clinical features of tuberculosis. *Acta Tuberc. Scand.* **33**, 37–92 (1957).
40. Poulsen, A. Some clinical features of tuberculosis. 1. Incubation period. *Acta Tuberc. Scand.* **24**, 311–346 (1950).
41. Gedde-Dahl, T. Tuberculous infection in the light of tuberculin matriculation. *Am. J. Hyg.* **56**, 139–214 (1952).
42. McCarthy, O. R. Asian immigrant tuberculosis–the effect of visiting Asia. *Br. J. Dis. Chest* **78**, 248–253 (1984).
43. Hatherell, H.-A. *et al.* Declaring a tuberculosis outbreak over with genomic epidemiology. *Microb. Genom.* https://doi.org/10.1099/mgen.0.000060 (2016).

44. Cobelens, F. *et al.* From latent to patent: rethinking prediction of tuberculosis. *Lancet Respir. Med.* **5**, 243–244 (2017).
45. Gupta, R. K. *et al.* Concise whole blood transcriptional signatures for incipient tuberculosis: a systematic review and patient-level pooled meta-analysis. *Lancet Respir. Med.* **8**, 395–406 (2020).
46. du Plessis, N. *et al.* Increased frequency of myeloid-derived suppressor cells during active tuberculosis and after recent *Mycobacterium tuberculosis* infection suppresses T-cell function. *Am. J. Respir. Crit. Care Med.* **188**, 724–732 (2013).
47. Weiner, J. *et al.* Changes in transcript, metabolite and antibody reactivity during the early protective immune response in humans to *Mycobacterium tuberculosis* infection. *Clin. Infect. Dis.* **71**, 30. https://doi.org/10.1093/cid/ciz785 (2020).
48. Esmail, H. *et al.* Characterization of progressive HIV-associated tuberculosis using 2-deoxy-2-[18F]fluoro-D-glucose positron emission and computed tomography. *Nat. Med.* **22**, 1090–1093 (2016).
49. Hwang, E. J. *et al.* Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. *Clin. Infect. Dis.* **69**, 739–747 (2019).
50. Coleman, M. T. *et al.* Early changes by 18 fluorodeoxyglucose positron emission tomography coregistered with computed tomography predict outcome after *Mycobacterium tuberculosis* infection in cynomolgus macaques. *Infect. Immun.* **82**, 2400–2404 (2014).
51. Lin, P. L. & Flynn, J. L. The end of the binary era: revisiting the spectrum of tuberculosis. *J. Immunol.* **201**, 2541–2548 (2018).
52. Singhania, A., Wilkinson, R. J., Rodrigue, M., Haldar, P. & O'Garra, A. The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis. *Nat. Immunol.* **19**, 1159–1168 (2018).
53. Van Der Meeren, O. *et al.* Phase 2b controlled trial of M72/AS01E vaccine to prevent tuberculosis. *N. Engl. J. Med.* **379**, 1621–1634 (2018).
54. Lu, L. L. *et al.* IFN-γ-independent immune markers of *Mycobacterium tuberculosis* exposure. *Nat. Med.* **15**, 17. https://doi.org/10.1038/s41591-019-0441-3 (2019).
55. Weiner, J. *et al.* Metabolite changes in blood predict the onset of tuberculosis. *Nat. Commun.* **9**, 5208 (2018).
56. Vesosky, B., Rottinghaus, E. K., Davis, C. & Turner, J. CD8 T cells in old mice contribute to the innate immune response to *Mycobacterium tuberculosis* via interleukin-12p70-dependent and antigen-independent production of gamma interferon. *Infect. Immun.* **77**, 3355–3363 (2009).
57. Cyktor, J. C. *et al.* Killer cell lectin-like receptor G1 deficiency significantly enhances survival after *Mycobacterium tuberculosis* infection. *Infect. Immun.* **81**, 1090–1099 (2013).
58. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
59. Liaw, A. & Wiener, M. Classification and regression by random forest. *R News* **2**, 18–22 (2002).
60. Kuhn, M. *et al. caret: Classification and Regression Training*. (https://CRAN.R-project.org/package=caret, 2018).
61. Wright, M. N. & Ziegler, A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* https://doi.org/10.18637/jss.v077.i01 (2017).
62. Greenwell, B., Boehmke, B., Cunningham, J. & Developers (https://github.com/gbm-developers), G. B. M. *gbm: Generalized Boosted Regression Models* (https://CRAN.R-project.org/package=gbm, 2018).
63. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab—an S4 package for kernel methods in R. *J. Stat. Softw.* **11**, 1–20 (2004).
64. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
65. Sweeney, T. E., Braviak, L., Tato, C. M. & Khatri, P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *Lancet Respir. Med.* **4**, 213–224 (2016).
66. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
67. Li, S. *et al.* Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* **15**, 195–204 (2014).
68. Domaszewska, T. & Weiner, J. *disco: Discordance and Concordance of Transcriptomic Responses*. (https://CRAN.R-project.org/package=disco, 2018).
69. Weiner 3rd, J. & Domaszewska, T. *tmod: an R package for general and multivariate enrichment analysis*. https://peerj.com/preprints/2420 (2016) https://doi.org/10.7287/peerj.preprints.2420v1.
70. NCAR - Research Applications Laboratory. *verification: Weather Forecast Verification Utilities*. (https://CRAN.R-project.org/package=verification, 2015).
71. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* **12**, 77–84 (2011).
72. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
73. Ault, R. C. *et al.* Blood RNA Signatures Predict Recent Tuberculosis Exposure in Mice, Macaques and Humans. *bioRxiv* (2019) https://doi.org/10.1101/830794.

Acknowledgements

We thank the personnel of the Ohio State University Animal BSL3 laboratory for assistance, including J. Cyktor for technical help with the mouse experiments. We acknowledge the efforts of the Baylor Institute for Immunology Research Genomics core for assistance with sample preparation and microarray processing. We thank E. Kautto and Q. Hassan for consultation with the mouse data analysis. We thank L. Schlesinger and L. Barreiro for critical review of the manuscript. This manuscript has been released as a Pre-Print at bioRxiv[73].

Author contributions

R.C.A. conceived the idea, designed the study and data analysis, analyzed the data and wrote the manuscript. C.A.H. performed the IPA analysis. A.E.H. contributed to the macaque data analysis. B.J.C. performed the mouse experiment. B.J.C. and A.M. prepared samples for RNA microarray analysis. J.T. oversaw the study and data analysis. R.C.A., C.A.H. and J.T. revised the manuscript. All authors commented on the manuscript.

Funding

This work was supported by NIH grant no. AI064522 (to J.T.) and a Texas Biomedical Forum grant (to J.T.). R.C.A. was supported by the Ohio State University Dean's Distinguished University Fellowship.

Competing interests

R.C.A. is inventor on pending patents filed by Texas Biomedical Research Institute for using RNA expression to determine duration of mycobacterial infection, US provisional Patent No. 62/768,708, PCT Patent Application No. PCT/US19/61895. All other authors have no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-73942-z.

Correspondence and requests for materials should be addressed to J.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.