A blood RNA signature for tuberculosis disease risk: a prospective cohort study


Summary

Background Identification of blood biomarkers that prospectively predict progression of Mycobacterium tuberculosis infection to tuberculosis disease might lead to interventions that combat the tuberculosis epidemic. We aimed to assess whether global gene expression measured in whole blood of healthy people allowed identification of prospective signatures of risk of active tuberculosis disease.

Methods In this prospective cohort study, we followed up healthy, South African adolescents aged 12–18 years from the adolescent cohort study (ACS) who were infected with M tuberculosis for 2 years. We collected blood samples from study participants every 6 months and monitored the adolescents for progression to tuberculosis disease. A prospective signature of risk was derived from whole blood RNA sequencing data by comparing participants who developed active tuberculosis disease (progressors) with those who remained healthy (matched controls). After adaptation to multiplex quantitative real-time PCR (qRT-PCR), the signature was used to predict tuberculosis disease in untreated adolescent samples and in samples from independent cohorts of South African and Gambian adult progressors and controls. Participants of the independent cohorts were household contacts of adults with active pulmonary tuberculosis disease.

Findings Between July 6, 2005, and April 23, 2007, we enrolled 6363 participants from the ACS study and 4466 from independent South African and Gambian cohorts. 46 progressors and 107 matched controls were identified in the ACS cohort. A 16 gene signature of risk was identified. The signature predicted tuberculosis progression with a sensitivity of 66·1% (95% CI 63·2–68·9) and a specificity of 80·6% (79·2–82·0) in the 12 months preceding tuberculosis diagnosis. The risk signature was validated in an untouched group of adolescents (p=0·018 for RNA sequencing and p=0·0095 for qRT-PCR) and in the independent South African and Gambian cohorts (p values <0·0001 by qRT-PCR), the signature was used to predict tuberculosis disease in untreated adolescent samples and in samples from independent cohorts of South African and Gambian adult progressors and controls. The whole blood tuberculosis risk signature prospectively identified people at risk of developing active tuberculosis disease, opening the possibility for targeted intervention to prevent the disease.

Interpretation The whole blood tuberculosis risk signature prospectively identified people at risk of developing active tuberculosis disease during their lifetime; most individuals will remain healthy.

Introduction A third of the population worldwide is infected with Mycobacterium tuberculosis,1 but less than 10% of these individuals will progress to have active tuberculosis disease during their lifetime; most individuals will remain healthy.2–6 Risk of progression is associated with age,7 comorbidities such as HIV infection and diabetes mellitus, socioeconomic and nutritional compromise, and therapy with immune modulatory drugs such as tumour necrosis factor inhibitors, among others.7–9 Current assays for determining the presence of M tuberculosis infection, such as an interferon gamma release assay (IGRA) or tuberculin skin test (TST), cannot predict which infected individuals will develop active tuberculosis.

Previous systems biology approaches have identified diagnostic signatures that discriminate tuberculosis disease from latent M tuberculosis infection and from other disease states.10–21 For example, Berry and colleagues22 identified and validated a 393 gene signature that allowed differentiation of people with active tuberculosis disease and latent infection. Anderson and colleagues23 identified and validated a 53 gene signature that distinguished active tuberculosis from other diseases in African children with or without HIV infection. By contrast with the published diagnostic studies, our focus was on prospective signatures of risk that could be identified in healthy individuals up to 2 years before clinical tuberculosis disease manifests.

Knowledge gained from this signature could lead to targeted antimicrobial therapy to prevent tuberculosis disease, as treating all people who are latently infected in endemic countries for 6–9 months is not feasible.
Evidence before this study
We searched the PubMed database for studies published before July 1, 2015, using the search criteria “tuberculosis AND risk AND blood AND (RNA OR microarray OR transcriptome OR RNA-Seq)”. The resulting scientific literature included several substantial analyses comparing the blood RNA profiles of individuals with active tuberculosis disease and healthy individuals. These important studies have established that the tuberculosis disease state is reflected in the blood RNA profile of the patient with tuberculosis with ongoing disease. The resulting literature also included several small studies reporting candidate host markers for tuberculosis disease risk that have not been rigorously assessed in independent cohorts. Repeating the search without the “(RNA OR microarray OR transcriptome OR RNA-Seq)” term yielded literature with established risk factors for tuberculosis disease, which have been summarised in several reviews. Despite these important studies and known tuberculosis risk factors, it is not possible to predict which individuals infected with Mycobacterium tuberculosis will develop active tuberculosis with tools.

Added value of this study
Our study expands the previous findings by being the first large-scale search for prospective correlates of risk of tuberculosis in healthy individuals before the onset of disease. We used unbiased high-throughput screening of host blood RNA profiles to identify new signatures of risk for tuberculosis. These signatures were confirmed in the original cohort with targeted assays; these targeted assays were then successfully used to predict tuberculosis disease progression in two independent cohorts. Further meta-analyses of published datasets showed that the prognostic signatures might simultaneously serve as diagnostic signatures for tuberculosis.

Implications of all the evidence
Our study provides the first demonstration that host blood RNA signatures can be used to predict progression to active tuberculosis disease in healthy individuals that are latently infected with or exposed to the M tuberculosis pathogen. These findings will result in further follow-up studies to assess whether the prognostic signatures can be used to prevent tuberculosis disease through targeted prophylactic treatment. Additional follow-up studies might focus on optimising the practical measurement of the signatures and understanding the biological significance of the host genes implicated by the signatures.

Research in context

Evidence before this study
We searched the PubMed database for studies published before July 1, 2015, using the search criteria “tuberculosis AND risk AND blood AND (RNA OR microarray OR transcriptome OR RNA-Seq)”. The resulting scientific literature included several substantial analyses comparing the blood RNA profiles of individuals with active tuberculosis disease and healthy individuals. These important studies have established that the tuberculosis disease state is reflected in the blood RNA profile of the patient with tuberculosis with ongoing disease. The resulting literature also included several small studies reporting candidate host markers for tuberculosis disease risk that have not been rigorously assessed in independent cohorts. Repeating the search without the “(RNA OR microarray OR transcriptome OR RNA-Seq)” term yielded literature with established risk factors for tuberculosis disease, which have been summarised in several reviews. Despite these important studies and known tuberculosis risk factors, it is not possible to predict which individuals infected with Mycobacterium tuberculosis will develop active tuberculosis with tools.

Added value of this study
Our study expands the previous findings by being the first large-scale search for prospective correlates of risk of tuberculosis in healthy individuals before the onset of disease. We used unbiased high-throughput screening of host blood RNA profiles to identify new signatures of risk for tuberculosis. These signatures were confirmed in the original cohort with targeted assays; these targeted assays were then successfully used to predict tuberculosis disease progression in two independent cohorts. Further meta-analyses of published datasets showed that the prognostic signatures might simultaneously serve as diagnostic signatures for tuberculosis.

Implications of all the evidence
Our study provides the first demonstration that host blood RNA signatures can be used to predict progression to active tuberculosis disease in healthy individuals that are latently infected with or exposed to the M tuberculosis pathogen. These findings will result in further follow-up studies to assess whether the prognostic signatures can be used to prevent tuberculosis disease through targeted prophylactic treatment. Additional follow-up studies might focus on optimising the practical measurement of the signatures and understanding the biological significance of the host genes implicated by the signatures.

Other potential applications of biomarkers of risk of tuberculosis disease include assessment of response to drug therapy and targeted enrolment into efficacy trials of new tuberculosis vaccines and drugs. In view of the fact that a third of the world’s population is latently infected with M tuberculosis, our approach constitutes an opportunity to lessen the burden of disease. To this end, we aimed to assess whether global gene expression measured in whole blood of healthy people allowed identification of prospective signatures of risk of active tuberculosis disease.

Methods

Study design and participants
We included participants from several cohorts in this analysis. First, we assessed participants aged 12–18 years from the South African adolescent cohort study (ACS) who were infected with M tuberculosis to identify and validate a tuberculosis risk signature (figure 1). All adolescents whose parents or legal guardians provided written, informed consent and who provided written, informed consent themselves were enrolled. About half the participants from the ACS cohort were assessed at enrolment and every 6 months during 2 year follow-up; the other half were assessed at baseline and at 2 years. At enrolment and at each visit, clinical data were collected and 2·5 mL blood was collected directly into PAXgene blood RNA tubes (PreAnalytiX, Hombrechtikon, Switzerland), which were stored at −20°C.

Only adolescents with latent M tuberculosis infection at enrolment were included in the analysis aimed at identification of a tuberculosis risk signature. Latent M tuberculosis infection was diagnosed by a positive QuantiFERON TB gold in-tube assay (Cellestis, Chadstone, Australia; >0·35 IU/mL) or a positive tuberculin skin test (0·1 mL dose of purified protein derivative RT-23, 2-TU, Staten Serum Institute, Denmark; >10 mm), or both. According to South African policy, adolescents positive on these tests were not given therapy to prevent tuberculosis disease. Adolescents who developed active tuberculosis disease during follow-up were included as progressors. Tuberculosis was defined as intrathoracic disease, with either two sputum smears positive for acid-fast bacilli or one positive sputum culture confirmed as M tuberculosis complex (mycobacterial growth indicator tube, BD BioSciences, NJ, USA). For each progressor, two matched controls that remained healthy during follow-up were selected and matched by age at enrolment, sex, ethnic origin, school of attendance, and presence or absence of previous episodes of tuberculosis disease. Participants were excluded if they developed tuberculosis disease within 6 months of enrolment to exclude early asymptomatic disease that could have been present at the time of assessment, or if they were HIV-positive. Before analysis, the ACS progressors and controls were randomly divided into training and test sets, at a ratio of 3:1 using the randomisation function in Excel.
Figure 1: The adolescent cohort study and the Grand Challenges 6–74 study cohorts for the discovery and validation of the tuberculosis risk signature

(A) Inclusion and exclusion of participants from the adolescent cohort study and assignment of eligible progressors and controls to the training and test sets.

- 6363 participants included from the adolescent cohort study
- 3595 were positive for either QFT or TST, or both
- 2349 were negative for both QFT and TST
- 22 excluded as QFT and TST unknown
- 60 developed intrathoracic tuberculosis
- 27 developed intrathoracic tuberculosis
- 20 excluded as not per-protocol tuberculosis
- 17 excluded as developed tuberculosis within 6 months of QFT or TST conversion
- 6 excluded, incomplete sample set
- 46 tuberculosis progressors
- 46 tuberculosis progressors
- 6 had tuberculosis >6 months after Mycobacterium tuberculosis infection
- 60 developed intrathoracic tuberculosis
- 37 progressors
- 77 controls
- 9 progressors
- 30 controls

(B) Inclusion and exclusion of adult household contacts of patients with lung tuberculosis from the Grand Challenges 6–74 study cohorts, and assignment of eligible progressors and controls.

- 4466 participants included from the Grand Challenges 6–74 cohort study
- 1948 from The Gambia
- 1197 from South Africa
- 34 developed intrathoracic tuberculosis
- 43 South Africa progressors
- 30 The Gambia progressors
- 1914 remained healthy
- 172 matched healthy controls to progressors (4:1) by age, sex, and recruitment year
- 172 South Africa controls
- 172 South Africa progressors
- 1321 from other African countries
- 129 The Gambia controls

QFT=QuantiFERON TB gold in-tube assay. TST=tuberculin skin test.
The other cohorts consisted of South African and Gambian participants from the Grand Challenges 6-74 study (GC6-74) who were enrolled to independently validate the tuberculosis risk signature (figure 1). Briefly, from a parent GC6-74 cohort, HIV-negative people aged 10–60 years who had household exposure to an adult with sputum smear positive tuberculosis disease were enrolled to this study. At baseline (both sites), at 6 months (The Gambia only), and at 18 months (both sites), participants were assessed clinically and blood was collected and stored in PAXgene tubes. Follow-up continued for 2 years, and concluded on Nov 18, 2012. Among GC6-74 participants, progressors had intrathoracic tuberculosis, defined on the basis of sputum culture, smear microscopy, and clinical signs. For each progressor, four controls were matched according to recruitment region, age category (±18 years, 19–25 years, 26–35 years, or ≥36 years), sex, and year of enrolment.

The study protocols were approved by the relevant human research ethics committees (appendix 1). Written informed consent was obtained from participants. For adolescents, consent was obtained from parents or legal guardians of adolescents and written informed assent from each adolescent. In both studies, participants with diagnosed or suspected tuberculosis disease were referred to a study-independent public health physician for treatment according to national tuberculosis control programmes of the country involved.

Procedures

The analytical approach to identify and validate the tuberculosis risk signature is shown (figure 2). The tuberculosis risk signature was derived from mining RNA sequencing data generated from the ACS training set. The RNA sequencing based tuberculosis risk signature was then adapted to the quantitative real-time PCR (qRT-PCR) platform. The RNA sequencing and qRT-PCR-based signature of risk was validated by blind prediction on untouched samples from the ACS test set. The qRT-PCR-based tuberculosis risk signature was also validated by blind prediction on independent GC6-74 cohort samples from South Africa and The Gambia.

For RNA sequencing analysis of the ACS training and test sets, RNA was extracted from PAXgene tubes of the ACS training set. Globin transcript depletion (GlobinClear, ThermoFisher Scientific, MA, USA) was followed by cDNA library preparation using the Illumina mRNA-Seq sample preparation kit (Illumina, CA, USA). RNA sequencing was done by Expression Analysis, Q2 Solutions, NC, USA, at 30 million 50 bp paired-end reads, on Illumina HiSeq-2000 sequencers. Read pairs were aligned to the hg19 human genome using gsnapti, which generated a table of gene expression abundances for each sample. This gene expression abundance was measured at the level of splice junction counts, which quantifies the frequency of specific mRNA splicing events in expressed genes; this approach would help with translation to qRT-PCR. For simplicity, splice junction expression levels are referred to as “gene expression levels” throughout.

The tuberculosis risk signature was generated and was adapted from the original RNA sequencing-based platform to qRT-PCR by directly matching splice junctions in the signature to commercial TaqMan primer sets (Thermo Fisher Scientific; appendix 1). A complete set of qRT-PCR data for selected primers was generated for ACS training set samples with the BioMark HD multiplex microfluidic instrument (Fluidigm, CA, USA). Variables in the qRT-PCR-based version of the tuberculosis signature were then assigned by fitting the model to the dataset.

RNA sequencing and qRT-PCR analysis of samples from the ACS test set were done as described. Before analysis, all test set samples were assigned random numerical codes that masked study timepoints and progressor and control status. Prediction of tuberculosis risk on the masked ACS test set samples was then done in a fully blinded manner, in parallel with RNA sequencing and qRT-PCR-based versions of the signature.

We blindly predicted the independent GC6-74 validation cohorts using the qRT-PCR-based signature of risk. qRT-PCR analysis of samples from the South African and Gambian cohorts of GC6-74 was done as described less than 1 year after ACS validation analysis. Before analysis, all samples were assigned random numerical codes. Fully blinded predictions were then made with the qRT-PCR-based signature of risk.

To allow assessment of the risk signature for diagnosis of active disease, results from published microarray-based studies of active tuberculosis versus latent disease or other disease were used.19,20 The signature was adapted from RNA sequencing to the Illumina platform and parameterised using tuberculosis cases and controls latently infected with M tuberculosis from the UK training cohort of Berry and colleagues21 (appendix 1). The locked-down Illumina microarray-based risk signature was used to make predictions in the independent test and validation cohorts from the study by Berry and colleagues,22 and from the subsequent studies.23,24,25

Statistical analysis

Statistical and machine learning approaches were applied to discover the signatures of tuberculosis risk. To generate the tuberculosis risk signature, we used an extension of the k-top-scoring pairs (k-TSP) method, which has been used successfully to identify cancer biomarkers.26,27 The k-TSP approach identifies pairs of genes that discriminate better than either gene would individually.28 We replaced the k-TSP rank-based gene pair models with the so-called support vector machine (SVM)-based gene pair models for greater flexibility in predictions. This modification is similar to the k-TSP modification proposed by Shi and colleagues,29 but holds the advantage of retaining the fault-tolerance and parsimony of k-TSP. For analysis, prospective RNA sequencing data of progressors was realigned to the timepoint at which active tuberculosis
was diagnosed (figure 2), thereby synchronising the cohort with respect to outcome.

The genes that comprise the final tuberculosis risk signature were selected in two stages, with data from the ACS training set. First, a large set of genes was identified by comparing gene expression in progressors at the most proximal timepoint to diagnosis with that in matched controls. SVM models were trained on these datapoints for all possible pairwise combinations of risk-associated genes. Second, the models were filtered for predictive accuracy with the remaining prospective progressor and control samples. Surviving SVM models comprised the tuberculosis risk signature, which computes a “tuberculosis risk score” based on blood gene expression levels measured at a single timepoint. The algorithm is fully described (appendix 1).

These analyses were executed using R or custom programs written in C++. Application of final
signatures to predict tuberculosis risk was carried out using scripts written in R or an Excel spreadsheet (appendix 2). Statistical evaluation of prediction performance was done by analysis receiver operating characteristic curves (ROCs) using the R package pROC.27

A

B

C ACS training

D ACS test

E GC6-74

F GC6-74
Role of the funding source
The funders of this study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. DEZ, AP-N, TJ S, ET, LMA, AA, and WAH had full access to all the data in the study; The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results
Between July 6, 2005, and April 23, 2007, we enrolled 6363 healthy adolescents from the ACS cohort; follow-up was completed by February, 2009 (appendix 1). Between Feb 27, 2005, and Dec 14, 2010, (South Africa) and between March 5, 2007, and Oct 21, 2010 (The Gambia), we enrolled 4466 healthy individuals from the GC6-74 cohort, of these 1197 were enrolled in the South African and 1948 enrolled in The Gambia (appendix 1).

46 ACS participants with microbiologically confirmed tuberculosis were identified as progressors (figure 1; appendix 2). For progressors, the time between sample collection and diagnosis with active tuberculosis (“time to diagnosis”) ranged from 1 to 894 days (figure 2; appendix 2). 107 control participants who were infected with *M tuberculosis* at enrolment but who remained healthy during 2 years of follow-up were matched to progressors. Before analysis, progressors and controls were randomly divided into a training set of 37 progressors and 77 controls, and a test set of nine progressors and 30 controls (figures 1, 2). Progressor and control participants in GC6-74 were household contacts of newly diagnosed index cases with pulmonary tuberculosis disease (figure 2). Two GC6-74 sites, South Africa and The Gambia, had sufficient numbers of progressors and controls to allow analysis, and were therefore included in this study (figures 1, 2). 43 progressors and 172 controls were identified at the South African site, whereas 30 progressors and 129 controls were identified at the Gambian site (figure 1; appendix 2).

RNA was isolated from all progressor and matched control samples of the ACS training set and analysed by RNA sequencing (figure 2; appendix 2). Data mining of the RNA sequencing data derived a candidate signature of risk for tuberculosis disease progression. The signature comprised paired splice junction data from 16 genes (appendices 1, 2). Expression of signature genes in samples from progressors increased as tuberculosis diagnosis approached (figure 3). Robust discrimination between progressors and controls based on the expression of the gene pairs in the signature was readily apparent (figure 3).

The predictive potential of the tuberculosis risk signature was shown within the ACS training set by cross validation (figure 2; appendix 1); the risk signature achieved 71·2% sensitivity in the 6 month period immediately before diagnosis, and 62·9% sensitivity 6–12 months before diagnosis, at a specificity of 80·6% (figure 3, table 1). During the 12–18 month period before diagnosis, the signature achieved 47·7% sensitivity.

To help with broad application, the tuberculosis risk signature was adapted to a practical platform, qRT-PCR (figure 2; appendix 2). The RNA sequencing and qRT-PCR versions of the tuberculosis risk signature were used to predict tuberculosis risk in the untouched ACS test set samples. This was done in a fully blinded manner, with all sample meta-data masked before making predictions. The ability of both versions of the signature to predict tuberculosis progression in healthy subjects was validated (figure 3, table 2).

To determine whether inclusion of a larger number of genes would have increased accuracy of predictions, the performance of a random forest-based classifier comprised of 631 genes was assessed; the outcome was

<table>
<thead>
<tr>
<th>By 6 month period</th>
<th>ROC AUC (95% CI)</th>
<th>Sensitivity (95% CI)</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-180</td>
<td>0.79 (0.76–0.82)</td>
<td>71.2% (66.6–75.2)</td>
<td>61%</td>
</tr>
<tr>
<td>181-360</td>
<td>0.77 (0.75–0.79)</td>
<td>62.9% (59.0–66.4)</td>
<td>61%</td>
</tr>
<tr>
<td>361-540</td>
<td>0.72 (0.70–0.76)</td>
<td>47.7% (42.9–52.5)</td>
<td>61%</td>
</tr>
<tr>
<td>541-720</td>
<td>0.54 (0.49–0.59)</td>
<td>29.1% (23.1–35.9)</td>
<td>61%</td>
</tr>
<tr>
<td>&gt;720</td>
<td>0.49 (0.43–0.56)</td>
<td>5.4% (2.4–13.0)</td>
<td>61%</td>
</tr>
<tr>
<td>By 12 month period</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-360</td>
<td>0.779 (0.76–0.80)</td>
<td>66.1% (63.2–68.9)</td>
<td>61%</td>
</tr>
<tr>
<td>360-720</td>
<td>0.647 (0.62–0.673)</td>
<td>37.5% (33.9–41.2)</td>
<td>61%</td>
</tr>
<tr>
<td>Total time period</td>
<td>0.743 (0.73–0.76)</td>
<td>58.4% (56.1–60.7)</td>
<td>61%</td>
</tr>
</tbody>
</table>

Sensitivity values are reported at a specificity of 80% (95% CI). ROC AUC=area under receiver operating characteristic curve. ACS=adolescent cohort study.
equivalent to when the tuberculosis risk signature was used for classification (appendices 1, 2).

For independent validation, the qRT-PCR-based signature was used to predict tuberculosis progression using samples collected from healthy participants in the GC6-74 adult household contact cohorts from South Africa and The Gambia (figure 2). Predictions were made in a blinded manner. The ability of the signature to predict tuberculosis progression in healthy participants were validated in these independent cohorts, irrespective of whether these were analysed individually or collectively (p<0·0001; figure 3, table 2). As for the ACS, the signature had greater sensitivity for predicting tuberculosis in samples collected closer to the time of diagnosis (figure 3, table 2).

Because the sensitivity of the tuberculosis risk signature increased as the time of diagnosis approached, we assessed performance of the risk signature for diagnosis of active tuberculosis disease. We did these analyses after adapting the signature to Illumina microarrays using data from the UK training cohort of Berry and colleagues (appendices 1, 2),27 which enabled use of published datasets.28–31 The signature readily differentiated active tuberculosis from latent infection in adult cohorts from the UK, South Africa, and Malawi, including in populations that were co-infected with HIV (appendices 1, 2). The signature also discriminated active tuberculosis from other pulmonary diseases (appendices 1, 2). Despite being derived from adolescents, the signature discriminated active, culture confirmed, tuberculosis from latent M tuberculosis infection and from other diseases in childhood (appendices 1, 2). Finally, applying the signature to data from a treatment study32 showed that the active tuberculosis signature gradually disappears during 6 months of therapy (appendix 1, 2).

**Discussion**

Roughly one third of the world’s population might harbour latent M tuberculosis infection and is at risk of active disease. We have identified a gene expression signature for predicting the risk of tuberculosis disease progression. This signature was discovered in a longitudinal analysis of South African adolescents with latent M tuberculosis infection who either developed tuberculosis disease or remained healthy. The signature was then validated in blinded samples from untouched adolescents of the same parent cohort. The signatures were again validated, in independent cohorts of longitudinally followed up household contacts of patients with tuberculosis disease from South Africa and The Gambia, who either developed tuberculosis disease or remained healthy. These results show that it is possible to predict progression from latent to active disease with whole blood gene expression measurements at any single timepoint up to 18 months before tuberculosis disease manifests.

The tuberculosis risk signature was discovered using RNA sequencing, a transcriptome analysis technology that is quantitative, sensitive, and unbiased.28 The signature was formulated using a framework termed SVM, an extension of the k-TSP approach,24 which robustly generates a tuberculosis risk score from gene expression data, with simple arithmetics (appendix 2). The signature was adapted from RNA sequencing to qRT-PCR, a more targeted and affordable technology. The power of the approach was shown by blinded validation of the qRT-PCR-based signature in the independent cohorts.

The tuberculosis risk signature predicted tuberculosis disease progression despite marked diversity between the ACS and GC6-74 cohorts. This result is encouraging in view of the different age ranges (adolescents vs adults), different infection or exposure status, distinct ethnic origin and genetic backgrounds,33–35 differing local epidemiology,1 and differing circulating strains of M tuberculosis36 between South Africa and The Gambia. Distinct mechanisms of progression might be elucidated when specific subgroups of progressors are analysed (eg, early vs late progressors in GC6-74). Targeted analyses to identify distinct mechanisms of progression are underway.
To explore potential application of the signature for targeted preventive therapy, we estimated the relative risk for tuberculosis disease between signature positive and negative people from a representative adult population from South Africa, where tuberculosis is endemic. The relative risk of tuberculosis disease is about 2 when IGRA or TST is used, whereas the relative risk using our risk signature was between 6 and 14. Moreover, this risk signature would aid in detection of asymptomatic or undiagnosed tuberculosis disease, or both. For example, when applied to combined data from four studies of HIV-uninfected South African adults involving 130 prevalent tuberculosis cases and 230 controls, the signature discriminated between patients with active tuberculosis and uninfected or healthy controls infected with M tuberculosis with 87% sensitivity and 97% specificity.

Although our focus was on prospective prediction of tuberculosis disease, we also showed that the risk signature was excellent for differentiating tuberculosis disease from latent infection and from other disease states. This ability to diagnose tuberculosis disease was not markedly affected by HIV status. The risk signature could also diagnose culture positive childhood tuberculosis, but not culture negative childhood disease. These results suggest that the risk signature might represent bacterial load in the lung because culture positive childhood tuberculosis is likely associated with higher bacterial loads compared with culture negative disease. An association between the risk signature and bacterial load was further supported by meta-analysis of a published treatment study, in which the signature relaxed during 6 months of antimicrobial therapy. It is presently not known whether the risk signature will be useful for predicting treatment failure or recurrence.

While enrichment analysis of published blood signatures for active tuberculosis implicates various biological processes in the disease, the sole gene module that was over-represented in the risk signature was interferon response (appendix 2). Overlap between 15 of the 16 genes in our prospective tuberculosis risk signature and the 393 gene signature of active tuberculosis disease from Berry and colleagues suggests that chronic peripheral activation of the interferon response precedes the onset of active disease and the inflammatory manifestations of tuberculosis disease shown by previously published gene expression studies. Although additional research is needed to understand the functional role of interferon responses during tuberculosis progression, pathogen induction of type I interferons and their detrimental effects on immunity to tuberculosis have been shown in several in vivo studies in mice and in-vitro experiments of human cells. Nevertheless, not all interferon response genes in the risk signature might be associated with a poor outcome, since genes such as GBP1, STAT1, and TAP1 might have a protective role during tuberculosis infection (appendix 2).

Our predictive signature was obtained from transcriptional analysis of peripheral whole blood. This compartment, although conveniently sampled, might not accurately represent pathogenic events in the lung, mainly affected by tuberculosis disease. Irrespective of this, circulating white blood cells can serve as sentinels of lung pathophysiology, as transcriptional changes occur when the cells migrate through this organ. To explore a possible cell-type specific origin of the risk signature, we used data from published global gene expression in whole blood and sorted peripheral blood mononuclear cells, monocytes, neutrophils, and T cells from healthy controls and patients with tuberculosis.

Differential expression of the risk signature genes between healthy controls and patients with tuberculosis was similar in whole blood and peripheral blood mononuclear cells (appendix 2), suggesting that contribution of granulocytes to the risk signature is redundant. Consistent with this hypothesis, differential expression of risk signature genes was apparent in monocytes and neutrophils. When compared to the diagnostic signature of Berry and colleagues, reported to be derived from neutrophils, these results suggest that progression to active tuberculosis involves more diverse cell types.

So far, Sloot and colleagues published the only report of prospective associations between blood gene expression and tuberculosis disease risk. With a predefined 141 gene panel, peripheral blood mononuclear cell RNA expression in 15 HIV-positive drug users who developed active tuberculosis disease was compared with 16 who did not develop tuberculosis. Four genes assayed showed nominal expression differences (unadjusted p<0.05) and, when combined, two genes, IL-13 and AIRE, fit the data (area under receiver operating characteristic curve fit=0.8). The association between these genes and tuberculosis progression was not validated in a test set or independent cohort; none of the four genes showed differences between progressors and controls in our whole blood RNA sequencing datasets.

Our results, showing that blood-based signatures in healthy individuals can predict progression to active tuberculosis disease, pave the way for the establishment of diagnostic methods that are scalable and inexpensive. An important first step would be to test whether the signature can predict tuberculosis disease in the general population, rather than the select populations included in this project; for example, the risk of tuberculosis disease in our populations was much higher than the lifetime risk of 10% encountered in the general population. The newly described signature holds potential for highly targeted preventive therapy, and therefore for interrupting the worldwide epidemic.

**Contributors**

We acknowledge European Commission (EC) FP7 NEWTBVAC contract number HEALTH.F3.2009 247495 and EC HORIZON2020 project TBVAC2020. AP-N and SsA were supported by Postdoctoral Research Awards from The Carnegie Corporation of New York. AP-N was also supported by The Claude Leon Foundation and the Columbia University-Southern African Fogarty AIDS International Training and Research Program (AITRP) through the Fogarty International Center, National Institutes of Health (grant # D43 TW000231). SHEK also acknowledges support from European Union’s Seventh Framework Programme (EU FP7) project ADITEC (HEALTH-F4-2011-288073; the Innovative Medicines Initiative Joint Undertaking “Biomarkers for Enhanced Vaccine Safety” project BioVacSafe (grant number 115308) and the European Union’s Horizon 2020 project TBVAC2020 (grant number 643181).

References


