Four-Gene Pan-African Blood Signature Predicts Progression to Tuberculosis

Suliman, S.; Thompson, E.G.; Sutherland, J.; Weiner, J.; Ota, M.O.C.; Shankar, S.; ... ; Adolescent Cohort Study Grp

Title: Four-gene pan-African blood signature predicts progression to tuberculosis

Authors list:
Sara Suliman*[1], Ethan Thompson*[2], Jayne Sutherland[3], January Weiner 3rd[4], Martin O.C. Ota[3], Smitha Shankar[2], Adam Penn-Nicholson[1], Bonnie Thiel[5], Mzwandile Erasmus[1], Jeroen Maertzdorf[4], Fergal J. Duffy[2], Philip C. Hill[6], E. Jane Hughes[1], Kim Stanley[7], Katrina Downing[1], Michelle L. Fisher[1], Joe Valvo[2], Shreemanta K Parida[4], Gian van der Spuy[7], Gerard Tromp[7], Ifedayo M.O. Adetifa[3], Simon Donkor[3], Rawleigh Howe[8], Harriet Mayanja-Kizza[9], W. Henry Boom[5], Hazel Dockrell[10], Tom H.M. Ottenhoff[11], Mark Hatherill[1], Alan Aderem[2], Willem A. Hanekom[1], Thomas J. Scriba**[1], Stefan H. E. Kaufmann**[4], Daniel E. Zak**[2], Gerhard Walzl**#[7], and the GC6-74¶ and ACS§ cohort study groups

* and ** Contributed equally

#Corresponding author:

Gerhard Walzl, DST/NRF Centre of Excellence for Biomedical TB Research and MRC Centre for TB Research, Division of Molecular Biology and Human Genetics, Stellenbosch University, Tygerberg, South Africa
Tel. +27-21-938-9401
gwalzl@sun.ac.za

Affiliations:

[1]South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine & Division of Immunology, Department of Pathology, University of Cape Town, Cape Town, South Africa
[2]The Center for Infectious Disease Research, Seattle, WA, USA
[3]Vaccines and Immunity, Medical Research Council Unit, Fajara, The Gambia
[4]Max Planck Institute for Infection Biology, Berlin, Germany
[5]Case Western Reserve University, Cleveland, OH, USA
[6]Centre for International Health, School of Medicine, University of Otago, Dunedin, New Zealand
[7]DST/NRF Centre of Excellence for Biomedical TB Research and MRC Centre for TB Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Heath Sciences, Stellenbosch University, Tygerberg, South Africa
[8]Immunology Unit, Armauer Hansen Research Institute, Addis Ababa, Ethiopia
[9]Department of Medicine and Department of Microbiology, Makerere University, Kampala, Uganda
[10]Department of Immunology and Infection, London School of Hygiene and Tropical Medicine, London, UK
[11]Department of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands

¶The GC6-74 cohort study team:

DST/NRF Centre of Excellence for Biomedical TB Research and MRC Centre for TB Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa:
Gerhard Walzl, Gillian F. Black, Gian van der Spuy, Kim Stanley, Magdalena Kriel, Nelita Du Plessis, Nonhlanhla Nene, Andre G. Loxton, Novel N. Chegou, Gerhardus Tromp, David Tabb

Department of Infectious Diseases, Leiden University Medical Centre, Leiden, The Netherlands:
Tom H.M. Ottenhoff, Michel R. Klein, Marielle C. Haks, Kees L.M.C. Franken, Annemieke Geluk, Krista E van Meijgaarden, Simone A Joosten

Tuberculosis Research Unit, Department of Medicine, Case Western Reserve University School of Medicine and University Hospitals Case Medical Center, Cleveland, Ohio, USA:
W. Henry Boom, Bonnie Thiel

Department of Medicine and Department of Microbiology, College of Health Sciences, Faculty of Medicine, Makerere University, Kampala, Uganda:
Harriet Mayanja-Kizza, Moses Joloba, Sarah Zalwango, Mary Nsereko, Brenda Okwera, Hussein Kisingo

Department of Immunology, Max Planck Institute for Infection Biology, Berlin, Germany:
Stefan H.E. Kaufmann (GC6-74 Principal Investigator), Shreemanta K. Parida, Robert Golinski, Jeroen Maertzdorf, January Weiner 3rd, Marc Jacobson

Department of Immunology and Infection, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom:
Hazel Dockrell, Steven Smith, Patricia Gorak-Stolinska, Yun-Gyoung Hur, Maeve Lalor, Ji-Sook Lee

Karonga Prevention Study, Chilumba, Malawi:
Amelia C Crampin, Neil French, Bagrey Ngwira, Anne Ben-Smith, Kate Watkins, Lyn Ambrose, Felanji Simukonda, Hazzie Mvula, Femia Chilongo, Jacky Saul, Keith Branson

South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine & Division of Immunology, Department of Pathology, University of Cape Town, Cape Town, South Africa:

Sara Suliman, Thomas J. Scriba, Hassan Mahomed, E. Jane Hughes, Nicole Bilek, Katrina Downing, Michelle Fisher, Adam Penn-Nicholson, Humphrey Mulenga, Brian Abel, Mark Bowmaker, Benjamin Kagina, William Kwong Chung, Willem A. Hanekom

Aeras, Rockville, MD, USA:
Jerry Sadoff, Donata Sizemore, S Ramachandran, Lew Barker, Michael Brennan, Frank Weichold, Stefanie Muller, Larry Geiter

Ethiopian Health & Nutrition Research Institute, Addis Ababa, Ethiopia:
Desta Kassa, Almaz Abebe, Tsehayenesh Mesele, Belete Tegbaru

University Medical Centre, Utrecht, The Netherlands:
Debbie van Baarle, Frank Miedema

Armauer Hansen Research Institute, Addis Ababa, Ethiopia:
Rawleigh Howe, Adane Mihret, Abraham Aseffa, Yonas Bekele, Rachel Iwnetu, Mesfin Tafesse, Lawrence Yamuah

Vaccines & Immunity Theme, Medical Research Council Unit, Fajara, The Gambia:
Martin Ota, Jayne Sutherland, Philip Hill, Richard Adegbola, Tumani Corrah, Martin Antonio, Toyin Togun, Ifedayo Adetifa, Simon Donkor

Department of Infectious Disease Immunology, Statens Serum Institute, Copenhagen, Denmark:
Peter Andersen, Ida Rosenkrands, Mark Doherty, Karin Weldingh

Department of Microbiology and Immunology, Stanford University, Stanford, California, USA:
Gary Schoolnik, Gregory Dolganov, Tran Van

[§]The ACS cohort study team:

South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine & Division of Immunology, Department of Pathology, University of Cape Town, Cape Town, South Africa:
Fazlin Kafaar, Leslie Workman, Humphrey Mulenga, Thomas J. Scriba, E. Jane Hughes, Nicole Bilek, Yolundi Cloete, Deborah Abrahams, Sizulu Moyo, Sebastian Gelderbloem, Michele Tameris, Hennie Geldenhuys, Willem Hanekom, Gregory Hussey

School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa:
Rodney Ehrlich

KNCV Tuberculosis Foundation, The Hague, and Amsterdam Institute of Global Health and Development, Academic Medical Centre, Amsterdam, The Netherlands:
Suzanne Verver

Aeras, Rockville, MD, USA:
Larry Geiter

Author's Contributions:

SS, EGT, SHEK, PCH, WAH, GW, TJS and DEZ designed the study

SS, EGT, SHEK, GW, TJS and DEZ drafted the manuscript

SS, EGT, JS, JW, SSh, BT, APN, ME, JM, FJD, EJH, KS, KD, MLF, JV, GS, GT, IA, SD, RH, HMK and WHB contributed to sample and data management as well as data acquisition

SS, EGT, JS, JW, MOCO, SSh, BT, APN, ME, JM, FJD, HD, TO, MH, AA, WAH, TJS, SHEK, DEZ, GW and various members of the GC6-74 and ACS cohort study groups contributed to data analysis and interpretation

All authors reviewed, provided feedback and approved the manuscript and are accountable for the accuracy and integrity of the work

Funding:

The study was funded by the Bill & Melinda Foundation grants OPP1065330 and OPP1023483, OPP1055806 and GC6-74 Grant no. 37772, and grants from the National Institutes of Health (NIH) grants: R01AI087915, U01AI115619 and NO1AI095383/AI070022. The study was also supported by the Strategic Health Innovation Partnerships (SHIP) Unit of the South African Medical Research Council with funds received from the South African Department of Science and

Technology. AP-N and SSu were supported by Postdoctoral Research Awards from The Carnegie Corporation of New York. SSu was also supported by the South African National Research Foundation. 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, NIH (D43 TW000231). We also acknowledge funding by EC HORIZON2020 TBVAC2020 (Grant Agreement No. 643381) to THMO and SHEK.

Short Title:
Trans-African Prospective TB Biomarker

Subject category descriptor number: 11.4 Mycobacterial Disease: Host Defenses

Total word count: 3,459

At a glance commentary

Intervention against the tuberculosis (TB) epidemic requires a multi-pronged approach, including treatment and prevention. TB exists in a dynamic spectrum from latent infection to disease, and only about 5 to 10% of infected individuals develop clinical TB. Therefore, the reservoir for TB is huge since 1.7 billion people globally are estimated to be infected with the causative pathogen, *Mycobacterium tuberculosis* (*M.tb*). Consequently, identifying asymptomatic individuals who are at high risk of progressing to TB would help prioritize preventative strategies, which would provide an important step forward towards

better TB control. We developed a blood test to predict progression towards active TB in multiple Sub-Saharan African populations, following exposure to an index (active) TB patient living in the same household. The test surpassed published signatures in its ability to predict TB progression in different African cohorts. This simple 4-marker test could be translated into a simple, rapid and affordable point-of-care test for field application in resource-limited settings where TB and *M.tb* infection are endemic to identify individuals at high risk of developing TB. High-risk TB contacts could then be prioritized for prophylactic interventions.

Online data supplement: This article has an online data supplement, which is accessible from this issue's table of content online at www.atsjournals.org

Abstract

Rationale: Contacts of tuberculosis patients constitute an important target population for preventative measures as they are at high risk of infection with *Mycobacterium tuberculosis* and progression to disease.

Objectives: We investigated biosignatures with predictive ability for incident tuberculosis.

Methods: In a case-control study nested within the Grand Challenges 6-74 longitudinal African cohort of exposed household contacts, we employed RNA sequencing, polymerase chain reaction (PCR) and the Pair Ratio algorithm in a training/test set approach. Overall, 79 progressors, who developed tuberculosis between 3 and 24 months following exposure, and 328 matched non-progressors, who remained healthy during 24 months of follow-up, were investigated.

M*easurements and Main Results*: A four-transcript signature (RISK4), derived from samples in a South African and Gambian training set, predicted progression up to two years before onset of disease in blinded test set samples from South Africa, The Gambia and Ethiopia with little population-associated variability and also validated on an external cohort of South African adolescents with latent *Mycobacterium tuberculosis* infection. By contrast, published diagnostic or prognostic tuberculosis signatures predicted on samples from some but not all 3 countries, indicating site-specific variability.

Post-hoc meta-analysis identified a single gene pair, C1QC/TRAV27, that would consistently predict TB progression in household contacts from multiple

African sites but not in infected adolescents without known recent exposure events.

Conclusions: Collectively, we developed a simple whole blood-based PCR test to predict tuberculosis in household contacts from diverse African populations, with potential for implementation in national TB contact investigation programs.

Abstract word count: 244

MeSH key words: tuberculosis, gene expression, biomarkers

Introduction

Tuberculosis (TB), caused by infection with *Mycobacterium tuberculosis* (*M.tb*)[1,2], is the leading cause of death caused by a single pathogen globally[3]. Prior to development of symptomatic disease, latent *M.tb* infection can be detected by measuring immunological sensitization, using the tuberculin skin test (TST) and/or interferon gamma release assays (IGRA)[4]. Most infected individuals have effective defense mechanisms to control *M.tb*[5] as only 5-10% will progress to TB during their lifetime. Despite this, over 10 million new cases of TB are diagnosed each year and almost 2 million people die from the disease[3]. Although recent *M.tb* exposure and TST or IGRA conversion are associated with higher risk of TB progression[6], the positive predictive values of these tests are low, i.e. 1.5% and 2.7%[7], falling short of current WHO supported guidelines. Thus, the number of TST or IGRA-positive individuals requiring treatment to prevent progression to a single incident case of TB is prohibitively high[8].

Factors associated with elevated risk of progression to TB include age, sex, comorbidities[9,10], and especially being in recent contact with a patient with active pulmonary TB[11,12]. A biomarker that identifies HHC who will progress to TB would provide an opportunity to arrest disease progression through targeted prophylactic intervention[13,14]. Such prognostic biomarkers would be most impactful as point-of-care tests for resource-limited settings, such as those in Sub-Saharan Africa. Test performance should not be adversely affected by geographical diversity, as seen in Africa, which has a diversity of ethnic backgrounds[15] and circulating *M.tb* lineages[16]. A ‘TB-risk’ test must be practical

for field application and therefore based on accessible biological samples routinely used in clinical settings, such as peripheral blood[17].

Transcriptional profiling of blood cells has emerged as a powerful platform to discover potential TB biomarkers discriminating TB patients from healthy uninfected and/or latently *M.tb*-infected individuals[18-23]. We previously defined a 16-gene blood transcriptional correlate of risk (COR) signature that predicts risk of progression to TB in *M.tb*-infected HIV-negative South African adolescents and HHC from South Africa and The Gambia[24]. However, given that this COR signature was developed using a single cohort of latently *M.tb*-infected South African adolescents, the predictive accuracy for HHC in diverse African populations may be sub-optimal[24]. It would also be desirable to reduce the number of transcripts in the signature, to facilitate implementation of a low-cost point-of-care test.

In this study, we developed a simple blood RNA-based, four host-transcript signature (RISK4) for predicting risk of TB progression in HHC from diverse African cohorts. RISK4 was validated independently in distinct African populations from The Gambia, Ethiopia and two cohorts from South Africa. Furthermore, our study uniquely highlights signatures, as small as single transcript pairs, which were regulated in opposite directions in progressors and controls following HHC. These simple tests pave the way for cost-effective identification of individuals at highest risk for progression.

Methods

Study design and participants

All clinical sites adhered to the Declaration of Helsinki and Good Clinical Practice guidelines. Ethical approvals were obtained from institutional review boards **(Supplementary Table 1, and online supplement)**. The HHC study included participants from four African sites: South Africa, The Gambia, Ethiopia and Uganda, under the Bill and Melinda Gates Grand Challenges 6-74 (GC6-74) program **(Figure 1 and Supplementary Table 2)**. The Adolescent Cohort Study was described previously[24,25] and included IGRA+ and/or TST+ South African adolescents aged 12-18 years old with *M.tb* infection, occurring at unspecified times. Adult participants, or legal guardians of participants aged 10-17 years old, provided written or thumb-printed informed consent to participate after careful explanation of the study and potential risks.

Sample processing and RNA-sequencing

PAXgene (PreAnalytiX, Hombrechtikon, Switzerland) blood RNA samples were collected from all participants. Progressors were defined as individuals who developed TB 3-24 months post-HHC. Non-progressor samples were matched to the pre-diagnosis time points of each progressor by site, gender, age and recruitment year (**online supplement**). RNA-sequencing was performed by Beijing Genomics Institute (Shenzhen, China); additional details for processing and quality control are provided in the online supplement. FASTQ files have been deposited into the Gene Expression Omnibus[26] under accession GSE94438.

Identification of predictive signatures

Candidate site-specific signatures of risk for TB disease progression and final, simplified qRT-PCR-based candidate signatures were developed using the *Pair Ratios algorithm* (**online supplement**), which was previously described[27] and is a variation on the pairwise approach used to discover the ACS COR signature[24]. To summarize, the step-by-step procedure for computing the RISK4 signature scores using sample qRT-PCR measurements was:

1. Measure the cycle thresholds (Cts) for the four primer-probes (Applied Biosystems TaqMan Assays) listed in **Supplementary Table 3**.
2. For each of the four pairs of primer-probes, compute the difference in raw Ct, which produces the log-transformed ratio of expression.
3. Compare the measured ratio to ratios in the look-up table for the given pair of transcripts in **Supplementary Tables 4-7**. Find the minimal ratio in column 1 of the table that is greater than or equal to the measured ratio.
4. Assign the corresponding score in the second column of the look-up table to the ratio. If the measured ratio is larger than all ratios in column 1 of the look-up table, then assign a score of 1 to the ratio.
5. Compute the average over the scores generated from the set of pairs. If any assays failed on the sample, compute the average score over all ratios not including the failed assays. The resulting average is the final score for that sample.

Adaptation of published diagnostic signatures to qRT-PCR

The previously published signatures from Maertzdorf et al[28] and Sweeney et al[29] were adapted to the qRT-PCR platform, where we refer to them as DIAG4 and DIAG3, respectively. Primer-probe sets were selected for each gene in the respective signatures, and overall scores were computed for each sample as the difference in the mean of the up-regulated and the down-regulated transcripts (**Supplementary Tables 8-9).**

Results

We enrolled 4,466 HIV-negative healthy HHC of 1,098 index TB cases between 2006 and 2010 into the GC6-74 cohorts across 4 African sites **(Figure 1 and Supplementary Table 2)**. Samples were collected at enrolment/baseline, 6 and 18 months, with the exception of South Africa, where PAXgene blood RNA samples were collected at baseline and 18 months of follow-up, due to logistical limitations. Samples from Uganda were not available in sufficient quantities for this analysis **(Figure 1)**. TB incidence in HIV-negative healthy HHC was highest in South Africa, and lowest in Ethiopia **(Table 1)**, as defined by TB case classifications A-K in **Supplementary Table 10**. Incident cases (progressors) were defined as those who developed TB between 3 and 24 months following exposure. “Co-incident” cases, i.e. diagnosed with TB within 3 months of contact with the index case **(Methods)**, were not included in analysis. Prior TB was an exclusion criterion **(online supplement)**, thus progressors likely had their first TB episode during follow-up. Median age of progressors was comparable across the 4 African sites (Kruskal-Wallis $p=0.92$, **Table 1**). Median times to progression were 7 months in South Africa and Uganda, and 10.5 and 10 months in The Gambia and Ethiopia, respectively **(Table 1, and Supplementary Table 11A)**. Progressors, as defined by clinical symptoms, chest and other radiographs (CXR) consistent with TB and response to chemotherapy, without microbiological confirmation comprised 25% (4/12) of progressors in Ethiopia, 2% (1/43) in South Africa and 6% (3/34) in The Gambia (TB classification K, **supplementary Tables 10 and 11A)**.

A four-gene correlate of risk signature predicts TB progression in household contacts

We divided South African and Gambian HHC cohorts into training and test sets, while the entire Ethiopian cohort was assigned to the test set due to its small sample size (**Figure 1, and Supplementary tables 11A and 11B**). We utilized the South African and Gambian training sets to construct site-specific signatures of TB risk, using RNA-seq transcriptomes and the Pair Ratio approach, which uses ratios of transcripts that were regulated in opposite directions during TB progression, as a means to magnify TB-associated signals and simultaneously standardize for RNA concentration by focusing on regulation in opposite directions (**online supplement** and **Supplementary Tables 12 and 13**). Leave-one-out cross-validation analysis (LOOCV; applied to all samples from specific individuals) indicated strong potential for predicting TB progression in both cohorts (South Africa: **Figure 2A**; area under the receiver operating characteristic curve (AUC)=0.86 [95% CI: 0.79-0.94], $p=8.4x10^{-10}$; The Gambia: **Figure 2B;** AUC=0.77 [0.66-0.88]; $p=2.5x10^{-10}$). Applying the algorithm to the South African and Gambian cohorts generated two distinct risk signatures **(Figure 2C and D).** When measured by qRT-PCR using primer/probe sets that corresponded to the exons, predictive accuracy was maintained **(Supplementary Figure 1)**. Surprisingly, the two signatures were not strongly cross-predictive when applied to samples from the other country **(Figures 2A and B).** The South Africa signature weakly validated on Gambian samples

147 (**Figure 2B**; AUC=0.66 [0.54-0.76], $p=8.8X10^{-3}$), while The Gambia signature
148 failed to validate on samples from South Africa (**Figure 2A**; AUC=0.59 [0.46-
149 0.73], *p=0.061*), suggesting site-specific progression signatures in South Africa
150 and The Gambia.

151 The poor cross-prediction of the South Africa and The Gambia signatures
152 motivated explicit development of a multi-cohort signature using a training set
153 that combined samples from both sites. We pooled the PCR-based transcript
154 pairs that comprised all the South Africa (38 transcripts), and The Gambia (35
155 transcripts) signatures (**Figure 2C and D, and Supplementary Tables 12 and**
156 **13**) and sought to identify transcript pairs that were significantly predictive of TB
157 progression in both cohorts. This analysis on RT-PCR data was also carried out
158 using the "Pair Ratios" framework **(online supplement)**. We started by
159 identifying a single pair of transcripts that best fitted the entire training set, and
160 then successively added the next best pair to the ensemble and re-assessed the
161 predictive power at each stage **(Supplementary Table 14)**. This procedure was
162 carried out until addition of pairs led to no further increase in predictive power.
163 This resulted in the RISK4 signature comprising two transcript pairs constructed
164 from four unique genes: GAS6 and SEPT4 were up-regulated, whereas CD1C
165 and BLK were down-regulated in progressors vs. matched controls **(Figure 3A)**.

166 Having developed a multi-site PCR-based signature of risk, we validated it
167 by blind prediction of TB progression on the multi-cohort test sets from South
168 Africa, The Gambia and Ethiopia **(Figure 1)**. The RISK4 signature successfully
169 predicted progression in the entire combined test set (AUC=0.67 [0.57-0.77],

170 *p=2.6X10^{-4}*, **Figure 3B**), and on each individual site (South Africa, The Gambia,
171 and Ethiopia with AUCs: 0.66-0.72, p<0.03, **Figure 3B**). Surprisingly,
172 performance of the signature on combined test set samples within a year of TB
173 diagnosis (AUC=0.66 [0.55-0.78], *p=1.9X10^{-3}*, **Figure 3C**) was comparable to
174 samples collected more than a year before diagnosis (AUCs=0.69 [0.51-0.86],
175 *p=0.015*). Deployment of such a risk signature in a screen-and-treat strategy in
176 TB HHC would most likely entail testing early after exposure. Therefore, we
177 assessed the predictive performance of RISK4 on samples from HHC collected
178 within two months of diagnosis of the index case, and indeed it also validated in
179 this setting (**Figure 3D**; AUC=0.69 [0.52-0.86], *p=4.8X10^{-3}*). Finally, to further
180 corroborate the robustness of RISK4, we performed blinded predictions on
181 samples from an external cohort of IGRA+/TST+ South African adolescents (the
182 "ACS" cohort), where the time of TB exposure was unknown[24]. RISK4 also
183 significantly predicted risk of TB progression in this cohort (**Figure 3E**; AUC=0.69
184 [0.62-0.76], *p=3.4X10^{-7}*).
185
186 **Comparison of RISK4 with published diagnostic TB signatures**
187 To benchmark the predictive performance of the RISK4 signature, we
188 compared it to qRT-PCR-based versions of three published transcriptional
189 signatures for TB diagnosis: "DIAG3"; the 3-gene diagnostic signature by
190 Sweeney et al[29], and "DIAG4"; the 4-gene diagnostic signature by Maertzdorf et
191 al[28], and our own previously-reported 16-gene COR signature for TB progression
192 ("ACS COR", Zak et al[24]). The three signatures predicted TB progression in the

combined test set with comparable accuracy to RISK4 (**Figure 4A**, AUCs of 0.64-0.68, $p<3X10^{-3}$). However, unlike RISK4 (**Figure 3B**), the three other signatures did not validate on all sites when evaluated individually (**Figures 4B-D**), suggesting that RISK4 represents a more generally applicable prognostic signature.

After unblinding the South African, Gambian, and Ethiopian test sets, we interrogated whether the RISK4 signature could be reduced to a single pair of transcripts without a loss of predictive accuracy. We applied each of the four ratios in the RISK4 signature to each of the test set cohorts individually, and compared the performance to the entire RISK4 signature **(Supplementary Table 15)**. The ratio between the SEPT4 and BLK primers reproduced the performance of the RISK4 signature on all three test set cohorts, demonstrating feasibility of a highly simplified, 2-gene host RNA-based signature for identifying HHC at greatest risk of progressing to active TB.

Meta-analysis identifies gene pairs that predict TB progression across Africa

Overall, predictions for TB progression were the least accurate for the Ethiopian cohort, which was not used to develop the initial RISK4 signature **(Figures 1, 3 and 4)**. To determine whether further improved accuracy could be achieved for a signature performing well at all sites, we performed a meta-analysis of RNA-seq profiles for the combined training and test datasets from all our three cohorts. This analysis was focused on identifying better predictive gene

pairs, given that the single transcript pair SEPT4/BLK performed equivalently to the RISK4 signature (**Supplementary Table 15**).

We combined RNA-seq data from all training and test cohorts, thus merging the three independent cohorts from South Africa, The Gambia and Ethiopia. Pairs of up-regulated and down-regulated transcripts were formed from all transcripts that individually discriminated progressors from controls in at least one cohort (**Supplementary Tables 16 and 17;** Wilcoxon FDR<0.05 in at least one of the three cohorts). Each pair was then analyzed on each of the three sites. We identified nine transcript pairs that discriminated progressors from controls with AUC>0.75 on all three sites (**Supplementary Table 18)**. The optimal pair consisted of C1QC (up-regulated) and TRAV27 (down-regulated) and achieved AUC>0.76 on all three sites. We performed logistic regression analysis to determine whether the remaining eight pairs **(Supplementary Table 19, Supplemental Methods)** captured information about TB progression that was redundant or complementary to the signals detected by C1QC/TRAV27. The ratio between ANKRD22 (up-regulated with TB progression) and OSBPL10 (down-regulated with progression) led to significantly increased discrimination between progressors and controls when it was combined with the C1QC/TRAV27 ratio in HHC cohorts **(Figures 5A-C)**, increasing the ROC AUC on all three HHC cohorts individually to AUC>0.79 **(Supplementary Table 20)**. Thus, the ratios C1QC/TRAV27 and ANRKD22/OSBPL10 capture distinct aspects of TB progression signals in HHC that are shared across three distinct African sites.

To determine whether the C1QC/TRAV27 and ANKRD22/OSBPL10 signatures captured universal aspects of TB progression rather than HHC-associated biology, we evaluated them using data from the cohort of IGRA+TST+ South African adolescents[24]. The ANKRD22/OBSPL10 ratio strongly predicted TB progression among the *M.tb*-infected adolescents (**Figure 5D**; AUC=0.75 [0.68-0.81], $p=2.86x10^{-11}$), but the C1QC/TRAV27 ratio was poorly predictive in the adolescent cohort (**Figure 5D**; AUC=0.57 [0.49-0.64], p=0.042). In contrast to the HHC, combining the two ratios did not lead to improved discrimination of progressors and controls in the adolescent cohort (AUC=0.69 [0.61-0.76]; **Figure 5D and Supplementary Figure 2A**). To further understand the disparity in the predictive performance for the HHC cohorts and the *M.tb*-infected adolescents, we evaluated the longitudinal behavior of the transcript ratios for progressor samples in the HHC and adolescent cohorts (**Figures 5F and 5G**). The ANKRD22/OSBPL10 pair exhibited similar behavior in the HHC and ACS, with a steady up-regulation during progression and no significant difference between GC6-74 and adolescent participants in any 6-month time window preceding TB diagnosis **(Figure 5F)**. In contrast, the C1QC/TRAV27 ratio was significantly higher in HHC progressors than in *M.tb*-infected adolescents 19-24 months before TB diagnosis ($p=3X10^{-3}$, **Figure 5G)**. Importantly, samples from HHC progressors were collected mostly at enrolment, immediately following exposure to the respective TB index cases, thus possibly representing a signature of *M.tb* exposure.

Discussion

We identified and validated a simple, easily implementable, PCR-based transcriptomic signature, “RISK4”, to predict risk of progression to active TB disease in diverse African cohorts of recently exposed HHC of index TB cases. This four-gene signature predicted risk of progression with similar accuracy in 4 cohorts from 3 Sub-Saharan African populations with heterogeneous genetic backgrounds, TB epidemiology and circulating *M.tb* strains[30]. Importantly, RISK4 exhibited consistent predictive performance in all test set cohorts, while previously reported signatures[24,28,29] exhibited cohort-specific variability in performance. We previously reported that the ACS COR signature validated on the entire South African and Gambian HHC cohorts, which were not separated into training and test sets[24]. Failure of the ACS COR to predict TB progression on The Gambian test set, as reported here, is likely a function of the sample distribution in the small test set compared with the full Gambian HHC cohort[24].

The signatures reported herein represent significant and translational improvements over currently used biomarkers for predicting risk of TB, such as IGRAs or TST[13,14]. Recent estimates suggest the TB incidence of South Africa and The Gambia to be 0.8%[3] and 0.3%[31], respectively. However, IGRA and TST-positive prevalence can reach up to 50% in The Gambia and 80% in South Africa[3] and although IGRA and TST have a high (approximately 80%) sensitivity for *M.tb* infection, they have poor positive predictive values (PPV) of 2.7% and 1.5%, respectively for TB progression. Therefore, dozens of individuals would require prophylactic treatment to prevent progression to TB in a single

individual[32]. The target product profile for a non-sputum based TB risk test states that it should be a rule-out test with high sensitivity, such that individuals at high risk of TB progression are unlikely to be falsely excluded[7,17] and are referred for additional investigation for TB or offered prophylactic treatment[33]. At sensitivities of 81, 71, 62 and 50% the RISK4 signature achieves specificities of 34, 52, 63 and 77% in healthy asymptomatic individuals, respectively, by selection of different thresholds **(Supplementary Table 21)**. Although RISK4 has a similar poor PPV of 3% as IGRA tests or the TST, it importantly has lower positivity rates in the target population. To achieve a test performance similar to IGRAs (between 70 to 80% sensitivity and the number to harm (NTH) to prevent one case of approximately 85), the RISK4 threshold would identify between 38 and 54% of household contacts for preventative measures, compared to 78% for IGRA **(Supplementary Table 21)**. The performance of RISK4 will, however, have to be confirmed in larger studies. Importantly, RISK4 fulfills the need for a test based on accessible samples, such as blood and could yield rapid results as it does not require antigen stimulation. Computing the score requires basic arithmetic and the pair-ratio structure eliminates the need for housekeepers or other standardization methods. Measurement of the transcript levels can therefore be easily translated to field-friendly PCR devices for simple qRT-PCR-based point-of-care tests.

We identified several transcript pairs that recapitulated the predictive performance of the RISK4 signature and reflected complementary signals in predicting risk of TB progression. The most universal pair defined in this meta-

analysis showed up-regulation of the complement C1q C-chain (C1QC), and down-regulation of T-cell receptor alpha variable gene 27 (TRAV27). Interestingly, complement pathway genes are markedly up-regulated following *M.tb* infection of non-human primates[34], consistent with the up-regulation of C1QC/TRAV27 at baseline in the HHC. Complement activation is also observed early during human progression to TB[35] while C1q is down-regulated early after starting TB treatment[21], suggesting that C1q may be a proxy of early TB pathology. Conversely, down-regulation of TRAV27, and several other T-cell genes **(Supplementary Table 17)**, is likely associated with the overall decrease in peripheral T-cell frequencies and their associated gene expression modules during TB progression, potentially due to migration of T-cells to the disease site[18,20,35]. The simple C1QC/TRAV27 signal may thus be a read-out of TB risk following initial exposure to a pulmonary TB case, which is more synchronized in a HHC study design, even though prior exposure to *M.tb* cannot be ruled out in our GC6-74 study, and progression to TB disease within the first three months of the observation period were excluded from the analysis. This may explain why C1QC/TRAV27 signal was less predictive in the natural history cohort of *M.tb*-infected adolescents, where the time of *M.tb* exposure was unspecified. Early clinical studies suggest that recent exposure to *M.tb*, indicated by TST conversion, can correlate with symptoms consistent with febrile disease, such as fever and erythema nodosum[36,37], markers of systemic inflammation. C1QC/TRAV27 may reflect this inflammatory response induced by failed containment of *M.tb* following recent exposure.

Overall, our study identifies and validates a simple cost-effective PCR-based test from accessible blood samples that predicts TB in heterogeneous African populations with intermediate to high TB burdens[13,14]. The test can be used to screen for risk of progression during TB contact investigation, implemented by national public health structures[12,32]. The next steps include assessment of the performance of RISK4 and the 2-transcript C1QC/TRAV27 signature in other settings, including non-African populations and to determine the feasibility of developing a point-of-care test for targeted intervention.

340 **Table 1:** Baseline demographic characteristics of progressors enrolled and
341 matched non-progressor controls in the 4 African household contact cohorts. n:
342 number of individuals, IQR: interquartile range.

Site	South Africa	The Gambia	Ethiopia	Uganda
HIV- HHC, n	1,197	1,948	818	499
Progressors, n Incidence, %	43 3.6	34 1.7	12 1.5	11 2.2
Median age, years (IQR)				
Progressors	25 (18-41)	22.5 (20-30.75)	23 (19.75-27)	23 (18-36)
Non-progressors	24 (18-38)	24 (18-30.25)	25 (20-35)	27 (19-38.75)
Male, %				
Progressors	41.9	44.1	33.3	54.5
Non-progressors	40.7	44.1	35.4	54.5
Median time to TB, months (IQR)				
Progressors	7 (5-17)	10.5 (7-18.75)	10 (6.5-15)	7 (5-11)

343

344

345 **Figure Legends**

346 **Figure 1: Consort diagram describing the inclusion and exclusion of**
347 **participants from the different African cohorts in the Grand Challenges 6-74**
348 **household contact study:** Stellenbosch University in South Africa (SUN),
349 Armauer Hansen Research Institute in Ethiopia (AHRI), Makerere University in
350 Uganda (MAK), Medical Research Council in The Gambia (MRC), and the
351 external validation natural history study of South African Adolescents (ACS) in
352 training predictive transcriptomic biomarker for TB progression.

353

354 **Figure 2: Site-specific Feature Selection and Translation to RT-PCR. (A)**
355 Receiver Operating Characteristic (ROC) Curve for Leave-One-Out Cross-
356 Validation (LOOCV) of South Africa (blue; AUC=0.86 [0.79-0.94], $p=8.4x10^{-10}$) vs.
357 The Gambia-trained prospective signature (red; AUC=0.59 [95% CI: 0.46-0.73],
358 $p=0.06$) in South African training set; samples listed in Supplementary Tables
359 11A and 11B. **(B)** ROC curves for LOOCV of The Gambia (blue; AUC=0.77
360 [0.66-0.88], $p=2.5x10^{-5}$) vs. South Africa prospective signature (red; AUC=0.66
361 [0.54-0.77], $p=8.8X10^{-3}$) in The Gambia training set containing 26 progressor and
362 76 non-progressor samples. **(C and D)** Heatmaps showing the expression of
363 each splice junction in the South Africa **(C)** and The Gambia **(D)** signatures in
364 non-progressors (left columns), progressors 1-2 years before diagnosis (middle
365 columns), and progressors 0–1 years before diagnosis (right columns). For each
366 group of samples, the central column is the mean fold expression change vs non-
367 progressors, while left/right columns in each group correspond to mean -/+

368 standard error of the mean. Each row corresponds to a splice junction, and
369 genes with multiple rows are represented by multiple splice junctions in the
370 signature.
371
372 **Figure 3: Validation of a multi-cohort 4-gene (RISK4) signature derived from**
373 **the South African and Gambia training sets. (A)** Expression ratio of gene
374 pairs in the RISK4 signature, in South Africa (top) and The Gambia (bottom)
375 training set: non-progressors (left columns), progressors 1–2 years before
376 diagnosis (middle columns), and progressors 0–1 (right columns) years before
377 diagnosis. In each group, the central column is the mean fold expression over
378 non-progressors, while left/right columns in each group correspond to mean -/+
379 standard error of the mean. **(B)** ROC curves for blind predictions of RISK4 on
380 test set samples of all sites (black: AUC=0.67 [0.57-0.77], $p=2.6X10^{-4}$), South
381 Africa (red: AUC=0.72 [0.53-0.92], $p=6.3X10^{-3}$), The Gambia (blue: AUC=0.72
382 [0.55-0.88], $p=5.4X10^{-3}$), and Ethiopia (green: AUC=0.67 [0.5-0.83], $p=0.02$). **(C)**
383 Performance of RISK4 signature in test set samples taken within one year of
384 diagnosis (red; AUC=0.66 [0.55-0.78], $p=1.9X10^{-3}$; 30 progressor samples, 201
385 non-progressor samples) or 1-2 years before diagnosis (blue; AUC=0.69 [0.51-
386 0.86], $p=0.$015; 12 progressor samples, 201 non-progressor samples). **(D)** ROC
387 curve of RISK4 on all baseline test set samples (AUC=0.69 [0.52-0.86],
388 $p=4.8X10^{-3}$). **(E)** ROC curve blind prediction of RISK4 in latently *M.tb*-infected
389 South African adolescents (AUC=0.69 [0.62-0.76], $p=3.4X10^{-7}$).
390

Figure 4: Comparison of RISK4 and published small TB diagnostic signatures. (A) ROC curves for blind predictions of RISK4 (Black: AUC=0.67 [0.57-0.77], $p=2.6X10^{-4}$), DIAG3 (red: AUC=0.68 [0.59-0.78], $p=8.4X10^{-5}$), DIAG4 (blue: AUC=0.64 [0.53-0.74], $p=2.6X10^{-3}$) and ACS COR (green: AUC=0.66 [0.55-0.76], $p=5.8X10^{-4}$) in all test set samples. **(B-D)** Blind prediction of published small signatures: DIAG3 (B: South Africa AUC=0.66 [0.47-0.84], The Gambia AUC=0.6 [0.45-0.77] and Ethiopia AUC=0.78 [0.64-0.92]), DIAG4 (C: South Africa AUC=0.77 [0.62-0.91], The Gambia AUC=0.52 [0.33-0.71] and Ethiopia AUC=0.64 [0.46-0.83]) and RISK16 (D: South Africa AUC=0.82 [0.71-0.92], The Gambia AUC=0.56 [0.37-0.75] and Ethiopia AUC=0.6 [0.41-0.79]). South Africa, The Gambia and Ethiopia AUCs are depicted in red, blue and green, respectively.

Figure 5: Gene pairs to predict TB progression in African cohorts. Ratios of C1QC/TRAV27 and ANKRD22/OBSPL10 plotted on samples from South Africa **(A)**, The Gambia **(B)**, and Ethiopia **(C)** along with an optimal discriminant (dashed line; optimizes sum of sensitivity and specificity) separating progressors (orange) from non-progressors (blue). On each cohort, the two pairs provide complementary information; p-values correspond to Chi-square complementation analysis in Supplementary Table 15. **(D)** ROC curves showing the ability of the GC6-trained C1QC/TRAV27 (solid; AUC=0.57 [0.49-0.64], $p=0.042$), ANKRD22/OBSPL10 (dashed; AUC=0.75 [0.68-0.81], $p=2.86x10^{-11}$), and a linear combination of C1QC/TRAV27 and ANKRD22/OBSPL10 (dotted; AUC=0.69

414 [0.61-0.76], $p=4.3X10^{-07}$) models to predict TB disease progression on in the
415 ACS cohort. **(F and G)** Log-ratios of expression (mean +/- 95% confidence
416 interval) for ANKRD22/OBSPL10 **(F)** and C1QC/TRAV27 **(G)** are plotted as a
417 function of time to diagnosis, for both GC6 (blue) and ACS (red) progressor
418 samples. Comparison of C1QC/TRAV27 expression at 19-24 months before
419 diagnosis, between the GC6-74 HHC and ACS cohorts was statistically
420 significantly different ($p=3X10^{-3}$) using the Mann-Whitney *U* test.
421
422

References:

1. Riley, F. Tuberculosis in hospital nurses; five years figures. *Mon Bull Minist Health Public Health Lab Serv* **18**, 38-45 (1959).
2. Yates, T.A.*, et al.* The transmission of Mycobacterium tuberculosis in high burden settings. *Lancet Infect Dis* **16**, 227-238 (2016).
3. WHO. Global Tuberculosis Report 2017. *published online:* http://www.who.int/tb/publications/global_report/gtbr2017_main_text.pdf?ua=1 (2017).
4. Rieder, H. Annual risk of infection with Mycobacterium tuberculosis. *Eur Respir J* **25**, 181-185 (2005).
5. O'Garra, A.*, et al.* The immune response in tuberculosis. *Annu Rev Immunol* **31**, 475-527 (2013).
6. Andrews, J.R.*, et al.* Risk of progression to active tuberculosis following reinfection with Mycobacterium tuberculosis. *Clin Infect Dis* **54**, 784-791 (2012).
7. Cobelens, F.*, et al.* From latent to patent: rethinking prediction of tuberculosis. *Lancet Respir Med* (2016).
8. Sester, M., van Crevel, R., Leth, F. & Lange, C. Numbers needed to treat to prevent tuberculosis. *Eur Respir J* **46**, 1836-1838 (2015).
9. Wood, R.*, et al.* Burden of new and recurrent tuberculosis in a major South African city stratified by age and HIV-status. *PLoS One* **6**, e25098 (2011).
10. Blaser, N.*, et al.* Tuberculosis in Cape Town: An age-structured transmission model. *Epidemics* **14**, 54-61 (2016).
11. 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).
12. 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).
13. Abu-Raddad, L.J.*, et al.* Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. *Proc Natl Acad Sci U S A* **106**, 13980-13985 (2009).
14. Dye, C., Glaziou, P., Floyd, K. & Raviglione, M. Prospects for tuberculosis elimination. *Annu Rev Public Health* **34**, 271-286 (2013).
15. Coussens, A.K.*, et al.* Ethnic variation in inflammatory profile in tuberculosis. *PLoS Pathog* **9**, e1003468 (2013).
16. Comas, I.*, et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **45**, 1176-1182 (2013).
17. Denkinger, C.M.*, et al.* Defining the needs for next generation assays for tuberculosis. *J Infect Dis* **211 Suppl 2**, S29-38 (2015).

18. Berry, M.P., *et al.* An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**, 973-977 (2010).
19. Maertzdorf, J., *et al.* Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes Immun* **12**, 15-22 (2011).
20. Joosten, S.A., Fletcher, H.A. & Ottenhoff, T.H. A helicopter perspective on TB biomarkers: pathway and process based analysis of gene expression data provides new insight into TB pathogenesis. *PLoS One* **8**, e73230 (2013).
21. Cliff, J.M., *et al.* Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *J Infect Dis* **207**, 18-29 (2013).
22. Cliff, J.M., Kaufmann, S.H., McShane, H., van Helden, P. & O'Garra, A. The human immune response to tuberculosis and its treatment: a view from the blood. *Immunol Rev* **264**, 88-102 (2015).
23. Ottenhoff, T.H., *et al.* Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. *PLoS One* **7**, e45839 (2012).
24. Zak, D.E., *et al.* A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *Lancet* **387**, 2312-2322 (2016).
25. Mahomed, H., *et al.* Predictive factors for latent tuberculosis infection among adolescents in a high-burden area in South Africa. *Int J Tuberc Lung Dis* **15**, 331-336 (2011).
26. Saeed, S., *et al.* Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* **345**, 1251086 (2014).
27. Thompson, E.G., *et al.* Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis* **107**, 48-58 (2017).
28. Maertzdorf, J., *et al.* Concise gene signature for point-of-care classification of tuberculosis. *EMBO Mol Med* **8**, 86-95 (2016).
29. 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).
30. de Jong, B.C., *et al.* Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. *J Infect Dis* **198**, 1037-1043 (2008).
31. Adetifa, I.M., *et al.* A tuberculosis nationwide prevalence survey in Gambia, 2012. *Bull World Health Organ* **94**, 433-441 (2016).
32. Petruccioli, E., *et al.* Correlates of tuberculosis risk: predictive biomarkers for progression to active tuberculosis. *Eur Respir J* **48**, 1751-1763 (2016).
33. Penn-Nicholson, A., Scriba, T.J., Hatherill, M., White, R.G. & Sumner, T. A novel blood test for tuberculosis prevention and treatment. *S Afr Med J* **107**, 4-5 (2016).
34. 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).

35. Scriba, T.J., Penn-Nicholson, A., Shankar, S., Hraha, T., Thompson, E.G., Sterling, D., Nemes, E., Darboe, F., Suliman, S., Amon, L.M., Mahomed, H., Erasmus, M., Whatney, W., Johnson, J.L., Boom, W.H., Hatherill, M., Valvo, J., De Groote, M.A., Ochsner, U.A., Aderem, A., Hanekom, W.A., Zak, D.E., and other members of the ACS cohort study team. Sequential inflammatory processes define human progression from M. tuberculosis infection to tuberculosis disease. *PLoS Pathogens* **13**(11): e1006687 (2017).
36. Wallgren, A. The time-table of tuberculosis. *Tubercle* **29**, 245-251 (1948).
37. Poulsen, A. Some clinical features of tuberculosis. 1. Incubation period. *Acta Tuberc Scand* **24**, 311-346 (1950).