SUPPLEMENTARY METHODS:

Study Inclusion:

We hypothesized that the gene expression signature in response to TB included multiple molecular sub-groups, or endotypes. We therefore implemented a systemic review and individual participant data meta-analysis according to the PRISMA guidelines[1]. A priori hypotheses were established and shared with co-authors prior to implementation. The prospective analysis plan was to evaluate if molecular sub-groups, or endotypes, exist within a common TB clinical phenotype, with the gene expression signature of TB endotypes compared to that of healthy controls. A search of publicly available gene expression datasets of patients with TB was carried out using Pubmed, and the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository. The search was last performed on 10 January 2021, including the terms “tuberculosis,” “transcriptomics,” “microarray,” and “diagnostics.” We included studies assessing the gene expression signature of participants with microbiologically confirmed pulmonary TB and healthy, asymptomatic controls. Microbiologic tests of TB included either smear microscopy, culture, or Mtb-specific PCR such as Gene Xpert. Participants without microbiological confirmation were not included. We excluded studies that did not describe methods of microbiologic confirmation or evaluated cells other than whole blood (i.e. PBMC-derived gene expression or in vitro cell culture experiments). Tuberculin skin test (TST) and/or IFN-γ release assay (IGRA) results were not available in all studies; hence, our analytic control group included asymptomatic individuals with and without latent infection with Mtb.

| Deposited Data | **Full description of the cohorts evaluated can be found in supplemental Table S1** |
|----------------|----------------------------------------------------------------------------------|
| Microarray Discovery Cohort | GSE39939                                                                          |
| Microarray Discovery Cohort | GSE19491 superseries: GSE19435, GSE19439, GSE19442, GSE19444                     |
| Microarray Discovery Cohort | GSE83456                                                                           |
| Microarray Discovery Cohort | GSE40553                                                                           |
| Microarray Discovery Cohort | GSE42834 superseries: GSE42825, GSE42826, GSE42830                                 |
| Microarray Discovery Cohort | GSE37250                                                                           |
| Microarray Discovery Cohort | GSE73408                                                                           |
| German and Romanian Validation cohort | GSE147689, GSE147691                                                              |
| RNA-seq Validation Cohorts | GSE101705 GSE107995 superseries: GSE107991, GSE107992, GSE107994                  |

Data normalization and preprocessing:

Studies included both microarray[2-8] and genome-wide (RNA-seq)[9, 10] gene expression analyses (Supplemental Table 1). Transcriptomic profiles for the discovery cohort (assessed using microarray platforms) included 12,468 commonly evaluated genes. Gene expression datasets were normalized using the ComBat R package[11]. Transcriptomic profiles for the validation cohort (assessed using microarray platforms) included 21,754 commonly
evaluated genes and were normalized using upper quartile normalization as previously processed by the Borstel group[12]. RNA-Seq for the validation cohort, consisting of 297 samples (118 TB and 179 healthy controls) was mapped to the human genome UCSC hg38 using STAR[13]. Gene expression was quantified using featureCounts and the GENCODE gene model. RNA-Seq data was normalized using the EdgeR R package. The feature counts were normalized using upper quartile normalization and the normalized RNA-Seq data for validation was represented as log2(CPM+1), where CPM represents counts per million reads mapped.

**TB Clustering:**
To identify independent TB endotypes, we used recent transcriptomic clustering methods implemented in the single-cell RNA-Seq analysis packages Seurat[14]. Specifically, we first determined the top 50 Principal Component Analysis (PCA) components, then employed cell clustering using the findCluster method, that implements the Louvain network-based clustering algorithm[15]. To better assess the quality of the TB patient clusters detected, we employed findCluster using a wide range of resolutions, from R=0 to R=1.2. The tree of clusters at different resolutions was generated and visualized using the R package clustree (Supplemental Figure S1) [16].

**Pathway and Connectivity Map analysis:**
The gene expression signature of each TB patient cluster was compared to the group of 533 healthy controls. We used the limma package[17], and considered significance achieved for an FDR-adjusted p-value<0.05 and fold change < 0.5 or > 2. We carried out pathway enrichment using the Gene Set Enrichment Analysis Method (GSEA)[18] using rank files of the 12,468 genes, 1000 permutations, with significance achieved at FDR<0.25. GSEA was carried out against the Hallmark pathway compendium[19], as hosted by the Molecular Signature Database (MSigDB) resource[20]. Chemical compounds associated with the gene signatures were determined using the Library of Integrated Network-Based Cellular Signatures (LINCS) resource[21]. Based on analysis techniques routinely used in cancer cohorts[22], we determined pathway activity scores for each TB patient and each of the Hallmark pathways. First, the gene expression data was converted via a z-score transformation (achieving a mean of zero and standard deviation of 1 for each gene); next the z-scores for each gene in the pathway were added at individual patient level, thus generating a pathway activity score for each patient transcriptome profile. Differences between overall pathway scores in healthy controls and the two major TB endotypes were then assessed using the Mann-Whitney-Wilcoxon test, with significance achieved at p<0.05.

**Gene classifier:**
Given the fundamental difference between the microarray and RNA-seq profiling platforms, we independently normalized the microarray data and the RNA-seq data and used the former as the discovery cohort and the latter as the validation cohort. The first validation cohort consisted of 297 RNA-seq samples, including 118 microbiologically confirmed TB cases and 179 healthy controls (Supplemental Table S1). The second validation cohort consisted of 135 microarray samples, including 121 microbiologically confirmed TB cases and 14 healthy controls (Supplemental Table S1). Gene expression of the TB patients was normalized with respect to healthy controls using z-score normalization within each of the discovery and validation datasets, respectively. We further selected 11,568 genes commonly detected in the discovery and each of the two validation datasets. We applied the random forest classifier in the discovery cohort and determined the importance of each gene in the random forest model as measured by its mean decrease in Gini index. We next determined the top 500 genes in order
of importance. Since low cost would be beneficial for a classifier deployed in different economical and field condition, a search for an effective endotype A versus endotype B classifier was performed using different gene counts. Specifically, after selecting and sorting the top 500 genes in order of their random forest importance in the discovery cohort, classifiers were built and evaluated using a range of genes, as indicated in Supplemental Figure 2A. Classification performance was measured in the discovery cohort via the classification error rate, implemented in RandomForest as out of bag (OOB) error via repeated subsampling with replacement. The top three performance scores were achieved for 500 genes, 40 genes, and 50 genes, respectively. We next applied the 40 genes, 50 genes, and 500 genes random forest classifiers to the two validation datasets and computed the GSEA pathway enrichment using the Hallmark compendium. As indicated in Supplemental Figure 2B, very similar results were obtained in each validation dataset across 40, 50, or 500 genes. Given the significantly lower cost of measuring 40 genes versus 500 genes, and the minor difference in classification performance, measured either as OOB error rate or via enriched Hallmark pathways via GSEA, we selected 40 genes as the basis of our final random forest classifier.

**TB transcriptome profiles trajectory analysis**

Transcriptome profile trajectory analysis (pseudotime) was carried out over the 968 TB patients and healthy controls from the discovery dataset by utilizing the single-cell RNA-Seq analysis package Monocle over the normalized microarray data[23].

**Disease severity score:**

The endotypes were evaluated against six previously published scores that identified risk of treatment failure, with two of the scores also identifying risk for TB disease severity[24-29]. We determined signature activity scores for each TB patient and each of gene signatures. First, the gene expression data were converted via a z-score transformation (achieving a mean of zero and standard deviation of 1 for each gene); next the z-scores for each up signature gene were added at individual patient level, and the z-scores for each down signature gene were subtracted at individual patient level, thus generating a signature activity score for each patient transcriptome profile. Differences between overall gene signature activity scores in the healthy controls and the two major TB endotypes were then assessed using the Mann-Whitney-Wilcoxon test, with significance achieved at p<0.05.

**Statistics:**

A one-sided Chi-squared test (assuming Endotype A outcome is worse than Endotype B, based on findings in Figure 3) was used to assess the incidence of various clinical variables over the TB patient endotypes groups, with significance achieved at p<0.05.

Patient profiles for the ELISA validation cohort were sorted by the rank-sum of the cytokine/chemokine levels that were also present in the TB endotype A/TB endotype B gene signature. They were further stratified in the top 50% of the patients (TB Top) and the bottom 50% of the patients (TB Bottom). Differences between all the profiled proteins were assessed using the Mann-Whitney-Wilcoxon rank sum test, with significance achieved for p-value<0.05.

**Software and Algorithms**

| ComBat     | (Johnson et al., 2007) | https://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html |
External Multiplex ELISA Validation Cohort:

To validate the gene expression data, a cohort of TB patients and asymptomatic household contacts from Eswatini was evaluated (n = 79). The study protocol was reviewed by the Baylor College of Medicine Children’s Foundation-Swaziland, Baylor College of Medicine and Eswatini National Health Research and Review Board. All participation was voluntary and implemented in accord with the institutional and international guidelines for the Protection of Human Subjects in concordance with the Declaration of Helsinki including written informed consent from participants or their guardians. TB patients were defined by symptoms consistent with TB (chronic cough, fever, night sweats, anorexia) and microbiologic confirmation by culture or PCR (Gene Xpert). All TB participants in this analysis were microbiologically positive by culture or PCR. Controls were asymptotic household contacts of TB patients and remained asymptomatic for 12 months after enrollment. Forty-eight percent of the controls had a positive Quantiferon test (Qiagen, Hilden, Germany).

| Reagents          | Source          | Catalog   |
|-------------------|-----------------|-----------|
| LegendPlex IFNg   | BioLegend       | 740352    |
Further information and requests for reagents may be directed to, and will be fulfilled by the senior co-authors Cristian Coarfa (coarfa@bcm.edu), Anna Mandalakas (Anna.Mandalakas@bcm.edu), or Andrew DiNardo (Andrew.Dinardo@bcm.edu).

REFERENCES

1. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, Tierney JF, Group P-ID. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015: 313(16): 1657-1665.

2. Berry MP, Graham CM, McNab FW, Xu Z, Bloch SA, Oni T, Wilkinson KA, Banchereau R, Skinner J, Wilkinson RJ, Quinn C, Blankenship D, Dhawan R, Cush JJ, Mejias A, Ramilo O, Kon OM, Pascual V, Banchereau J, Chaussabel D, O’Garra A. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 2010: 466(7309): 973-977.

3. Bloom CI, Graham CM, Berry MP, Wilkinson KA, Oni T, Rozakeas F, Xu Z, Rossello-Urgell J, Chaussabel D, Banchereau J, Pascual V, Lipman M, Wilkinson RJ, O’Garra A. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. *PloS one* 2012: 7(10): e46191.

4. Bloom CI, Graham CM, Berry MP, Rozakeas F, Redford PS, Wang Y, Xu Z, Wilkinson KA, Wilkinson RJ, Kendrick Y, Devouassoux G, Ferry T, Miyara M, Bouvry D, Valeyre D, Gorochov G, Blankenship D, Saadatian M, Vanhems P, Beynon H, Vancheeswaran R, Wickremasinghe M, Chaussabel D, Banchereau J, Pascual V, Ho LP, Lipman M, O’Garra A. Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PloS one* 2013: 8(8): e70630.

5. Walter ND, Miller MA, Vasquez J, Weiner M, Chapman A, Engle M, Higgins M, Quinones AM, Rosselli V, Canono E, Yoon C, Cattamanchi A, Davis JL, Phang T, Stearman RS, Datta G, Garcia BJ, Daley CL, Strong M, Kecharis K, Fingerlin TE, Reves R, Geraci MW. Blood Transcriptional Biomarkers for Active Tuberculosis among Patients in the United States: a Case-Control Study with Systematic Cross-Classifier Evaluation. *Journal of clinical microbiology* 2016: 54(2): 274-282.

6. Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluha G, Crampin AC, Dockrell HM, French N, Hamilton MS, Hibberd ML, Kern F, Langford PR, Ling L, Motha R, Ottenhoff TH, Pienaar S, Pillay V, Scott JA, Twahir H, Wilkinson RJ, Coin LJ, Heyderman RS, Levin M, Eley B, Consortium I, Group KTS. Diagnosis of childhood tuberculosis and host RNA expression in Africa. *The New England journal of medicine* 2014: 370(18): 1712-1723.

7. Kaforou M, Wright VJ, Oni T, French N, Anderson ST, Bangani N, Banwell CM, Brent AJ, Crampin AC, Dockrell HM, Eley B, Heyderman RS, Hibberd ML, Kern F, Langford PR, Ling L, Mendelson M, Ottenhoff TH, Zgambo F, Wilkinson RJ, Coin LJ, Levin M. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS medicine* 2013: 10(10): e1001538.
8. Blankley S, Graham CM, Turner J, Berry MP, Bloom CI, Xu Z, Pascual V, Banchereau J, Chaussabel D, Breen R, Santis G, Blankenship DM, Lipman M, O’Garra A. The Transcriptional Signature of Active Tuberculosis Reflects Symptom Status in Extra-Pulmonary and Pulmonary Tuberculosis. *PloS One* 2016: 11(10): e0162220.
9. Singhania A, Verma R, Graham CM, Lee J, Tran T, Richardson M, Lecline P, Leissner P, Berry MPR, Wilkinson RJ, Kaiser K, Rodriguez M, Wolfram G, Haldar P, O’Garra A. A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection. *Nature Communications* 2018: 9(1): 2308.
10. Leong S, Zhao Y, Joseph NM, Hochberg NS, Sarkar S, Pleskunas J, Hom D, Lakshminarayan S, Horsburgh CR, Jr., Roy G, Ellner JJ, Johnson WE, Salgame P. Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India. *Tuberculosis* 2018: 109: 41-51.
11. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007: 8(1): 118-127.
12. Heyckendorf J, Marwitz S, Reimann M, Avsar K, Dinardo A, Gunther G, Hoelscher M, Ibraim E, Kalsdorf B, Kaufmann SHE, Kontsevaya I, van Leth F, Mandalakas AM, Maurer FP, Muller M, Nitschowsky D, Olaru ID, Popa C, Rachow A, Rolling T, Rybniker J, Salzer HJF, Sanchez-Carballo P, Schuermann M, Schau B, Spinu V, Suarez I, Terhalle E, Unnewehr M, Weiner J, 3rd, Goldmann T, Lange C. Prediction of anti-tuberculosis treatment duration based on a 22-gene transcriptomic model. *The European respiratory journal* 2021.
13. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013: 29(1): 15-21.
14. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015: 33(5): 495-502.
15. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019: 9(1): 5233.
16. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* 2018: 7(7).
17. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004: 3: Article3.
18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005: 102(43): 15545-15550.
19. Liberzon A, Birger C,Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 2015: 1(6): 417-425.
20. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011: 27(12): 1739-1740.
21. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM, Piccioni F, Johnson SA, Lyons NJ, Berger AH, Shamji AF, Brooks AN, Vrcic A, Flynn C, Rosains J, Takeda DY, Hu R, Davison D, Lamb J, Ardlie K, Hogstrom L, Greenside P, Gray NS, Clemons PA, Silver S, Wu X, Zhao WN, Read-Button W, Wu X, Hagarty SJ, Ronco LV, Boehm JS, Schreiber SL, Doench JG, Bittker JA, Root DE, Wong B, Golub TR. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 2017: 171(6): 1437-1452 e1417.
22. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, Socci
23. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 2017: 14(10): 979-982.

24. Penn-Nicholson A, Mbandi SK, Thompson E, Mendelsohn SC, Suliman S, Chegou NN, Malherbe ST, Darboe F, Erasmus M, Hanekom WA, Bilek N, Fisher M, Kaufmann SHE, Winter J, Murphy M, Wood R, Morrow C, Van Rhijn I, Moody B, Murray M, Andrade BB, Sterling TR, Sutherland J, Naidoo K, Padayatchi N, Walzl G, Hatherill M, Zak D, Scriba TJ, Adolescent Cohort Study t, Consortium GC, Clinical S, Laboratory T, Screen TBC, Consortium A-T, Re PBT, Peruvian Household Contacts Cohort T, team Cl. RISK6, a 6-gene transcriptomic signature of TB disease risk, diagnosis and treatment response. *Sci Rep* 2020: 10(1): 8629.

25. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *The Lancet Respiratory medicine* 2016: 4(3): 213-224.

26. Ronacher K, Chegou NN, Kleynhans L, Djoba Siawaya JF, du Plessis N, Loxton AG, Maasdorp E, Tromp G, Kidd M, Stanley K, Kriel M, Menezes A, Gutschmidt A, van der Spuy GD, Warren RM, Dietze R, Okwera A, Thiel B, Belisle JT, Cliff JM, Boom WH, Johnson JL, van Helden PD, Dockrell HM, Walzl G. Distinct serum biosignatures are associated with different tuberculosis treatment outcomes. *Tuberculosis* 2019: 118: 101859.

27. Sivakumaran D, Jenum S, Vaz M, Selvam S, Ottenhoff THM, Haks MC, Malherbe ST, Doherty TM, Ritz C, Grewal HMS. Combining host-derived biomarkers with patient characteristics improves signature performance in predicting tuberculosis treatment outcomes. *Commun Biol* 2020: 3(1): 359.

28. Thompson EG, Du Y, Malherbe ST, Shankar S, Braun J, Valvo J, Ronacher K, Tromp G, Tabb DL, Alland D, Shenai S, Via LE, Warwick J, Aderem A, Scriba TJ, Winter J, Walzl G, Zak DE, Catalysis TBBC. Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis* 2017: 107: 48-58.

29. Kumar NP, Moideen K, Nancy A, Viswanathan V, Thiruvengadam K, Nair D, Banurekha VV, Sivakumar S, Hissar S, Kornfeld H, Babu S. Plasma chemokines are baseline predictors of unfavorable treatment outcomes in pulmonary tuberculosis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2020.