Supplementary Figures

Supplementary Figure 1: Details of the tuberculosis risk signature. (A) Network representation of the tuberculosis risk signature. Nodes (circles) represent splice junctions from genes that comprise the signature. Node colors indicate to which genes the splice junctions belong. Lines indicate splice junction discriminant pairs that make up the overall signature. Node size is proportional to the number of discriminant pairs that contain the splice junction. All splice junctions were expressed at higher levels in progressors than controls. (B) Network representation of the tuberculosis risk scores during the transition from latent to active tuberculosis for a given progressor. Pair-wise discriminants that vote "progressor" are indicated as red lines and those that vote "control" are indicated as green lines. The overall tuberculosis risk score is the proportion of all pair-wise discriminants that vote progressor. The risk scores, expressed as percentages, are shown for progressor samples at the indicated time points before disease diagnosis, or from a sample from a single matched healthy control at study day 0.

Supplementary Figure 2: Benchmarking the predictive performance of the SVM-based tuberculosis risk signature against a risk signature derived from use of Random Forests. To benchmark prediction performance of the risk signature against an alternative data mining method, a Random Forest model consisting of 100,000 trees (100k RF) was generated from ACS Training set RNA-Seq data. (A-D) Performance of the Random Forest model was compared to the PSVM-based tuberculosis risk signature in cross-validation of the ACS Training set, stratified by time preceding tuberculosis disease onset: (A) 1-180 days before tuberculosis, (B) 181-360 days before tuberculosis, (C) 361-540 days before tuberculosis, and (D) 541-720 days before tuberculosis. (E) Despite the superiority of 100k RF in cross-validation, the prediction accuracy on the ACS test set was indistinguishable for the two methods.

Supplementary Figure 3: Performance of the tuberculosis risk signature for diagnosing active adult and childhood tuberculosis disease using published microarray datasets. (A) Analytical strategy for evaluating the tuberculosis risk signature as a signature for active tuberculosis disease using published microarray datasets. The RNA-Seq-based signature was re-parameterized using whole blood microarray data from the UK training set of Berry, *et al.*,[11]

651 which includes samples from tuberculosis cases and latently *M. tuberculosis* infected controls.
652 The fully locked-down microarray-based signature was then employed to make quasi-blind
653 predictions on data from other cohorts from the same and independent studies.[9-13] (B-H) ROC
654 curves depicting prediction performance of the Illumina microarray-based risk signature on
655 cohorts from published studies are shown. (B) Discrimination of active tuberculosis disease from
656 latent *M. tuberculosis* infection and healthy controls in the South African test set and UK test set
657 from Berry, *et al.*[11]; (C) Discrimination of active tuberculosis disease from latent *M. tuberculosis*
658 infection in the presence or absence of HIV co-infection, using data from Kaforou, *et al.*[12]; (D)
659 Discrimination of active tuberculosis disease from lung cancer, pneumonia, and sarcoidosis
660 using data from Bloom, *et al.*[10]; (E) Discrimination of active tuberculosis disease from other
661 pulmonary diseases in the presence or absence of HIV co-infection, using data from Kaforou, *et*
662 *al.*[12]; (F) Discrimination of culture positive and culture negative childhood tuberculosis disease
663 from latent *M. tuberculosis* infection, using data from Anderson, *et al.*[13]; (G) Discrimination of
664 culture positive childhood tuberculosis disease from other diseases in presence or absence of
665 HIV co-infection, using data from Anderson, *et al.*[13]; (H) Discrimination of active tuberculosis
666 disease from healthy individuals over the course of antimicrobial treatment (red lines), and
667 discrimination of baseline tuberculosis disease from tuberculosis disease after 2 weeks of
668 treatment (purple line), using data from Bloom, *et al*.[9]

Supplementary Appendices

1. Supplementary Appendix 1: Exclusion criteria for the adolescent cohort study (ACS)
2. Supplementary Appendix 2: Exclusion criteria for the GC6-74 study
3. Supplementary Appendix 3: Regulatory authorities that approved the ACS and GC6-74
4. Supplementary Appendix 4: Supplementary Methods

Supplementary Appendix 1: Exclusion criteria for the adolescent cohort study

Parent study:

- Pregnant or lactating women
- Any reported acute or chronic medical condition resulting in hospitalization within 6 months prior to enrollment.

Case-control study:

- Refer to main text

Supplementary Appendix 2: Exclusion criteria for the GC6-74 study

Parent study:

- Resident in the study area for less than 3 months prior to enrollment
- No permanent address
- Previous treatment for tuberculosis
- Previous or current anti-retroviral therapy
- Current participation in a drug or vaccine trial, or participation within 6 months of enrollment
- Concomitant cancer
- Diabetes mellitus
- Chronic bronchitis/emphysema/asthma requiring systemic steroid therapy
- Other steroid therapy within 6 months of enrollment
- Current pregnancy, or pregnancy within 3 months of enrollment

Case-control study:

- Refer to main text

Supplementary Appendix 3: Regulatory authorities that approved the ACS and GC6-74 protocols

ACS:

- University of Cape Town Research Ethics Committee, Cape Town, South Africa

GC6-74 South African and Gambian sites:

- Scientific and Ethics Committee of the Faculty of Health Sciences of Stellenbosch University, Cape Town, South Africa
- Joint Medical Research Council and Gambian Government ethics review committee, Banjul, The Gambia

Supplementary Appendix 4: Supplementary Methods

A. Detailed definition of cases and controls in the ACS for identifying and validating signatures of tuberculosis risk
B. Sequencing of whole blood transcriptomes
C. Derivation of the tuberculosis risk signature from the ACS training RNA-Seq dataset
D. Adaptation of the tuberculosis risk signature from RNA-Seq to qRT-PCR
E. Adaptation of the tuberculosis risk signature to the Illumina microarray platform

A. Detailed definition of cases and controls in the ACS and in GC6-74 for identifying and validating signatures of tuberculosis risk

The ACS determined the prevalence and incidence of *M. tuberculosis* infection and disease among adolescents from the Cape Town region of South Africa.[30,37] Overall, 53% of ACS participants had latent *M. tuberculosis* infection at enrollment. For the ACS signature of risk study, adolescents with latent *M. tuberculosis* infection at enrollment were eligible; *M. tuberculosis* infection was diagnosed by a positive QuantiFERON TB GOLD In-Tube Assay (QFT, Cellestis; >0·35 IU/mL) and/or a positive tuberculin skin test (TST, 0·1mL dose of Purified Protein Derivative RT-23, 2-TU, Staten Serum Institute; >10mm). According to South African policy, these QFT and/or TST positive adolescents were not given therapy to prevent tuberculosis disease.[21]

Adolescents who developed active tuberculosis disease during 2 years of follow-up were included as "progressors" (cases). Participants that were either exposed to tuberculosis patients, or had symptoms suggestive of tuberculosis, were evaluated clinically and by sputum smear, culture and chest roentgenography. 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). Participants who developed tuberculosis within 6 months of enrolment were excluded on the basis that they may represent individuals with active but as yet asymptomatic tuberculosis disease.

Five ACS participants who were not infected with *M. tuberculosis* at enrollment but who converted to a positive QFT and/or TST, and ultimately developed tuberculosis disease at least 6 months post QFT/TST conversion, were also included as progressors. As a subset of ACS participants from the parent study, who had a negative QFT at baseline, were followed for incident tuberculosis for up to 3 years after the last QFT (5 years in total) through biannual study visits and passive surveillance of health facility records, the follow-up of these participants was longer than the 2 years applying to most participants [38].

All ACS patients with tuberculosis disease were offered a HIV test; HIV infected patientswere excluded from the case controls study. HIV testing of healthy study participants was not permitted by the human research ethics committee of the University of Cape Town; this committee also did not allow post-hoc, anonymous HIV testing. Regardless, the HIV incidence rate in adolescents diagnosed with active tuberculosis was <2% (1 out of 61 who were offered and accepted testing), and since HIV is a risk factor for tuberculosis, we expect the HIV prevalence among healthy adolescents (from whom controls were identified) to be minute[30].

For each ACS progressor, two matched controls were identified. Controls were selected from ACS participants that remained healthy during follow-up, and were matched to progressors by age at enrolment, gender, ethnicity, school of attendance, and presence or absence of prior episodes of tuberculosis disease.

Among GC6-74 participants, progressors were defined as having intrathoracic tuberculosis by one of three categories: First, two positive sputum cultures (MGIT); second, one positive sputum culture and/or a positive sputum smear, and clinical signs and symptoms compatible with tuberculosis and/or a chest roentgenogram compatible with active pulmonary tuberculosis; third, two positive sputum smears with clinical signs and symptoms compatible with tuberculosis or a chest roentgenogram compatible with active pulmonary tuberculosis. Progressors were excluded if they developed disease within 90 days of enrollment, for reasons mentioned above. Controls were matched to progressors based on age category (<18, 18-25, 26-35, ≥36 years of age), gender and year of enrollment.

Time to tuberculosis: For the ACS study, two *time to tuberculosis* values were calculated for each progressor. First, original values were assigned early after sample collection and were employed throughout signature construction. Second, per protocol values were assigned during manuscript preparation when it was revealed that some original *time to diagnosis* assignments had been wrong. All prediction results are reported in terms of the per protocol values.

B. Sequencing of whole blood transcriptomes in the ACS

RNA-Seq: Whole blood was collected in PAXgene Blood RNA Tubes (PreAnalytiX), frozen, and RNA was later extracted using of PAXgene Blood RNA Kits. Globin transcripts were depleted and RNA was sequenced by Expression Analysis Inc. (Durham, NC) using a 30M read, 50bp paired-end sequencing strategy (60M reads/sample).

Alignment: Read pairs were preprocessed to adjust base calls with phred scores <5 to N and to remove read pairs for which either end has fewer than 30 unambiguous (non-N) base calls. This method indirectly removes pairs containing adaptor sequences. The median depth of RNA sequencing after post-processing was 31 million read pairs. Read pairs were aligned to the human genome using the gsnap program[22], allowing for novel splice junction detection. The mean percentage of reads mapped was 90% and the mean GC content was 50·6%.

Splice junction filtering: Gene expression abundance was measured at the level of splice junction counts, which quantifies the relative frequency of specific mRNA splicing events in expressed genes. This facilitates adaptation of the signatures from RNA-Seq to qRT-PCR because, in practice, PCR primer sets are designed to target splice junctions as a means to prevent amplification of contaminating genomic DNA. Splice junctions that were detected by at least five reads in at least ten samples were retained for signature analysis, leaving 141,140 splice junctions. In total, 355 samples from the ACS cohort were analyzed by RNA-Seq (Supplementary Table 6): 264 samples from the ACS training set and 91 samples from the ACS test set. RNA-Seq analysis of the ACS training and test sets was performed independently, over a year apart.

Normalization: Splice junction counts for each sample were first pre-normalized for library size by adding "1" to the raw counts, dividing the counts in a given sample by the sum of all counts in that sample, and then taking the logarithm (base 2). "Reference junctions" for use as internal controls in all subsequent analyses were then identified from the 20 splice junctions with the smallest coefficient of variance computed across all samples from the pre-normalized table. The final normalized log2-based splice junction table was finally constructed by subtracting the mean of the reference junction counts for each sample. Reference junctions were identified by using the 264 samples that comprise the full ACS training set RNA-Seq sample set (Supplementary Table 6), which included a small number of samples that were collected after the initiation of treatment (Supplementary Table 2). The set of reference junctions is provided in Supplementary Table 11. Expression levels of the reference junctions and junctions comprising the tuberculosis risk signature are provided in Supplementary Table 10. The Microsoft Excel-based worksheet for computing tuberculosis risk scores (Supplementary Table 11) includes reference junction normalization.

C. Derivation of the tuberculosis risk signature from the ACS training RNA-Seq dataset

Mining the RNA-Seq data from the ACS training set to generate the tuberculosis risk signature required (1) selecting the appropriate **mathematical framework** in which the signature would be formulated, and (2) exploiting the **longitudinal structure** of the progression cohort to extract and train the signature.

Mathematical framework – SVM: The mathematical framework for the signatures is a generalization of the k-top-scoring pairs (k-TSP) methodology, was developed for discovery of cancer biomarkers from microarray datasets. [23] Signatures derived using the k-TSP approach are collections of gene-pair discriminators that can vote "progressor" (1) or "control" (0) (for example). For a given sample, the classification "score" is the average of all of the "0" or "1" votes computed for the whole collection of discriminators for that sample. In this manner, k-TSP combines many "weak" discriminators to improve the reliability of the predictions. The pair-

wise discriminators underlying k-TSP are very simple, involving only a pair of genes for which gene1 > gene2 in progressors and the reverse is true in controls (for example).

The k-TSP framework is desirable in the present study for three reasons. First, it has the potential to identify combinations of genes that better predict progression than either gene individually, a characteristic common to bivariate approaches.[24] Second, being based on an ensemble of models, rather than a single model, the methodology is tolerant to failed measurements. For example, if a particular primer fails for a particular sample, the overall score can still be computed from the unaffected pairs. In this regard, k-TSP is similar to Random Forests.[39] Third, the underlying models, involving only two genes, are parsimonious and are therefore unlikely to suffer from overfitting. [40]

In the present study, we replaced the simple rank-based gene pair models in k-TSP with linear SVM gene pair discriminant models, and call the approach "PSVM" (pair-wise support vector machine ensembles). This generalization allows for greater flexibility in the selection of gene expression patterns that predict tuberculosis progression. While the k-TSP approach requires the relative ranking of the genes to change between the two conditions (effectively favoring gene pairs that are differentially expressed in opposite directions) any pair of genes that provides non-redundant information for predicting tuberculosis can be combined in a linear SVM discriminant. This was important for tuberculosis progression, where genes with the largest magnitude expression differences between progressors and controls tend to be expressed higher in progressors. By merging the k-TSP approach with SVMs, PSVM is similar to the k-TSP modification proposed by Shi et al., (2011). [25]. The difference between the method of Shi et al. (2011) [25] and PSVM is that the former replaces the ensemble-based structure with a single SVM model, while PSVM retains the ensemble structure and replaces the rank-based pairs with SVMs internally.

Extracting and training the tuberculosis risk signature from the ACS training set:

The PSVM approach was applied to the ACS training set in a manner designed to optimally extract a predictive risk signature from the data. At a high level, the strategy involves three steps (1) changing the cohort time scale to reflect the time before diagnosis with tuberculosis

instead of time since enrolment (since the enrolment time is arbitrary and not related to clinical outcome); (2) deriving a set of candidate predictor genes by comparing gene expression in controls with gene expression in progressors at the time points most proximal to diagnosis; and (3) filtering the candidate predictor genes to retain only those that, in pairwise combinations, robustly discriminate progressors from controls at time points that are more distal to tuberculosis diagnosis. Although this derivation of the tuberculosis risk signature employs multiple time points in the gene selection, applying the signature to assess tuberculosis disease risk only requires expression measurements for the final selected set of genes at single time point and is done in a blinded manner, without any requirement for sample metadata or participant information.

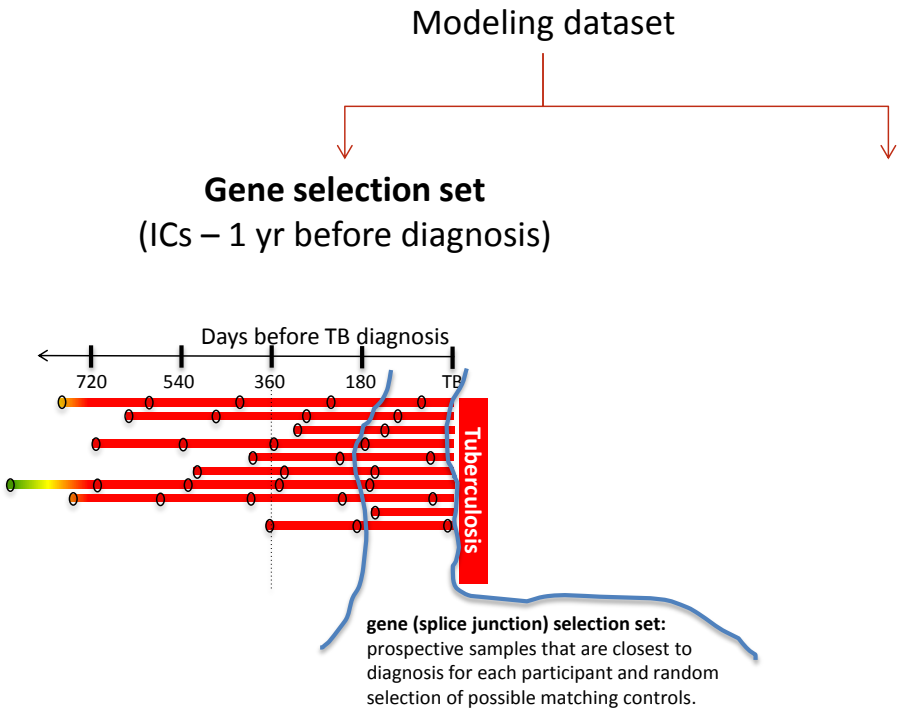
Although it may appear counter-intuitive that samples collected proximally to diagnosis are used to generate candidate genes for the signature of risk, initial cross-validation analyses demonstrated that this approach yields a more predictive signature than approaches that rely solely on samples that are distal to tuberculosis to generate the signature. One interpretation is that while a strong coherent signal is necessary to discover a predictive signature, once developed, the signature can correctly predict tuberculosis on samples where the signal is weaker. This is because less signal is required to recognize an established gene expression pattern than to discover a new one.

Having described the signature development process at a high level above, the following is a description of the process in detail. Development of the signature starts from the normalized splice junction count table of gene expression for the ACS training set. This same procedure was employed for both development of the final tuberculosis risk signature and for cross-validation of the signature generation method. In the case of the former, the signature generation procedure is applied to the entire ACS training set. For the latter, the procedure is applied to the ACS training set after having excluded 1/5 of progressors and matched controls. For simplicity, the dataset that is used for generating the signature will be referred to as the "Modeling dataset" for either case.

(1) ***Re-aligning the Modeling dataset time scale in terms of diagnosis with tuberculosis:*** The dataset was first synchronized according to the tuberculosis diagnosis time point instead of the study enrollment time point (Figure 2B). This allowed identification of which progressor samples (and matched control samples) that had been collected most proximally and most distally to diagnosis.

(2) ***Defining the gene (splice junction) selection set***: This first partition of the data defines which samples are be used to select the genes (splice junctions) that are evaluated as candidates predictors in the signature. The partition was made in the following manner:

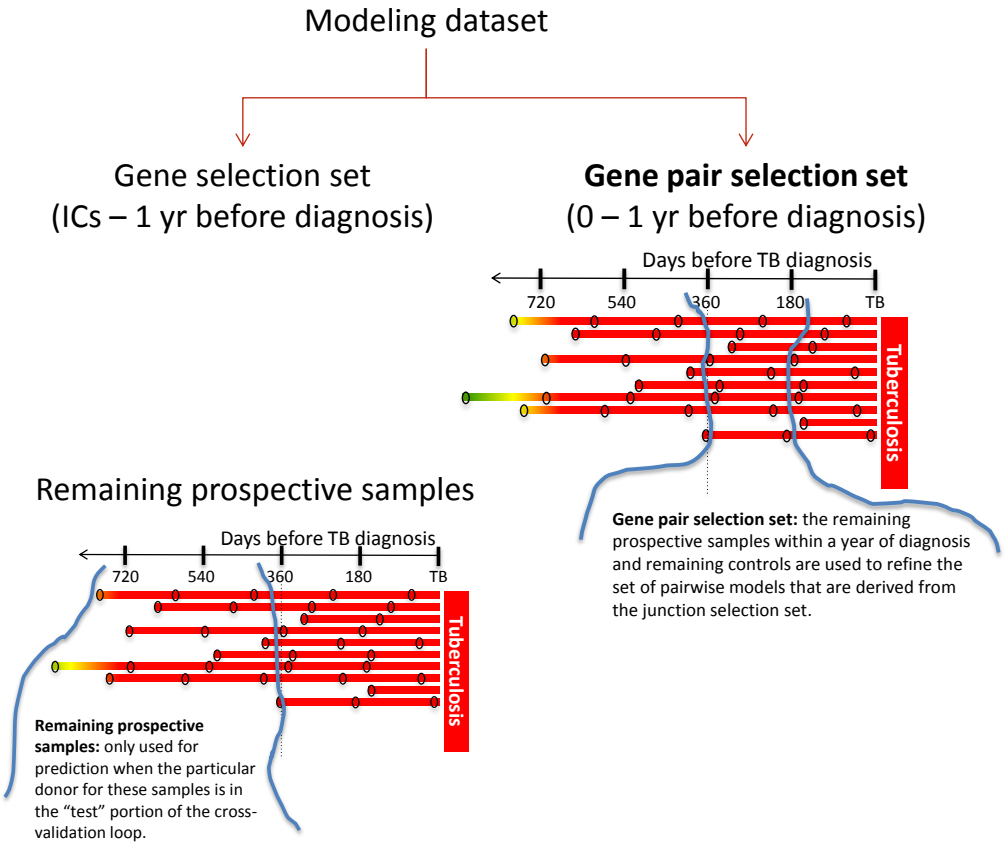
- Only one sample from any progressor or control was included in the junction selection set.
- For subjects with at least one sample less than a year before diagnosis (original estimates, Supplementary Table 2) or at the time of diagnosis (IC = incident case samples, Supplementary Table 2), the sample closest to or at diagnosis was put into the junction selection set.
- Matching control samples for each progressor were randomly selected from possible matches (same blood draw time and demographic bin as the progressor).



(3) *Defining the gene pair selection set:*

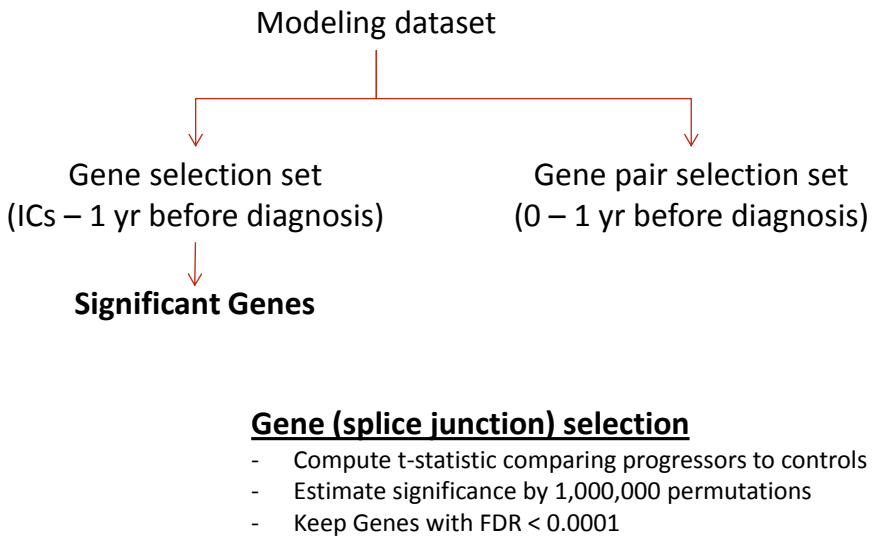
This step defines which samples are used to filter gene pairs for accurate discrimination of progressors from controls at time points that are more distal to diagnosis.

- All control samples and pre-diagnosis progressor samples within a year of tuberculosis diagnosis (by original estimates, Supplementary Table 2) that remained after defining the junction selection set comprised the pair selection set.
- Pre-diagnosis progressor samples that were collected over a year before tuberculosis diagnosis were evaluated in cross-validation but were not used for gene or gene pair selection.



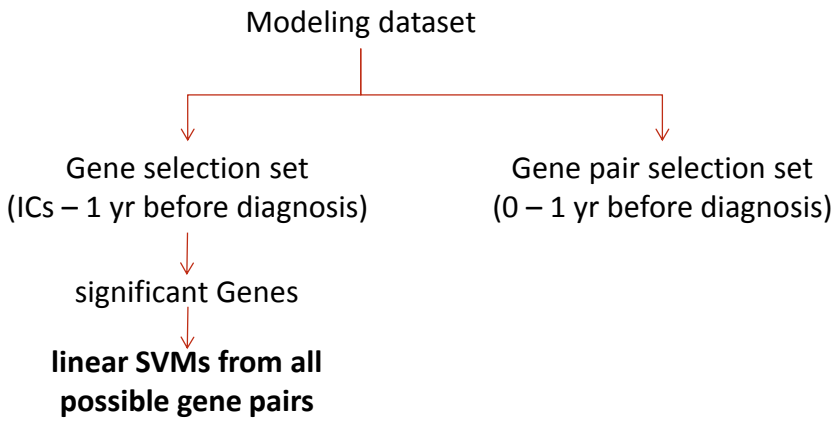
(4) *Gene (splice junction) selection*: Splice junctions that significantly discriminate progressor samples from matched control samples within the gene selection set were identified by permutation test. Briefly, t-statistics comparing progressors and controls were computed for each splice junction. The dataset was then randomly permuted 1,000,000 times. Splice junctions

with t-statistics that significantly deviated from the empirical distribution (FDR<0·0001) were retained.



(5) ***Gene pair model fitting***:

Linear models that discriminate progressors from controls for all possible pair-wise combinations of the selected genes were then fit to the gene selection data set using the Sequential Minimal Optimization SVM algorithm[41] in C++. Fitting of each pair-wise linear SVM to the entire gene selection set was performed independently.

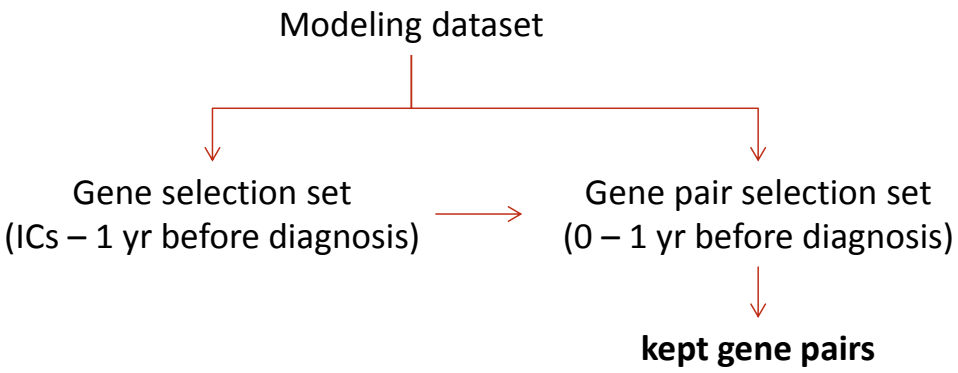


Gene pair model fitting

- Construct an ensemble of Support Vector Machine (SVM) models for all possible pairs
- SVM Gene pair models are trained using **Gene selection** set data

(6) ***Gene pair scoring***:

Each gene pair model was used to predict on the gene pair selection set. Pairs that correctly predicted 70% of control samples, 70% of progressor samples from within 6 months of diagnosis (by original estimates, Supplementary Table 2), and 70% of progressor samples from between 6 months and 1 year before diagnosis (by original estimates, Supplementary Table 2) in the pair selection set were given a score of "1". Junction pairs that did not meet these thresholds were given a score of "0".

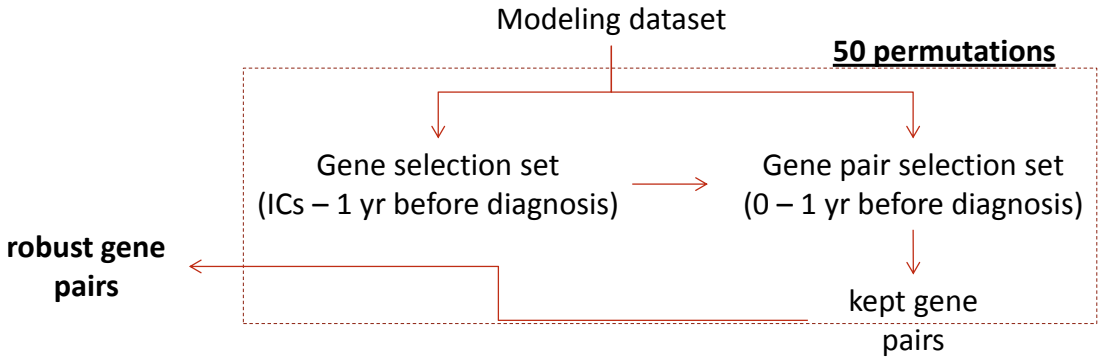


Gene pair scoring

- Test the ability of gene pair models to predict on the gene pair selection set
- Evaluate 3 groups: controls, 0-0.5 yr progressors, 0.5-1 yr progressors
- Retain only those gene pairs that achieve 70% accuracy in all groups

(7) *Gene pair selection*:

After completing one iteration of steps 1-6, the process was repeated starting from a different randomized matching of controls to progressors, yielding slightly different gene selection and gene pair selection sets (steps 1-2), and a different set of scores for the retained gene pairs (step 6). The process was repeated 50 times, with junction pairs receiving scores of 0-50. Gene pairs with scores ≥ 45 were retained for the final ensemble. In this manner, the ensemble is comprised of the most robustly predictive gene pairs.



Gene pair selection

- Repeat the Gene selection/Gene pair selection split 50X (randomized progressor/control matching)
- Retain Gene pairs that are selected in at least 45/50 permutations

(8) *Final model parameterization*:

The gene pairs that were retained after step 6 comprise the PSVM ensemble. Final linear discriminants for each of these pairs were parameterized using the entire set of progressor samples from the Modeling set that were collected less than a year before diagnosis (as indicated by original estimates, Supplementary Table 2) or at the time of diagnosis (IC = incident case samples, Supplementary Table 2) and matched control samples using the Sequential Minimal Optimization SVM algorithm[41] in C++ . As in step 4, fitting of each pair-wise linear SVM to the Modeling set was performed independently.

(9) ***Estimating the prediction accuracy of the TB risk signature by cross-validation***: Prior to making predictions on the ACS test set or adapting the TB risk signature to qRT-PCR, the predictive accuracy of the approach was estimated by performing 100 iterations of cross-validation involving random 4:1 splits of ACS training set participants into training and prediction sets. In each iteration, the entire procedure was repeated starting from the entire ACS training junction expression matrix, ensuring unbiased estimates of prediction accuracy.

(10) ***Benchmarking the signature performance against Random Forest.*** Prior to testing the PSVM signature on the GC6-74 dataset, prediction performance of the signature was benchmarked against the performance of a Random Forest [39] signature consisting of 100,000 trees. We compared the prediction performance of the 100k RF signature to the PSVM signature on two levels: cross-validation of the ACS training set RNA-Seq data (100 iterations of 5-fold splits), and prediction of the ACS test set RNA-Seq data. For both analyses, RF models were constructed in two-step procedure: (1) junction selection was performed by permutation analysis as for the PSVM signature, except that all progressor samples within one year of diagnosis and averaged matched controls were used to compute the t-statistics. (2) RF models were generated from the selected junctions using the R package *randomForest* [42], using default settings, except specifying 100,000 trees. Applying this approach to the entire ACS training set generated the final 100k RF model, which consisted of 1,712 junctions and 631 genes.

D. Adaptation of the tuberculosis risk signature from RNA-Seq to qRT-PCR for validation in the ACS and GC6-74

The signature was adapted to the qRT-PCR platform by matching each splice junction to Applied BioSystems TaqMan® gene expression assays in the following manner. Exact matches between splice junctions and TaqMan assays were identified first and selected. When exact matches were not commercially available, an attempt was made to construct custom Taqman assays spanning the desired junction. If the custom-design process failed, the commercial assay was selected that matched the splice junction of the same gene that was the most strongly correlated to the splice junction of interest. The resulting matches between RNA-Seq designed splice junctions and commercial Taqman assays are provided in Supplementary Tables 17-21. Parameterization of the qRT-PCR-based versions of the signature was performed using qRT-PCR expression data for the ACS Training set and retaining the ensemble structures. Normalization of the cycle threshold data was performed by comparing expression of the signature genes to the set of reference genes. The qRT-PCR-based risk signature was finally generated by re-training the pairwise SVM models to the normalized Ct data using the network structures obtained from RNA-Seq.

After validation of the signatures of risk on the ACS test set, qRT-PCR data for the PSVM primers and reference genes was generated from GC6-74 cohort RNA, in a blinded manner, as described above. Prior to predicting on GC6-74 RNA samples, two modifications were made to the signature. First, failure of one reference primer (GRK6) on the GC6-74 samples necessitated exclusion of this primer and re-parameterization of the signatures (using ACS training set data only). Second, post-hoc inspection of signature prediction on the ACS test set identified a subset of SVM pairs that always voted progressor or always voted control, irrespective of the sample. These pairs were pruned from the networks prior to predicting on GC6-74.

E. Adaptation of tuberculosis risk signature to the Illumina microarray and predicting on published microarray datasets

Uniform normalization of published microarray datasets: Microarray datasets from the published studies of interest were obtained from the Gene Expression Omnibus.[43] For most studies[11-13], the "Series Matrix File (TXT)" was downloaded. For Bloom, *et al.*[9], "GSE40553_non-normalized_SALong.txt.gz" was downloaded. For Bloom, *et al.*[10], "GSE42825_non-normalized.txt.gz", "GSE42826_non-normalized.txt.gz" and "GSE42830_non-

normalized.txt.gz" were downloaded and concatenated. Expression data matrices for each study were converted to a log2 format as needed. For each downloaded dataset, the Illumina microarray probes for all genes present in the set of reference splice junctions (Supplementary Table 11) were identified. Each sample was then normalized by subtracting the average expression of genes detected by these probes from the log2-based data.

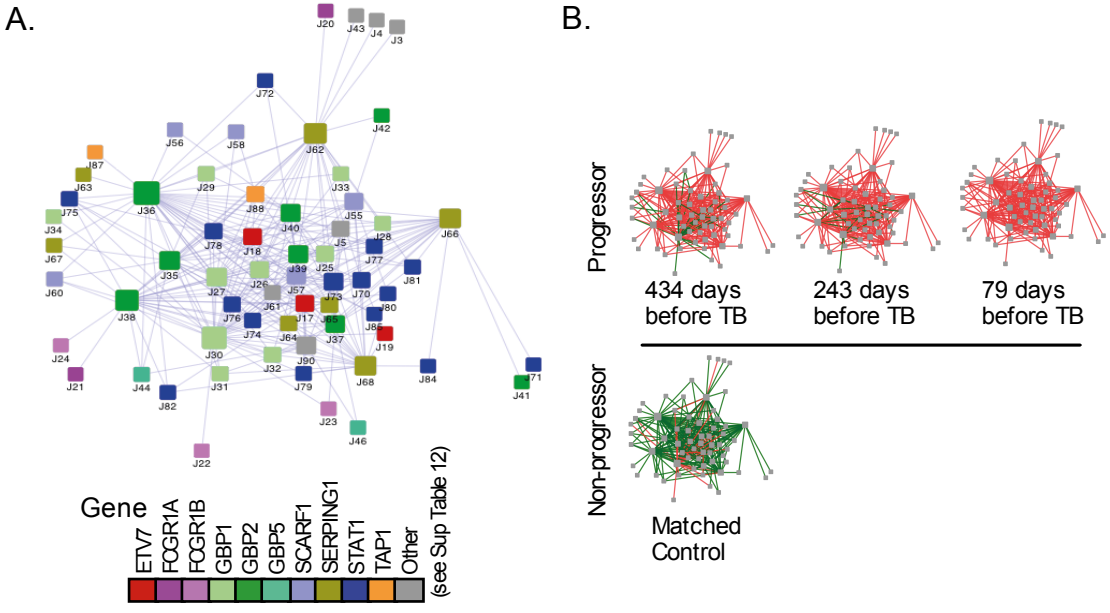
Microarray-based parameterization of the TB risk signature: The tuberculosis risk signature was adapted to the Illumina microarray platform by first collapsing each splice junction rule pair into a gene pair. All Illumina probes corresponding to the two genes in the rule were identified, and all possible probe pairs using these two sets of probes were constructed. Rules were then trained (using linear SVMs) on for each probe pair using the tuberculosis disease patient and latently infected control data from the UK Training Set from Berry, *et al.*[11] The resulting Illumina microarray-based signature was locked down and remained unmodified for making predictions on the other microarray cohorts from the same study[11] and other studies.[9,10,12,13]

Supplementary Tables

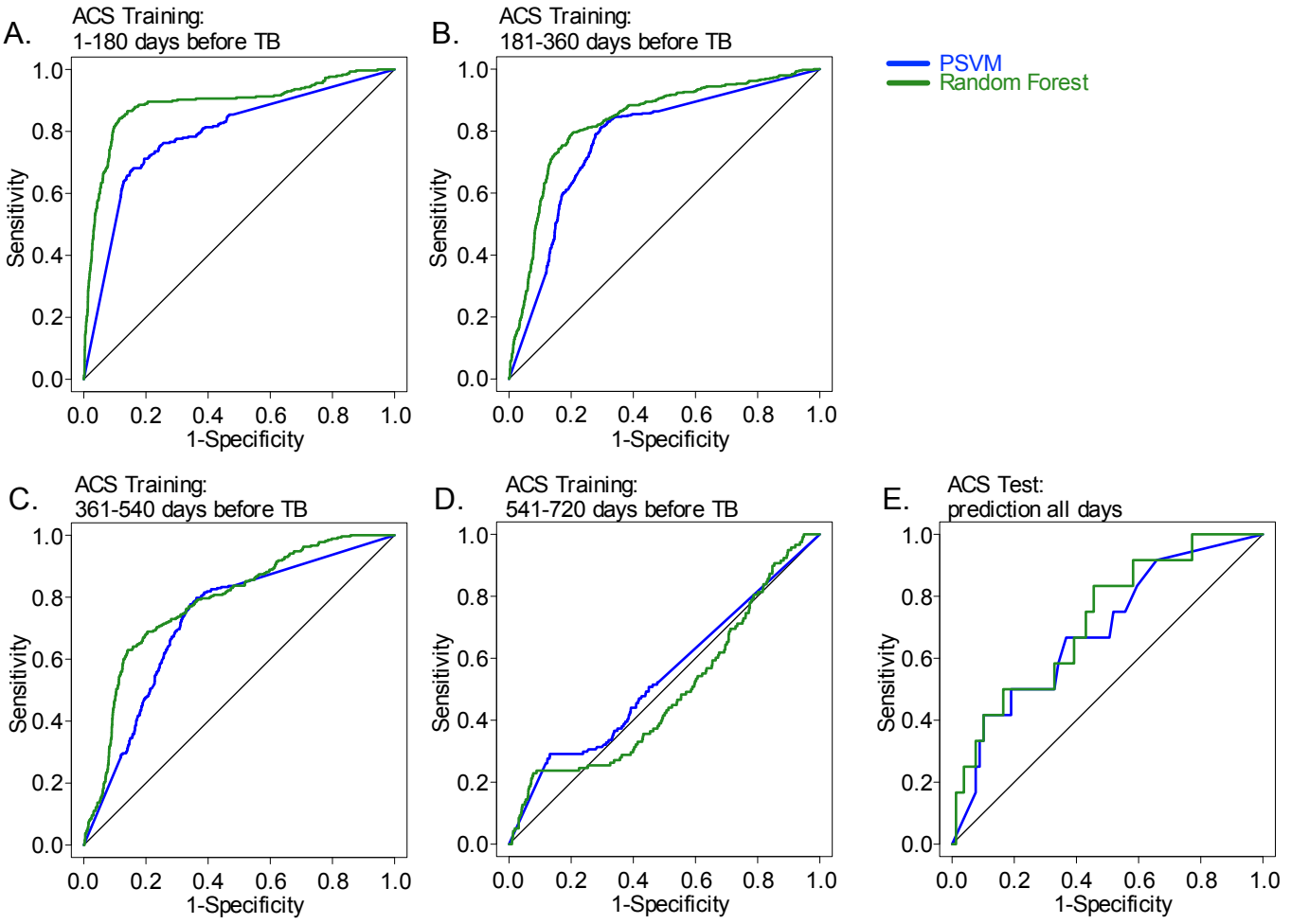
1. (SupTab1_ACS_Progressors): Tuberculosis diagnosis in ACS progressors
2. (SupTab2_ACS_TimeToDiagnosis): Original and per protocol (PP) sampling time points with respect to tuberculosis diagnosis for progressors from the ACS cohort.
3. (SupTab3_GC6Progressors): Tuberculosis diagnosis in GC6-74 progressors.
4. (SupTab4_GC6TimeToDiagnosis): Sampling time points with respect to tuberculosis diagnosis for progressors from the GC6-74 cohort.
5. (SupTab5_SampleNumbers): Number of progressors and controls and corresponding samples for identifying and validating signatures of risk using the ACS training, ACS test, and GC6-validation cohorts.
6. (SupTab6_RNASeqMetadata): Meta-data for RNA-Seq samples from ACS training and test cohorts.
7. (SupTab7_ReferenceJunctions): Reference junctions used for internal normalization of the signature.
8. (SupTab8_SignatureJunctions): Splice junctions, genes and chromosomal locations that comprise the signature
9. (SupTab9_SignatureJunctionPairs): RNA-Seq junction pairs and SVM parameters that comprise the signature.
10. (SupTab10_SignatureJunctionData): Raw RNA-Seq splice junction count data for all junctions in the signature.
11. (SupTab11_TBRiskComputationSheet): Worksheet that calculates risk scores upon input of raw splice junction count data (from SupTab10_SignatureJunctionData)
12. (SupTab12_RefJunctionPrimers): Primers corresponding to the reference junctions that were used in the signature
13. (SupTab13_Signatureprimers): Primers corresponding to the splice junctions contained in the signature.
14. (SupTab14_SignaturePrimerPairs): Signature primer pairs used for blind prediction of ACS Test set and GC6-74 Validation set.
15. (SupTab15_BenchmarkRF): ROC AUCs for the ACS training and test benchmark analysis against Random Forest

16. (SupTab16_SignatureScores): Per-sample TB risk score and correct classification for ACS training fit (RNA-Seq and qRT-PCR), ACS test predict (RNA-Seq and qRT-PCR), GC6 predict (qRT-PCR).
17. (SupTab17_MicroarrayAnalysis): Details and prediction statistics for the TB risk signature-based meta-analyses of published microarray datasets.
18. (SupTab18_TBAssociations): Genes in the TB risk signature and known associations with TB and/or relevant functions.
19. (SupTab19_SignatureGeneMatrix): Genes in the tuberculosis risk signature and diagnostic signatures (Berry *et al*., 2010; Kaforou *et al.,* 2013; Anderson *et al.,* 2014).
20. (SupTab20_SignatureSimilarity): Intersections between the tuberculosis risk signature and diagnostic signatures (Berry *et al*., 2010; Kaforou *et al*., 2013; Anderson *et al.,* 2014).
21. (SupTab21_ModuleEnrichments): Obermoser *et al*. 2013[44] module enrichments for genes that are common to TB risk and diagnostic signatures or specific to diagnostic signatures.
22. (SupTab22_CellTypes): Heatmap and significance counts for cell-type specific expression patterns of the PSVM signature transcripts based on data from Bloom *et al*., 2013.

Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3