Multinomial modelling of TB/HIV co-infection yields a robust predictive signature and generates hypotheses about the HIV+TB+ disease state.

Fergal J Duffy[1], Ethan G. Thompson[1], Thomas J. Scriba[2], and Daniel E Zak[1§]

[1]Center for Infectious Disease Research (formerly Seattle Biomedical Research Institute), Seattle WA, USA

[2]South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine & Department of Pathology, University of Cape Town, Cape Town, South Africa

[§]Corresponding author

Email addresses:

FJD: fergal.duffy@cidresearch.org

EGT: ethan.thompson@cidresearch.org

TJS: thomas.scriba@uct.ac.za

DEZ: daniel.zak@cidresearch.org

Abstract

Background. Current diagnostics are inadequate to meet the challenges presented by co-infection with *Mycobacterium tuberculosis* (Mtb) and HIV, the leading cause of death for HIV-infected individuals. Improved characterisation of Mtb/HIV coinfection as a distinct disease state may lead to better identification and treatment of affected individuals.

Methods. Four previously published TB and HIV co-infection related datasets were used to train and validate multinomial machine learning classifiers that simultaneously predict TB

and HIV status. Classifier predictive performance was measured using leave-one-out cross validation on the training set and blind predictive performance on multiple test sets using area under the ROC curve (AUC) as the performance metric. Linear modelling of signature gene expression was applied to systematically classify genes as TB-only, HIV-only or combined TB/HIV.

Results. The optimal signature discovered was a single 10-gene random forest multinomial signature that robustly discriminates active tuberculosis (TB) from other non-TB disease states with improved performance compared with previously published signatures (AUC: 0.87), and specifically discriminates active TB/HIV co-infection from all other conditions (AUC: 0.88). Signature genes exhibited a variety of transcriptional patterns including both TB-only and HIV-only response genes and genes with expression patterns driven by interactions between HIV and TB infection states, including the CD8+ T-cell receptor LAG3 and the apoptosis-related gene CERKL.

Conclusions. By explicitly including distinct disease states within the machine learning analysis framework, we developed a compact and highly diagnostic signature that simultaneously discriminates multiple disease states associated with Mtb/HIV co-infection. Examination of the expression patterns of signature genes suggests mechanisms underlying the unique inflammatory conditions associated with active TB in the presence of HIV. In particular, we observed that disregulation of CD8+ effector T-cell and NK-cell associated genes may be an important feature of Mtb/HIV co-infection.

Keywords: Tuberculosis; HIV; co-infection; machine learning; multinomial; interferon; microarray; blood transcription

Background

Almost ¼ of the global population is infected with *Mycobacterium tuberculosis* (Mtb) [1] and over 1,600,000 people succumbed to active tuberculosis disease (TB) in 2016 alone[2]. TB ordinarily requires at least six months of antibiotic treatment in order to remove all traces of the infection, with drug resistant strains requiring two years of intensive treatment[3]. Individuals with HIV/AIDS are at particularly high risk of active TB, up to 30 times higher than for HIV- individuals prior to the start of antiretroviral therapy(ART)[4]. This relative risk declines after the initiation of ART, but still remains 2-3 times higher than the general population, and the biological mechanisms underlying this increased risk remain unclear.

The current standard for diagnosis of active TB is microscopic or culture-based detection of *M. tuberculosis* bacteria in a patient-derived sputum sample. Sputum-based tests suffer from several major limitations, including the amount of time it takes to culture slow-growing TB, and the necessity of having sufficient TB bacteria in the sputum for detection. This is a further issue for children and TB cases in HIV+ patients[5] where low numbers of TB bacilli in the sputum may give a false-negative result. The Xpert MTB/RIF test[6] has enabled rapid TB diagnosis by detecting the presence of *M.tb*-specific DNA in sputum, but the sensitivity of this test is diminished in sputum culture-negative TB[7]. Sputum also represents a dangerous vector of infection for health-care workers analysing and handling sputum samples, due to the potential presence of live *M. tb* bacteria[8]. New TB diagnostic methods that do not necessitate the detection of large numbers of TB bacilli in sputum are therefore critically required to serve populations at high risk of TB. Blood-based signatures are an attractive alternative, as blood is a clinically accessible readout of the immunological state of the body.

Whole blood gene expression signatures that are diagnostic for TB have been described in many previous studies[9–12], but these signatures are generally focused on a single binary comparison, e.g. latent TB vs active TB or active TB vs other diseases. In this study, we

79 develop multi-class multinomial signatures that explicitly model the TB and HIV state of each
80 patient. Our analysis integrates published data from several cohorts and evaluate a range of
81 machine-learning approaches to generate a multinomial model that specifically discriminates
82 TB from non-TB disease states while simultaneously discriminating HIV+ TB as a unique
83 disease state.
84 Figure 1 shows the analytical plan for this work.
85

Methods

All computational and statistical analyses were performed using the R language for statistical computing[13].

Microarray Normalisation, Probe Filtering and Data Preparation

Microarray data from four TB/HIV cohorts were downloaded from GEO: GSE37250, GSE39941, GSE19491, GSE42834. The precise sample compositions of each of these datasets are provided in Table 1 (GSE37250) and Table 3 (GSE39941, GSE19491, GSE42834).

Microarray datasets were downloaded in the form of GEO series matrix files, background subtracted and quantile normalized. All reference and variable probe selection was performed using the GSE37250 Malawi adult data. 20 reference probes were selected by eliminating any probe with any expression value below the absolute value of the smallest expression value observed in the dataset. References were then selected as the 20 remaining probes with the smallest inter-quartile range (IQR) of expression in the dataset.

To pre-select probes likely to be discriminatory for TB, only probes with an IQR of above 1.5 (log2 normalised expression) were kept. This resulted in 554 candidate model-variable probes to be used in the model. Additional File 1 lists the variable and reference probes used in this work.

Before model training, each sample in every dataset was normalised by calculating the mean expression level for the reference probes, and subtracting this mean reference level from each model-variable probe. The GSE19491 dataset, which was measured using the Illumina 12v3 platform as opposed to the Illumina 12v4 platform used for the other datasets, was missing 4 of the 20 selected reference probes. The remaining 16 probes alone were therefore used to normalise this dataset.

Machine-Learning Model Training, Feature Shrinking and Model Selection

Five distinct machine-learning algorithms were used to train predictive models on the adult South African dataset, using the R caret[14]package. These were:

[1] Random Forest (RF) (R randomForest package[15]), is an algorithm based on training an ensemble of decision trees using randomly split subsets of the training samples and training variables, all of which then 'vote' to classify new samples.

[2] Support Vector Machine (SVM) using RBF kernel (R kernlab[16] package). SVMs attempt to find the optimal linear hyperplane decision boundary separating the two classes in n-dimensional space, where n is the number of features the SVM is trained on. The RBF, or Radial Basis Function kernel projects this n-dimensional feature space into a higher dimension to allow the identification of a linear decision boundary in a higher dimensional space if one cannot be found in the input n-dimensional space.

[3] Neural Networks (NN) (R nnet[17] package). NNs are comprised of a network of input nodes (1 per-feature), connected to output nodes (1 per possible outcome) via one or more 'hidden' layers of nodes. Each node represents a logistic regression function, based on the input value, and the weight given to each input node (which in turn determines the output classification) is determined during training.

[4] Elastic-net Logistic Regression (R glmnet[18] package), is a form of logistic regression with regularisation of the linear coefficients applied to control overfitting.

[5] K-Nearest Neighbor (KNN) (R caret[14] package), classifies samples by determining the 'k' most similar samples by Euclidian distance between sample genes and having them 'vote' on the classification.

All of these algorithms can be trained to produce binary (exactly 2 distinct classes, such as TB vs LTB) or multinomial (more than 2 classes) classifier models.

In this study, each algorithm was trained using normalised microarray data comprising all 554 most-variable probes (selected as described above) on each of four subsets of samples of the adult training data. These subsets were 1) the entire dataset, including TB, latent TB (LTB)

and other disease (OD) samples including HIV+ and HIV- samples, 2) All TB and LTB samples including HIV+ and HIV- samples (i.e. OD excluded), 3) All HIV+ TB and LTB samples, (i.e. all HIV- and OD excluded), 4) All HIV- TB and LTB samples (i.e. all HIV+ and OD excluded). Models were trained to simultaneously predict both the TB and HIV status of each training sample, i.e. models trained on subset 1) explicitly classified samples as one of 6 classes: TB:HIV+, TB:HIV-, LTB:HIV+, LTB:HIV-, OD:HIV+, or OD:HIV-; models trained on subset 2) classified samples as one of 4 classes TB:HIV+, TB:HIV-, LTB:HIV+, or LTB:HIV-, and models trained on subsets 3) and 4) were binary models that classified samples as TB or LTB only, as HIV status was constant in these subsets.

After models were trained on the initial 554 probes, the models were sequentially shrunk to obtain probe-reduced models that only comprised the most important 250, 50, 25, 15, or 10 probes from the initial set. Probe importance rankings were calculated using the *varImp* function supplied by the caret package. This function implements algorithm-specific methods for evaluating how much each probe contributes to the classification performance of the model. In the case of random forests, the difference in out-of-bag error[19] with and without the inclusion of a single probe was used to rank the probes in order of importance. For Elastic-net logistic regression models, the probe variable coefficient was used to rank the probes. For Neural Networks, Garson's algorithm[20] was used to calculate probe importance from network weights. For Random Forests, probe importance was measured as the difference in predictive performance comparing all trees that contain the probe with trees that lack that probe. For the remaining modelling approaches (SVM, KNN), the univariate predictive power of the individual probe was used to rank importance in an algorithm-independent way.

For each algorithm-subset combination, the most important 250 out of the original 554 probes were selected, the model re-trained on this subset of probes, and this probe-reduced model used to predict the left-out sample. This procedure was then repeated to iteratively shrink each model to contain the 50, 25, 15 and 10 most important probes from the previous step. This entire process was carried out for each held-out sample in the cross validation so that the

sequential shrinking and prediction steps were performed independently for every held out sample.

Training performance was assessed using leave-one-out cross validation (LOOCV). Initially, a single sample from the overall training set was held out, then a classifier model was trained on the remaining samples, and used to predict the status of the held-out sample. This procedure was repeated for every training sample, and model performance was then calculated using predictions on the held-out samples. Areas under the ROC curves (AUCs) for model discrimination of TB vs non-TB samples in the training subset were calculated for predictions on the held out samples, and this was used as the performance metric for initial model structure selection. For models that predicted more than 2 classes, TB predictions were calculated as the sum of TB-related prediction classes, e.g for 6 class models the overall TB prediction value was calculated as the TB:HIV+ prediction value plus the TB:HIV- prediction value.

Only "small" models (i.e., those consisting of 10, 15 or 25 probes) were considered for application to the test sets. To choose the algorithm/class-complexity/training set between these 3 probe sizes, initially the 10-probe model was selected. If either the 15 or 25-probe model showed significantly better LOOCV performance on the training set, that model was used. Significance was evaluated by comparing ROC AUCs using the *roc.test* function from the pROC[21] R package with a threshold of $p<0.05$.

Model Predictions on New Datasets.

After models were selected by recursive LOOCV evaluation as described above, each selected model was retrained using the most-commonly selected features from the LOOCV and re-parameterised on the entire relevant training data subset. Predictions were made using the *predict* function from the caret package to calculate class probabilities, and prediction accuracies were assessed by calculating TB vs non-TB ROC curves using the R pROC

193 package. As described for the LOOCV procedure above, in the case ofmodels that predicted
194 more than 2 classes, TB predictions were calculated as the sum of TB-related prediction
195 classes, e.g for 6 class models the overall TB prediction value was calculated as the TB:HIV+
196 prediction value plus the TB:HIV- prediction value.Performance of the new signatures on the
197 test sets was benchmarked against the performance of two previously-reported TB gene
198 signatures: the three gene multi-cohort diagnostic signature developed by Sweeney et al[27],
199 which we term the 'threeGene' signature; and our 16-gene correlate of TB risk [23], which
200 we term the 'ACS' signature (referring to the Adolescent Cohort Study from which the
201 signature was derived). For predictions using the threeGene signature, datasets were
202 downloaded in raw non-normalised format from GEO before being quantile normalised and
203 baseline corrected using the log-exponential method using the R limma package[22]. The
204 threeGene score was then directly calculated as (GBP5 + DUSP3)/2 – KLF2. For the ACS
205 model predictions, datasets were prepared and normalised and scored as described in[23].

206 **Linear modelling of disease state**

207 Linear regression models were fit to signature genes in order to assess the contribution of TB
208 and HIV status to gene expression. Expression of each gene was fit to a linear regression
209 model (R *lm* function) of the form:

210 $$Expr = a * TB_{status} + b * HIV_{status} + c * TB_{status}{:}\,HIV_{status}$$

211 TB and HIV status were encoded as binary variables with 1 meaning active TB/HIV+ and 0
212 meaning latent TB/HIV-. Non-TB other disease samples were excluded from this analysis.
213 The p-value of the model coefficients a, b and c were calculated using the R *summary.lm*
214 function, and a false discovery rate correction applied.
215

216 **Gene set enrichment analysis**

217 Gene set enrichment analysis was performed using the R tmod[24] package, using the blood
218 transcriptional gene sets previously described by Li et al[25] and Chaussabel et al[26]. P-

values were calculated using the hypergeometric test as implemented in the tmodHGtest function, using all included microarray gene symbols as the background.

Results

Development, cross-validation and selection of multinomial machine learning models for predicting TB and HIV.

We used data from a previously-published cohort [9] of 537 adults from Malawi and South Africa, comprised of samples from individuals diagnosed with active tuberculosis (TB), latent tuberculosis (LTB) or other non-TB diseases with clinical symptoms consistent with TB (OD). Roughly half of these individuals were also HIV+ (Table 1). These transcriptional profiles were used to develop and test multinomial machine learning approaches to specifically identify each symptomatic subset.

Machine learning models were trained on the South African adult dataset described in Table 1, with the Malawian adults used as an independent test set. In order to focus on the strongest signal probes, an initial down-selection step was performed where only probes with a log2 normalised expression interquartile range of at least 1.5 in the South African set were considered for model training (554 probes).

Models were trained to classify all or relevant subsets of the data into 2 (binary classifier), 4 (multinomial), or 6 (multinomial) classes, using a diverse panel of machine-learning algorithms. Two-class models were trained to classify a sample as either active or latent TB. Two different two class models were trained for each algorithm, one on HIV- TB and LTB samples only, and another on HIV+ TB and LTB samples only. Four-class models were trained to classify a sample as active or latent TB and as HIV+ or HIV- simultaneously, using all TB and LTB samples, both HIV+ and HIV-. Six-class models were trained to classify a sample as active TB, latent TB or other disease, and as HIV+ or HIV-, and were trained on the entire dataset, including TB, LTB and OD, both HIV+ and HIV-. Machine learning

algorithms used were Random Forests, Neural Networks, Support Vector Machines, Elastic-Net Logistic Regression, and k-Nearest Neighbours.

Initially, each model was trained using all 554 pre-selected probes. Starting from this initial model, the most important model probes were selected and the models recursively shrunk to use smaller numbers of probes (see Methods). Leave-one-out cross validation (LOOCV) performance on the training set was evaluated by measuring area under the ROC curve (AUC). Figure 2 shows the results of the LOOCV and recursive shrinking for each algorithm (

Figure 1). LOOCV AUCs are uniformly strong, almost all above 0.8. LOOCV performance of multinomial 4- and 6- class models is similar to that of the binary classification models.

An ideal model shows high predictive performance based on a small number of interpretable genes. A set of small models for further analysis were chosen by initially selecting the smallest (10 probe) model for each algorithm and classification complexity, and only selecting a larger model if it showed significantly stronger LOOCV performance. As performance, illustrated in

Figure 1 (a) and (b), was largely uniform across model sizes, the 10 probe models were universally selected. Table 2 lists the training cross-validation performance of each of these models, in terms of their area under the ROC curve. From Table 2, in three of the four complexity cases for the South Africa training set, Random Forest was the highest-performing algorithm, thus Random Forest models were selected for all further analyses.

A six-class multinomial model outperforms previously published signatures for identifying active TB in several independent test sets

The Malawian adults were used as an independent test set for the selected models.

Figure 3 (A) shows ROC curves representing the predictive ability of South Africa-derived models to specifically identity HIV- and HIV+ active TB samples vs latent TB and other diseases. These models were accompanied by two previously-reportedTB signatures: our 16-gene correlate of TB risk [23], termed here as the ‘ACS’ signature, and the three-gene multi-

cohort diagnostic signature developed by Sweeney et al[27], termed here as the 'threeGene' signature. The multinomial six-class Random Forest model outperformed all other models (AUC: 0.88, sensitivity 80%, specificity 82%), although performance of the threeGene model was very similar (AUC: 0.87 vs AUC 0.88).

To more thoroughly validate the six-class multinomial model, classification performance was evaluated using three additional previously-published whole-blood microarray datasets[10–12]. To reduce technical sources of variability as much as possible, test sets were selected that used the same Illumina HumanHT-12 microarray platform as was used for the Kaforou cohort. These additional test sets comprised a broad range of samples, including adult and childhood TB; TB vs other inflammatory, bacterial and pulmonary diseases; and samples taken from a range of geographical locations (Table 3). The two-class HIV- model was also included as a comparator for the six-class multinomial model.

Figure 3 (B) shows the ROC curves for the 10-gene six-class model, the 10-gene binary model and the two previously-described external signatures. Again, the six-class 10-gene signature was the overall top performer (AUC 0.88, sensitivity 80%, specificity 82%). This is significantly better predictive performance than the top external model, the threeGene signature ($p=0.006$ by a single tailed DeLong[28] test). Thus, multinomial modelling of TB disease states significantly improved the accuracy of discrimation of TB vs. non-TB samples.

The 10-gene multinomial signature identifies HIV+TB as a distinct disease state

To further evaluate the performance of the 10-gene six-class signature, particularly the potential of this signature to perform multi-class discrimination, we tested whether it can specifically identify HIV+ TB samples from all other samples in the Malawi test set (Figure 4 (A)). The signature accurately discriminated HIV+ TB from HIV- TB samples (AUC: 0.88) and HIV+ TB samples from all other samples (HIV+ TB, HIV-/+ latent TB, HIV-/+ other diseases, AUC: 0.86). This result suggests that HIV+ TB may exist as a distinct transcriptional state. Box- and dot-plots of normalised expression for the 10 genes in the

multinomial signature for active and latent TB individuals in the combined South-African and Malawian cohorts, stratified by TB and HIV status reveal a diverse pattern of transcriptional responses (
Figure 4 (B)).

Using HIV- latent TB samples as a baseline, 8 of the 10 signature genes are either downregulated in both active TB and HIV+ patients, or upregulated in both active TB and HIV+ patients. Two exceptions are LAG3 and CERKL. To more quantitatively determine the transcriptional patterns of these genes, linear models with gene expression as a function of TB and HIV status were fit, including a TB:HIV interaction term. Genes with significant (FDR<0.01) TB or HIV model coefficients were identified as TB- or HIV- independent disease signature genes, and genes with both TB and HIV significant coefficients were identified as overlapping signature genes (
Figure 4 (C)). Three of the ten genes in the signature were independent disease signature genes, with CD160 being the sole HIV-specific gene and CD36 and ZDHHC19 as the only TB-specific genes. The largest group of signature genes exhibits a unidirectional additive expression pattern, either downregulated in both TB and HIV (CD40LG, ID3) or upregulated in both TB and HIV (GBP6, C1QB). Interestingly, the CD8+ immune checkpoint gene LAG3 is upregulated in HIV+ individuals but downregulated in active TB. Two genes exhibit a significant interaction term: FCGR1B and CERKL. This interaction suggests crosstalk between the TB and HIV transcriptional response. In the case of FCGR1B, transcription reaches a saturated level in HIV- active TB that is not exceeded in HIV+ active TB. As FCGR1B is a cell surface receptor specific to macrophages, monocytes and neutrophils [EBI Expression Atlas, www.ebi.ac.uk/gxa], this saturation point may correspond with a maximum surface density of receptor or a maximal blood concentration for these cell types. CERKL, a negative regulator of apoptosis caused by oxidative stress[29], shows a more complex regulatory pattern where HIV- active TB is upregulated compared to all other states.

Biological pathways associated with divergent TB/HIV expression patterns reveal HIV+ TB as a distinct disease state.

Analysis of the genes comprising the 10-gene six-class signature identified LAG3 and CERKL as exhibiting distinct expression patterns. As the signature genes reflect a minimal set of genes necessary to classify disease states, we hypothesised that there may be other genes closely correlated with LAG3 and CERKL that could shed light on the biological processes driving the opposing regulation they exhibit.

The expression of LAG3 was correlated very tightly (Spearman $\rho > 0.8$, $p < 1e\text{-}32$), with a set of eight genes similarly downregulated in active TB and upregulated in HIV (Table S2, Figure 5). Mapping these genes to blood transcriptional genes sets[25,26] revealed significant enrichment for cytotoxic T-cell and NK-cell pathways (Table S3), suggesting that dysregulation of immune effector cells is a distinguishing characteristic of HIV+ active TB when compared to HIV- TB, or HIV+ latent TB.

In contrast, CERKL does not show similarly strong correlations (Spearman $\rho>0.8$) with any individual gene. At a more permissive correlation threshold ($\rho>0.6$), CERKL is correlated with 9 genes (Table S4), but this set of genes does not show significant enrichment for any gene set. Genes most strongly correlated with CERKL are the ribosomal-RNA processing gene HEATR1 and the ubiquitin ligase TRIM13.

Discussion

In a clinical setting, a major challenge faced regarding TB diagnosis is to discriminate active TB from other diseases presenting with similar symptoms. The ROC curves shown in Figure 3 shows that the 10 gene six-class signature identified in this work significantly improves on existing signatures for identifying active TB in a wide variety of contexts, i.e. active TB vs healthy samples, active TB vs latent TB and active TB vs other diseases, with or without the presence of HIV co-infection. A major advantage of the meta-analytical approach taken here is the testing of each signature on a combination of cohorts at once. While ROC

analysis can reveal the optimal classification performance on a single cohort, it is still necessary to choose an operating point or threshold to transform a continuous score into a dichotomous classifier. It is possible for a predictive signature to show a high sensitivity and specificity on many individual cohorts separately, but fail to recreate this performance when samples from all cohorts are combined. This is due to the signature score potentially having a differing optimal classification threshold on each cohort and will not be revealed by separate ROC analysis of each cohort. By combining cohorts, signatures with a stable "global" operating score are revealed. Thus, it can be seen in Figure 3 (B) that the across every cohort, the ten-gene six-class signature predicts with a sensitivity of 80% and specificity of 78%.

Explicit modelling of each cohort disease group has allowed us to hone in on transcriptional processes that specifically distinguish HIV+ active TB from HIV- active TB. Characterisation of a specific transcriptional state for HIV+ TB would improve understanding of how HIV increases TB risk, as well as illuminating on essential elements of an effective host response to TB missing from HIV+ TB patients. This analysis identified the CD8+ inhibitory checkpoint receptor LAG3. Linear modelling reveals significant upregulation of LAG3 in HIV infection, but also significant downregulated of LAG3 in active TB compared with latent TB (Figure 4 (B)). Upregulation of LAG3 is known to suppress T-cell activity in chronic HIV infection[30], and these exhausted T-cells show impaired production of the cytokines such as IL-2, IFNγ, and TNF, associated with an effective host response to TB[31]. LAG3 expression is also closely correlated with genes including the CD8A receptor; the CD8+ T-cell secreted chemokine CCL5/RANTES; the NK-cell granule gene NKG7, the CD8+ differentiation transcription factor EOMES, and the lysosomal membrane protein MCOLN2. All of these genes are involved in CD8+ or NK-cell effector activities. Thus, further investigation of a key signature gene has revealed that both innate (NK cell) and adaptive (CD8+ T-cell) effector

function appears to be suppressed in HIV+ active TB relative to HIV- active TB, suggesting at least one mechanism for increased TB risk in HIV+ individuals.

Interestingly, another CD8+ T-cell inhibitory checkpoint receptor, CD160, was also selected as a signature gene. CD160 shows a similar pattern of expression to LAG3: upregulated in HIV+ patients, but downregulated in active TB. However, this downregulation in active TB is much less pronounced than for LAG3, and the TB coefficient was not found to be significant in linear modelling (FDR=0.08).

Overexpression of CERKL has been shown to protect cells from apoptosis while under oxidative stress [29]. The expression pattern of CERKL, which shows lower expression in HIV+ active TB compared to both HIV- active TB and HIV+ latent TB indicates that HIV/Mtb co-infected patients may have impaired protection against cellular death due to oxidative stress. This expression pattern hints at a complex balance of apoptotic signalling in HIV/TB co-infection that does simply mirror the interferon-driven inflammatory response.

Conclusions

We have identified a broadly applicable active TB-specific 10-gene multinomial signature by validating candidate signatures with successively harder problems: training a diverse panel of candidate models on an adult test set; making blind predictions an independent adult test set from a different geographical cohort, albeit from the same study; making blind predictions on the combination of the adult test set with three additional independent cohorts; and finally testing for discrimination of HIV+ TB from HIV- TB.

While the signature shown here does not reach the diagnostic sensitivity required to be a practical alternative to sputum culture for clinical use (>98% sensitivity for culture positive TB)[3], it represents an incremental performance improvement over previously described signatures. All of the blood-based signatures evaluated in this work (the ten-gene six-class signature, the threeGene signature and the ACS signature) show similar performance on the test datasets examined here, performance which falls below that observed with traditional sputum culture. While whole blood gene expression signatures do not appear likely to

approach the performance of liquid culture, it is possible that whole blood signatures can be developed to improve diagnosis of TB cases who cannot produce sputum or who have paucibacillary disease, including HIV+ TB cases. Unfortunately, the lack of a "gold-standard" method of diagnosing TB when sputum culture cannot be obtained makes it extremely difficult to accurately evaluate blood transcriptional signatures in this context.

A possible practical application of this test is as a high-specificity "triage test" that can rule out patients unlikely to have TB, and identify persons who should receive a full sputum culture, thus reducing the necessity of working with difficult-to-acquire and potentially infectious sputum samples. At an operating point of 95% sensitivity, the ten-gene random forest shows a specificity of 47%. In a situation such as a medical clinic in a TB-endemic area, assuming 50% of patients presenting with symptoms consistent with TB have active TB, treating signature positive patients immediately would almost half the amount of sputum culture necessary.

List of Abbreviations

Mtb: Mycobacterium Tuberculosis; AUC: Area under the receiver-operator curve; ROC: receiver operator curve; LOOCV: Leave-one-out cross-validation. OD: other diseases; TB: active tuberculosis; LTB: latent tuberculosis; ACS: Adolescent Cohort Study; RF: random forest; SVM: support vector machine; NN: neural net; KNN k-nearest neighbours.

Declarations

Ethics approval and consent to participate.

Not applicable

Consent for publication:

All authors have read and contributed to this manuscript, and approve of its publication

Availability of data and material:

All microarray datasets used in this study are available on GEO

Competing interests

The authors declare that they have no competing interests

Funding

This work was 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. FJD was supported by the NCDIR, National Institutes of Health grant [U54 GM103511].

Authors' contributions

FJD, EGT, TJS and DEZ designed the analyses, assisted the interpretation of results, and revised the manuscript. FJD carried out the analysis and drafted the manuscript.

References

1. Houben RMGJ, Dodd PJ. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. PLoS Med. 2016;13:1–13.
2. World Health Organization. Tuberculosis Fact Sheet No 104 [Internet]. 2016. Available from: http://www.who.int/mediacentre/factsheets/fs104/en/#
3. Gardiner JL, Karp CL. Transformative tools for tackling tuberculosis. J. Exp. Med. [Internet]. 2015;212:1759–69. Available from: http://www.jem.org/cgi/doi/10.1084/jem.20151468
4. Lawn SD, Wood R, De Cock KM, Kranzer K, Lewis JJ, Churchyard GJ. Antiretrovirals and isoniazid preventive therapy in the prevention of HIV-associated tuberculosis in settings with limited health-care resources. Lancet Infect. Dis. [Internet]. Elsevier Ltd; 2010;10:489–98. Available from: http://dx.doi.org/10.1016/S1473-3099(10)70078-5
5. Elliott a. M, Namaambo K, Allen BW, Luo N, Hayes RJ, Pobee JOM, et al. Negative sputum smear results in HIV-positive patients with pulmonary tuberculosis in Lusaka, Zambia. Tuber. Lung Dis. 1993;74:191–4.
6. Boehme CC, Nicol MP, Nabeta P, Michael JS, Gotuzzo E, Tahirli R, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: A multicentre implementation study. Lancet [Internet]. Elsevier Ltd; 2011;377:1495–505. Available from: http://dx.doi.org/10.1016/S0140-6736(11)60438-8
7. Lawn SD, Mwaba P, Bates M, Piatek A, Alexander H, Marais BJ, et al. Advances in

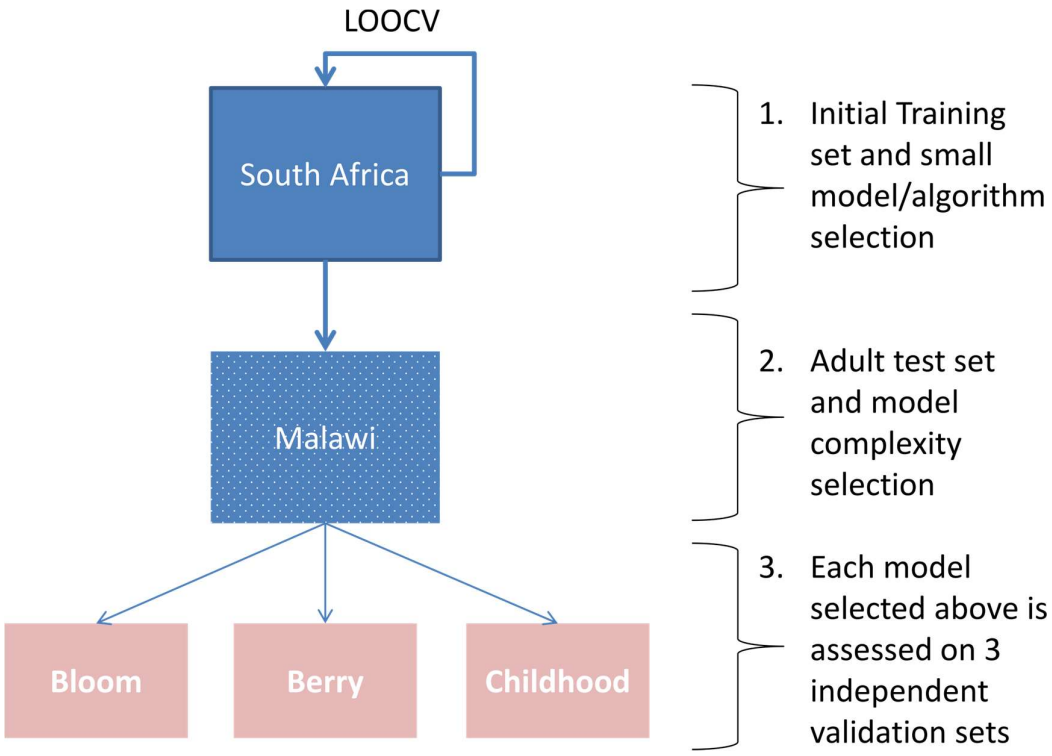
461 tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-
462 care test. Lancet Infect. Dis. [Internet]. Elsevier Ltd; 2013;13:349–61. Available from:
463 http://linkinghub.elsevier.com/retrieve/pii/S1473309913700082
464 8. Jensen P a, Lambert L a, Iademarco MF, Ridzon R. Guidelines for preventing the
465 transmission of Mycobacterium tuberculosis in health-care settings, 2005. MMWR.
466 Recomm. Rep. 2005.
467 9. Kaforou M, Wright VJ, Oni T, French N, Anderson ST, Bangani N, et al. Detection of
468 Tuberculosis in HIV-Infected and -Uninfected African Adults Using Whole Blood RNA
469 Expression Signatures: A Case-Control Study. PLoS Med. [Internet]. 2013;10. Available
470 from: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001538
471 10. Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluka G, et al.
472 Diagnosis of childhood tuberculosis and host RNA expression in Africa. N. Engl. J.
473 Med. [Internet]. 2014;370:1712–23. Available from:
474 http://www.ncbi.nlm.nih.gov/pubmed/24785206
475 11. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch S a a, Oni T, et al. An interferon-
476 inducible neutrophil-driven blood transcriptional signature in human tuberculosis.
477 Nature [Internet]. Nature Publishing Group; 2010;466:973–7. Available from:
478 http://dx.doi.org/10.1038/nature09247
479 12. Bloom CI, Graham CM, Berry MPR, Rozakeas F, Redford PS, Wang Y, et al.
480 Transcriptional Blood Signatures Distinguish Pulmonary Tuberculosis, Pulmonary
481 Sarcoidosis, Pneumonias and Lung Cancers. PLoS One. 2013;8.
482 13. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
483 Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from:
484 https://www.r-project.org/
485 14. Kuhn M. Building Predictive Models in R Using the caret Package. J. Stat. Softw.
486 [Internet]. 2008;28:1–26. Available from: http://www.jstatsoft.org/v28/i05/
487 15. Liaw A, Wiener M. Classification and Regression by randomForest. R news
488 [Internet]. 2002;2:18–22. Available from: http://cran.r-project.org/doc/Rnews/
489 16. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab – An S4 Package for Kernel
490 Methods in R. J. Stat. Softw. [Internet]. 2004;11:1–20. Available from:
491 http://www.jstatsoft.org/v11/i09/paper
492 17. Venables WN, Ripley BD. Modern Applied Statistics with S. Issues of Accuracy and
493 Scale [Internet]. 2002;868. Available from:
494 http://scholar.google.com/scholar?q=related:29252RNVomwJ:scholar.google.com/&hl
495 =en&num=100&as_sdt=0,5%5Cnpapers2://publication/uuid/E2522B53-755A-48D4-
496 925F-A02094784F2C

497 18. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear
498 Models via Coordinate Descent. J. Stat. Softw. [Internet]. 2010;33:1–22. Available from:
499 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2929880&tool=pmcentrez
500 &rendertype=abstract
501 19. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray
502 data using random forest. BMC Bioinformatics [Internet]. 2006;7:3. Available from:
503 http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-3
504 20. Garson DG. Interpreting neural-network connection weights. Artif. Intell. Expert.
505 1991;6:46–51.
506 21. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an
507 open-source package for R and S+ to analyze and compare ROC curves. BMC
508 Bioinformatics [Internet]. 2011;12:77. Available from:
509 http://www.biomedcentral.com/1471-2105/12/77
510 22. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers
511 differential expression analyses for RNA-sequencing and microarray studies. Nucleic
512 Acids Res. 2015;43:e47.
513 23. Zak DE, Penn-Nicholson A, Scriba TJ, Thompson E, Suliman S, Amon LM, et al. A
514 blood RNA signature for tuberculosis disease risk: a prospective cohort study. Lancet
515 [Internet]. Elsevier Ltd; 2016;6736:1–11. Available from:
516 http://linkinghub.elsevier.com/retrieve/pii/S0140673615013161
517 24. Weiner J. tmod: Feature Set Enrichment Analysis for Metabolomics and
518 Transcriptomics [Internet]. 2016. Available from: https://cran.r-
519 project.org/package=tmod
520 25. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, et al.
521 Molecular signatures of antibody responses derived from a systems biology study of
522 five human vaccines. Nat. Immunol. [Internet]. 2014;15:195–204. Available from:
523 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3946932&tool=pmcentrez
524 &rendertype=abstract
525 26. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A Modular
526 Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus
527 Erythematosus. Immunity. 2008;29:150–64.
528 27. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis
529 of pulmonary tuberculosis: a multicohort analysis. Lancet Respir. Med. [Internet].
530 Elsevier Ltd; 2016;2600:1–12. Available from:
531 http://linkinghub.elsevier.com/retrieve/pii/S2213260016000485
532 28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or

533 more correlated receiver operating characteristic curves: a nonparametric approach.
534 Biometrics. 1988;44:837–45.
535 29. Tuson M, Garanto A, Gonzàlez-Duarte R, Marfany G. Overexpression of CERKL, a
536 gene responsible for retinitis pigmentosa in humans, protects cells from apoptosis
537 induced by oxidative stress. Mol. Vis. [Internet]. 2009;15:168–80. Available from:
538 http://www.ncbi.nlm.nih.gov/pubmed/19158957
539 30. Tian X, Zhang A, Qiu C, Wang W, Yang Y, Qiu C, et al. The upregulation of LAG-3
540 on T cells defines a subpopulation with functional exhaustion and correlates with
541 disease progression in HIV-infected subjects. J. Immunol. [Internet]. 2015;194:3873–82.
542 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25780040
543 31. Jayaraman P, Jacques MK, Zhu C, Steblenko KM, Stowell BL, Madi A, et al. TIM3
544 Mediates T Cell Exhaustion during Mycobacterium tuberculosis Infection. PLoS Pathog.
545 [Internet]. 2016;12:e1005490. Available from:
546 http://www.ncbi.nlm.nih.gov/pubmed/26967901
547 32. Sweeney TE, Shidham A, Wong HR, Khatri P. A comprehensive time-course–based
548 multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic
549 gene set. Sci. Transl. Med. [Internet]. 2015;7:287ra71-287ra71. Available from:
550 http://stm.sciencemag.org/content/7/287/287ra71.abstract

551

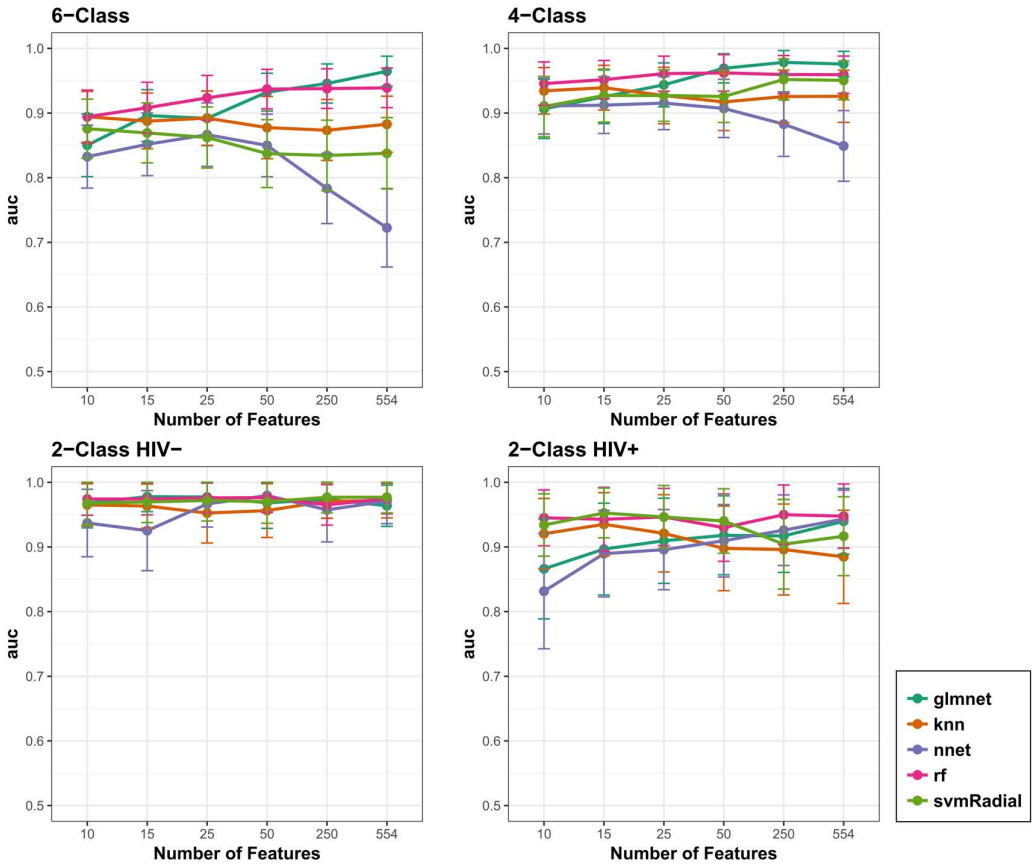
552 # Figures



553

554 **Figure 1: Analytical Plan.**

555 Outlines the model training, selection and prediction steps for the overall analysis



556

557 **Figure 2: Training Cross Validation Results on Adult TB Samples**

558 Leave-one-out cross validation results for models trained on South Africa adult data. Each
559 panel plots the area under the ROC curve for five machine learning algorithms (glmnet:
560 Elastic-Net logistic regression, knn: k-Nearest Neighbours, nnet: Neural Network, rf: Random
561 Forest, svmRadial, Support Vactor Machine with Radial Basis Function kernel) starting with
562 models trained using all 554 probes, and iteratively shrunk to models trained on 10 probes
563 only. Models were trained to classify the data into 6 (TB:HIV+, TB:HIV-, LTB:HIV+,
564 LTB:HIV-, OD:HIV+, OD:HIV-), 4 (TB:HIV+, TB:HIV-, LTB:HIV+, LTB:HIV-) and 2
565 (TB, LTB) classes. Two types of 2-class models were trained: using all HIV+ or all HIV-
566 samples. Error bars show bootstrap-estimated 95% confidence intervals around the AUC.

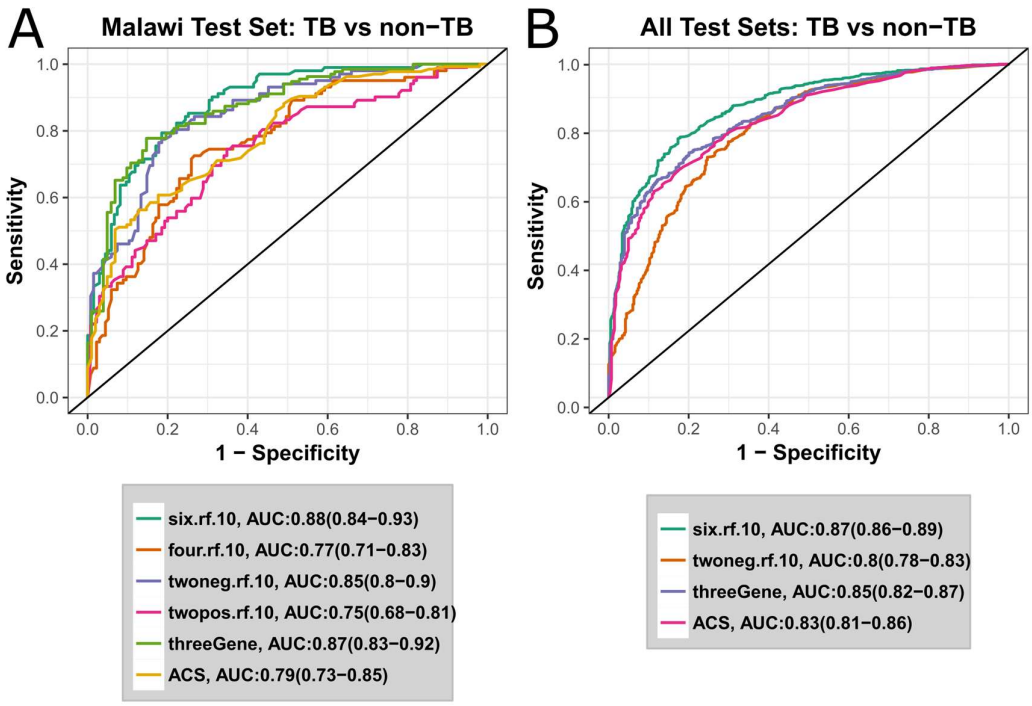
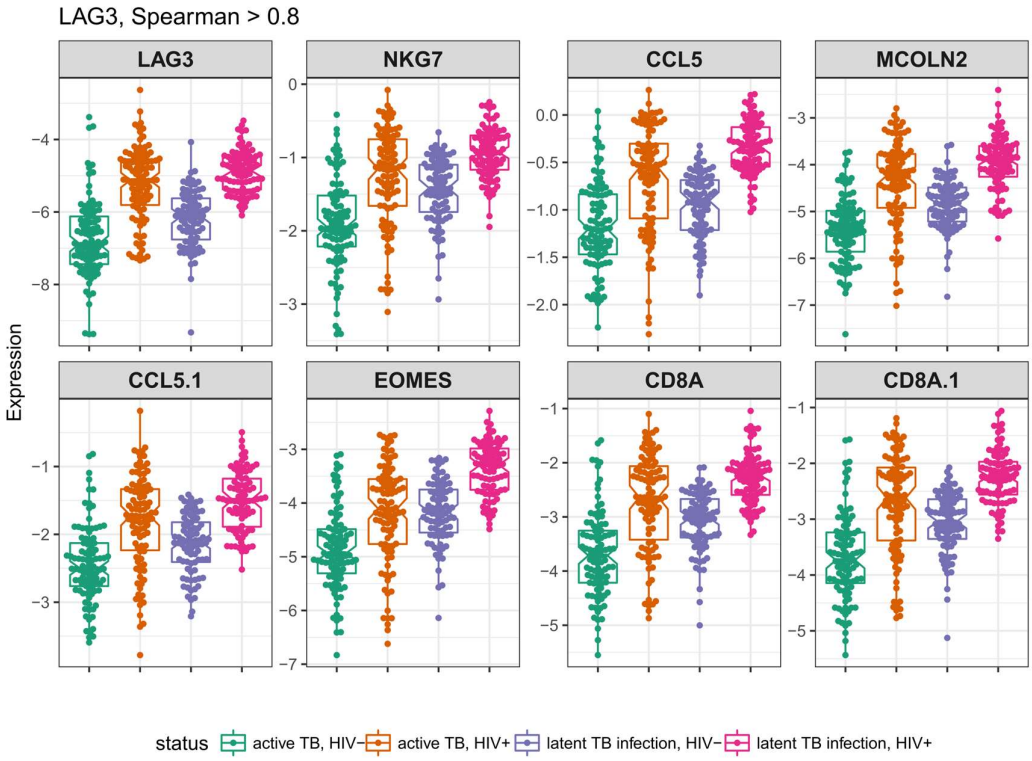


Figure 3: A six-class multinomial model optimally predicts 4 independent test sets.

ROC curves for active TB vs non-TB classification of independent test sets. Legends shows the AUC for each model, with the 95% confidence intervals in parentheses. Models developed in this study are named in the form <number-of-classes>.<algorithm>.<number-of-probes>. E.g. six.rf.10 is the 10 probe random forest model trained to predict 6 classes. twoneg and twopos refer to 2-class models trained on HIV- or HIV+ samples respectively. threeGene refers to the signature described by Khatri et al[32], and ACS refers to the signature described by Zak et al[23]. **A** ROC curves for classification of the Malawi test samples from the Kaforou cohort. **B** ROC curves for Malawi test set plus the three further independent test sets described in Table 3



579

580 **Figure 4: The six-class multinomial model identifies HIV+ TB as a distinct state**

581 **A** ROC curves for the 10-gene six-class multinomial model discriminating HIV+ active TB
582 samples from HIV- active TB samples, and HIV+ active TB samples from HIV- active TB
583 and HIV+/- LTB and HIV+/- OD samples in the Malawi test set. **B** Dot- and boxplots of
584 expression levels of six-class multinomial model genes in the entire Kaforou dataset. **C** Six-
585 class multinomial genes classified by their TB/HIV behaviour as determined by fitting linear
586 models to gene expression as a function of disease state. TB upregulated genes are indicated
587 in orange and downregulated genes shown in blue

588

589 **Figure 5: LAG3-correlated genes**

590 Dot and boxplots for each microarray primer, named as the corresponding gene, strongly
591 correlated with LAG3 (spearman correlation $\rho>0.8$) for latent and active TB samples from the
592 Kaforou dataset.

Tables

	Active TB		Latent TB		Other Diseases		**TOTAL**
	HIV+	**HIV-**	**HIV+**	**HIV-**	**HIV+**	**HIV-**	
Malawi	**51**	**51**	**36**	**35**	**30**	**34**	**237**
South Africa	**47**	**46**	**48**	**48**	**62**	**49**	**300**

Table 1: Adult Training and Test Set

	Number of Classes to Predict	Elastic-Net Logistic Regression	k-Nearest Neighbours	Neural Network	Random Forest	Support Vector Machine
South Africa	Four	0.91(0.86-0.95)	0.93(0.89-0.97)	0.91(0.87-0.95)	0.96(0.93-0.99)	0.91(0.86-0.96)
	Six	0.82(0.76-0.87)	0.89(0.85-0.93)	0.85(0.80-0.90)	0.92(0.88-0.95)	0.88(0.83-0.92)
	Two (HIV-)	0.98(0.94-1.0)	0.96(0.93-1.0)	0.97(0.93-1.0)	0.97(0.95-1.0)	0.97(0.94-1.0)
	Two (HIV+)	0.85(0.77-0.93)	0.91(0.86-0.97)	0.84(0.76-0.93)	0.95(0.90-1.0)	0.94(0.89-0.99)

596 **Table 2: Top LOOCV Model Structures**

597 Shows the area under the ROC curve (AUC) for the best small model at each combination of
598 machine-learning algorithm choice and number of predicted classes for models trained on
599 South Africa adult data. 95% confidence intervals of the AUC are shown in parentheses.

600

GEO ID	Description	Number of Samples	Reference
GSE39941	Whole blood microarray expression analysis of TB vs LTB and Other Diseases with potential HIV co-infection in Children from Kenya, Malawi and South Africa	122 TB (HIV-, 27 are culture-TB) 68 TB (HIV+, 17 are culture- TB) 68 LTB (HIV-) 140 OD (HIV-) 93 OD (HIV+)	Anderson et al, 2014[10]
GSE42834	Whole blood microarray expression analysis of TB vs Sarcoidosis, Pneumonia and Lung Cancer and Healthy Controls in Adults from the UK	35 TB 61 Sarcoidosis 121 Healthy 16 Lung Cancer 6 Pneumonia	Bloom et al, 2013[12]

GSE19491	Whole blood microarray expression analysis of TB vs LTB, Healthy controls and other bacterial and inflammatory diseases (Streptococcus, Staphylococcus, Still's Disease, Lupus) in patients from the UK and South Africa	54 TB 69 LTB 105 Healthy 110 Lupus 40 Staph 31 Still's 12 Strep	Berry et al, 2010[11]

Table 3: Further Validation Datasets