PLOS ONE

RESEARCH ARTICLE

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

Fergal J. Duffy[1,2], **Ethan G. Thompson**[2], **Thomas J. Scriba**[3], **Daniel E. Zak**[2]*

1 Seattle Children's Research Institute, Center for Global Infectious Disease Research, Seattle, WA, United States of America, **2** Center for Infectious Disease Research (formerly Seattle Biomedical Research Institute), Seattle, WA, United States of America, **3** South African Tuberculosis Vaccine Initiative, Institute of Infectious Disease and Molecular Medicine & Department of Pathology, University of Cape Town, Cape Town, South Africa

* danielzak000@gmail.com

Citation: Duffy FJ, Thompson EG, Scriba TJ, Zak DE (2019) Multinomial modelling of TB/HIV co-infection yields a robust predictive signature and generates hypotheses about the HIV+TB+ disease state. PLoS ONE 14(7): e0219322. https://doi.org/10.1371/journal.pone.0219322

Editor: Lars Kaderali, Universitatsmedizin Greifswald, GERMANY

Received: December 18, 2018

Accepted: June 20, 2019

Published: July 15, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

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 (award to TJS). FJD was supported by the National Center for Dynamic Interactome Research National Institutes of Health grant [P41 GM109824] (http://www.ncdir.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 characterization 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 10-gene random forest multinomial signature that robustly discriminated active tuberculosis (TB) from other non-TB disease states with improved performance compared with previously published signatures (AUC: 0.87), and specifically discriminated 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.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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 neighbors.

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 dysregulation of CD8+ effector T-cell and NK-cell associated genes may be an important feature of Mtb/HIV co-infection.

Introduction

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 analyzing 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 (LTB) vs active TB (TB) or active TB vs other diseases (OD). In this study, we develop multinomial signatures, defined as signatures that predict more than two classes simultaneously (as opposed to signatures limited to a binary classification). We train these signatures to explicitly model the TB and HIV state of each patient simultaneously.

Our hypothesis is that it is possible to improve the predictive performance and generalizability of a TB transcriptional signature by explicitly training a multinomial signature to predict multiple TB and HIV disease states. Our analysis integrates published data from several cohorts and evaluates a range of machine-learning approaches to generate a multinomial model that specifically discriminates TB from non-TB disease states while also discriminating HIV+ TB as a unique disease state. We validate this multinomial signature on transcriptional profiles incorporating a variety of disease states taken from both adults and children from geographically diverse locations.

Table 1. Study datasets. Whole-blood microarray profiles used in this study for training, testing, and validating signatures, as shown in Fig 1. Sample counts are shown for each dataset stratified by HIV and TB status.

GEO ID	Country	Active TB		Latent TB		Other Diseases		Healthy Controls	Total	Use	Reference	Other details
		HIV +	HIV-	HIV +	HIV-	HIV +	HIV-					
GSE37250	**South Africa**	47	46	48	48	62	49	-	300	Initial cross-validation set	Kaforou [9]	
GSE37250	**Malawi**	51	51	36	35	30	34	-	237	First test set and final model selection	Kaforou [9]	
GSE39941	**Malawi and South Africa**	68	122	-	68	93	140	-	491	Further validation set	Anderson [10]	Of the active TB patients, 17 HIV + and 27 HIV- have negative TB cultures
GSE42843	**UK**	-	35	-	-	-	83	121	239	Further validation set	Bloom [12]	Other disease breakdown: 61 sarcoidosis, 16 lung cancer, 6 pneumonia
GSE19491	**South Africa and UK**	-	54	-	69	-	192	105	420	Further validation set	Berry [11]	Other disease breakdown: 110 lupus, 40 staph, 31 still's, 12 strep

https://doi.org/10.1371/journal.pone.0219322.t001

Methods

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

Microarray normalization, probe filtering, and data preparation

Microarray data from four TB cohorts representing a range of geographical sites, including other diseases, and including both HIV- and HIV- individuals were downloaded from GEO: GSE37250, GSE39941, GSE19491, GSE42834. To reduce technical sources of variability as much as possible, training and test cohorts were selected that used the same Illumina HumanHT-12 microarray platform. 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. The precise sample compositions of each of these datasets are provided in Table 1 (GSE37250, GSE39941, GSE19491, GSE42834).

Microarray datasets were downloaded in the form of GEO series matrix files, background subtracted, and then quantile normalized were performed. All reference and variable probe selection was performed using the GSE37250 Malawi adult data. Probes with any expression values below the determined background noise level were removed.

To pre-select probes likely to be discriminatory for TB, only probes with an IQR of above 1.5 (log2 normalized expression) were kept. This resulted in 554 candidate model-variable probes to be used in the model. S1 File lists the variable and reference probes used in this work.

In order to facilitate inter-cohort comparisons, 20 constitutively expressed reference probes were selected representing the 20 probes with the smallest inter-quartile range (IQR) of expression in the dataset. Before model training, each sample in every dataset was normalized using the standard-score method by calculating the mean log2(expression) level for the reference probes and then subtracting this mean reference level from the log2(expression) value for each model-variable probe. Thus, expression of each probe in each dataset was calculated relative to the constitutively expressed reference probes. The GSE19491 dataset was measured using the Illumina 12v3 platform as opposed to the Illumina 12v4 platform used for the other datasets.

Four of the 20 reference probes were not available for the GSE19491 dataset. Therefore, the remaining 16 reference probes were used to normalize this dataset.

Machine-learning model training, reduction, and selection

Six 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 Radial Basis Function kernel (RBF) 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 regularization 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.
6. Extreme Gradient Boosting (xgbTree) (R *xgboost* [19] package) iteratively trains an ensemble of decision trees, before 'boosting' by fitting a new model to the residual error of the original model in a regularized manner.

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.

Microarray datasets consist of probe-based measurements, where each probe represents the expression of a single gene or gene splice-form variant. All machine learning training was done using microarray probes as input features. Each algorithm was trained using normalized 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.

Model prediction performance on the training set was assessed using leave-one-out cross validation (LOOCV). Initially, a single sample from the overall training set was held out and a classifier model was trained on the remaining samples and then used to predict the status of the held-out sample.

Hyperparameter tuning was performed automatically by passing the 'tuneLength = 3' option to the R *caret*::*train* function. This automatically selected a range of 3 predefined values for each hyperparameter, as implemented in the *caret*::*getModelInfo("modelName")$grid* function. Grid search was used to select the combination of hyperparameters that maximized the AUC for each model structure in the training set. Hyperparameter optimization was carried out for each fold in the training CV, with only predictions for the optimal hyperparameter set being retained. The full hyperparameter search was re-run on the entire training set to parameterize the final model.

Following training on the entire set of variable probes (554 probes), probe importance scores were calculated using the *caret*::*varImp* function. 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 and extreme gradient boosting, the difference in out-of-bag error [20] 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 [21] 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 calculate importance in an algorithm-independent way.

During LOOCV, each model was reduced to obtain models that were comprised of only the 250 most important probes from the initial set; and this reduced model was also used to predict the left-out sample. This procedure was then repeated to iteratively to reduce each model to contain the 50, 25, 15 and 10 most important probes from the previous step, and then used to predict the held-out sample. This recursive model-reduction procedure was repeated for every training sample.

Model performance was evaluated using area under the ROC curve (AUC) for discrimination of TB vs non-TB in the held-out samples for each algorithm-model size combination. 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 evaluation on 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 *pROC*::*roc.test* function [22] package with a threshold of $p<0.05$.

In this manuscript, models are names according to the pattern <classes-predicted>.<algorithm>.<number-of-probes> as defined in S2 File.

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-parameterized on the entire relevant training data subset. Predictions were made using the

caret::*predict* function to calculate class probabilities, and prediction accuracies were assessed by calculating TB vs non-TB ROC curves using the R *pROC* package. As described for the LOOCV procedure above, in the case of 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. Performance of the new signatures on the test sets was benchmarked against the performance of two previously-reported TB gene signatures: the three gene multi-cohort diagnostic signature developed by Sweeney et al [23], which we term the 'threeGene' signature; and our 16-gene correlate of TB risk [24], which we term the 'ACS' signature (referring to the Adolescent Cohort Study from which the signature was derived). For predictions using the threeGene signature, datasets were downloaded in raw non-normalized format from GEO before being quantile normalized and baseline corrected using the log-exponential method using the R *limma* package [25]. The threeGene score was then directly calculated as (GBP5 + DUSP3)/2 –KLF2. For the ACS model predictions, datasets were prepared and normalized and scored as described in [24].

Linear modelling interpretation of signature-probe expression patterns

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

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

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

Gene set enrichment analysis

Gene set enrichment analysis was performed using the R *tmod* [26] package, using the blood transcriptional gene sets previously described by Li et al [27] and Chaussabel et al [28]. P-values were calculated using the hypergeometric test as implemented in the *tmod*::*tmodHGtest* function, using all included microarray gene symbols as the background.

Results

Fig 1 shows the overall analytical plan for signature development, including cross-validation, testing and independent validation.

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

We selected a previously-published cohort [9] of 537 adults from Malawi and South Africa that was 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

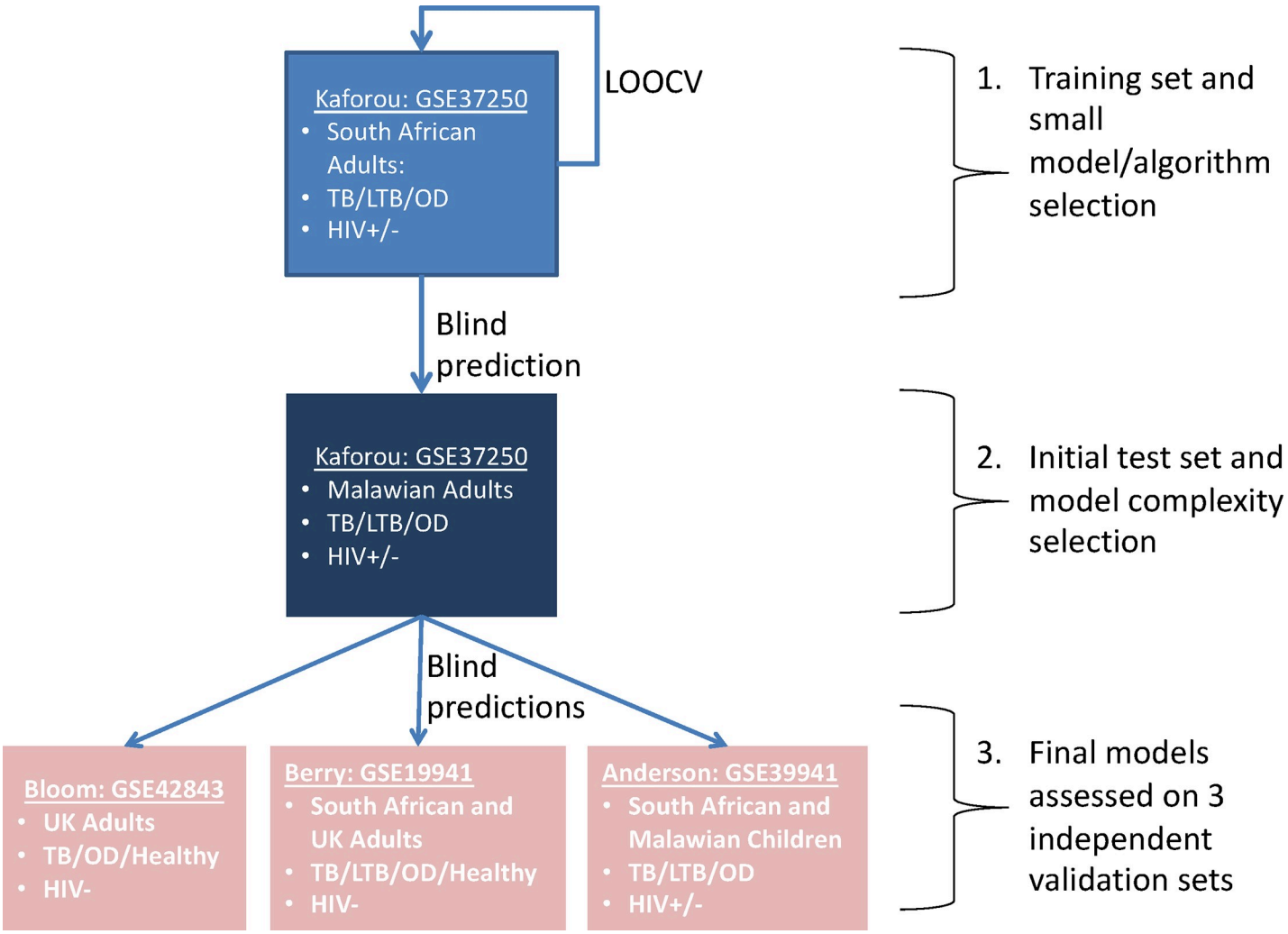

Fig 1. Analytical plan. Illustrates the workflow and whole-blood microarray datasets used to train, evaluate (using leave-one-out cross validation, LOOCV) and test predictive models. Each box represents a dataset, identified by GEO ID and first author name. Bullet points list the available geographical sites and TB and HIV status available for samples making up each dataset. Sample counts for each site are also shown in Table 1.

https://doi.org/10.1371/journal.pone.0219322.g001

probes with a log2 normalized 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 (rf), Neural Networks (nnet), Support Vector Machines (svmRadial), Elastic-Net Logistic Regression (glmnet), k-Nearest Neighbors (knn), and Extreme Gradient Boosting (xgbTree) (Table 2, S2 File).

Table 2. Key to model structure names. Models trained in the manuscript are named in the form: <classes-predicted>.<algorithm>.<number-of-probes>, where classes-predicted is one of the options specified under Model Complexities and algorithm is one of the options specified under Model Algorithms. Thus, the model named *six.rf.25* indicates a six-class multinomial random forest model based on 25 microarray probes. For comparison, two external models have been included, as indicated under External Models.

Model Complexities	
Classes Predicted	**Description**
six	Six-class multinomial model, predicts the following classes
	[TB.HIV+, TB.HIV-, LTB.HIV+, LTB.HIV-, OD.HIV+, OD.HIV-]
four	Four-class multinomial model, predicts the following classes
	[TB.HIV+, TB.HIV-, LTB.HIV+, LTB.HIV
twopos	Two-class binary model, predicts TB or LTB, trained on HIV+ samples only
twoneg	Two-class binary model, predicts TB or LTB, trained on HIV- samples only
Model Algorithms	
Algorithm	**Description**
glmnet	Logistic regression with elastic-net regularization (L1 and L2)
knn	K-nearest neighbours
nnet	Neural network
rf	Random forest
svmRadial	Support vector machine with radial basis function kernel
xgbTree	Extreme gradient boosting
External Models	
External Model Name	**Description**
threeGene	Three-gene TB diagnostic signature, published by Sweeney et al (2016), consisting of the genes *GBP5*, *KLF2* and *DUSP3*.
ACS	Signature of risk TB progression derived from South African adolescents with latent TB. Based of splice-junctions expression from 16 genes.

https://doi.org/10.1371/journal.pone.0219322.t002

Initially, each model was trained using all 554 pre-selected probes. In order to obtain parsimonious models in which the number of model probes was less than the number of training samples, a model reduction procedure was performed. Starting from the initial 554-probe model, the most important model probes were selected and the models recursively shrunk to use sequentially 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). Fig 2 shows the results of the LOOCV and recursive shrinking for each algorithm. The LOOCV AUCs were all above 0.8. LOOCV performance of multinomial 4- and 6-class models were similar to those obtained using 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. Performance, illustrated in Fig 2(A) and 2(B), was largely comparable across model sizes, indicating that small models performed comparably to larger models. Thus, the parsimonious 10-probe models were selected for all further analysis. Table 3 lists the training cross-validation performance of each of these models, in terms of their area under the ROC curve. In three of the four cases for the South Africa training set, Random Forest was the highest-performing algorithm and was thereby selected for all further analyses.

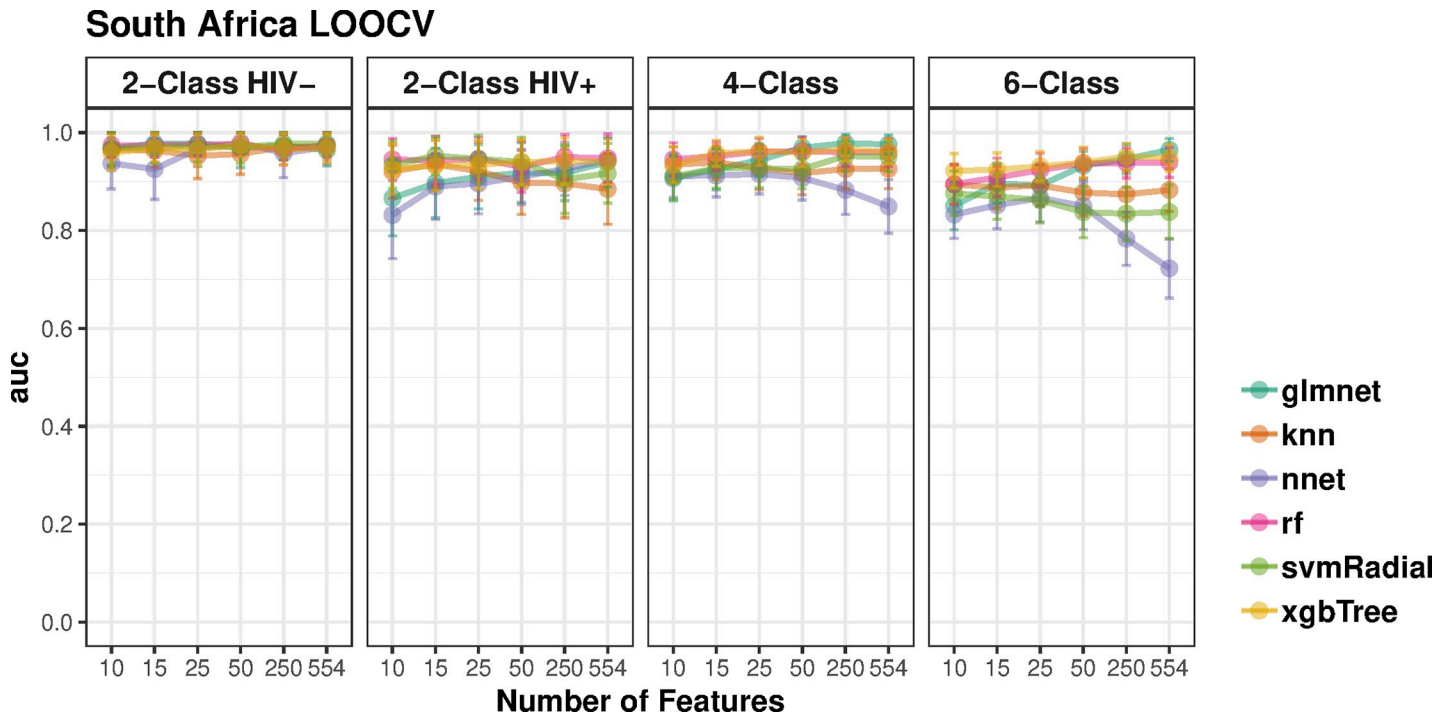

Fig 2. Training cross-validation results on adult TB samples. Leave one out (LOOCV) areas under the receiver operating curves (AUCs) for models on the South Africa adult data. Each panel plots the AUC curve for six machine learning algorithms (glmnet: Elastic-Net logistic regression, knn: k-Nearest Neighbors, nnet: Neural Network, rf: Random Forest, svmRadial, Support Vector Machine with Radial Basis Function kernel; xgbTree: Extreme Gradient Boosting) starting with models trained using all 554 probes, and iteratively shrunk to models trained on 10 probes only. Models were trained to classify the data into 6 (TB:HIV+, TB:HIV-, LTB:HIV+, LTB:HIV-, OD:HIV+, OD:HIV-), 4 (TB:HIV+, TB:HIV-, LTB:HIV+, LTB:HIV-) and 2 (TB, LTB) classes. Two types of 2-class models were trained: using either HIV+ or HIV- samples. Error bars show bootstrap-estimated 95% confidence intervals around the AUC.

https://doi.org/10.1371/journal.pone.0219322.g002

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.

Fig 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-reported TB signatures: our 16-gene correlate of TB risk [24], termed here as the 'ACS' signature, and the three-gene multi-cohort diagnostic signature developed by Sweeney et al [23], termed here as the 'threeGene' signature. The multinomial six-class Random Forest model outperformed all other models (AUC: 0.88,

Table 3. Training cross-validation results. Leave one out cross-validation (LOOCV) Areas under the receiver operating curves (AUC) and 95% confidence intervals (CI) for the best 10-probe model for each combination of machine-learning algorithm and number of predicted classes for models evaluated on the South Africa adult data (Fig 2).

	4-Class	6-Class	2-Class HIV-	2-Class HIV+
glmnet	0.91(0.86–0.95)	0.85(0.8–0.9)	0.97(0.93–1)	0.87(0.79–0.94)
knn	0.93(0.9–0.97)	0.89(0.85–0.93)	0.97(0.93–1)	0.92(0.87–0.97)
nnet	0.91(0.87–0.95)	0.83(0.78–0.88)	0.94(0.88–0.99)	0.83(0.74–0.92)
rf	0.95(0.91–0.98)	0.89(0.85–0.94)	0.97(0.95–1)	0.95(0.9–0.99)
svmRadial	0.91(0.86–0.96)	0.88(0.83–0.92)	0.97(0.93–1)	0.93(0.89–0.98)
xgbTree	0.93(0.9–0.97)	0.92(0.89–0.96)	0.96(0.92–1)	0.93(0.87–0.98)

https://doi.org/10.1371/journal.pone.0219322.t003

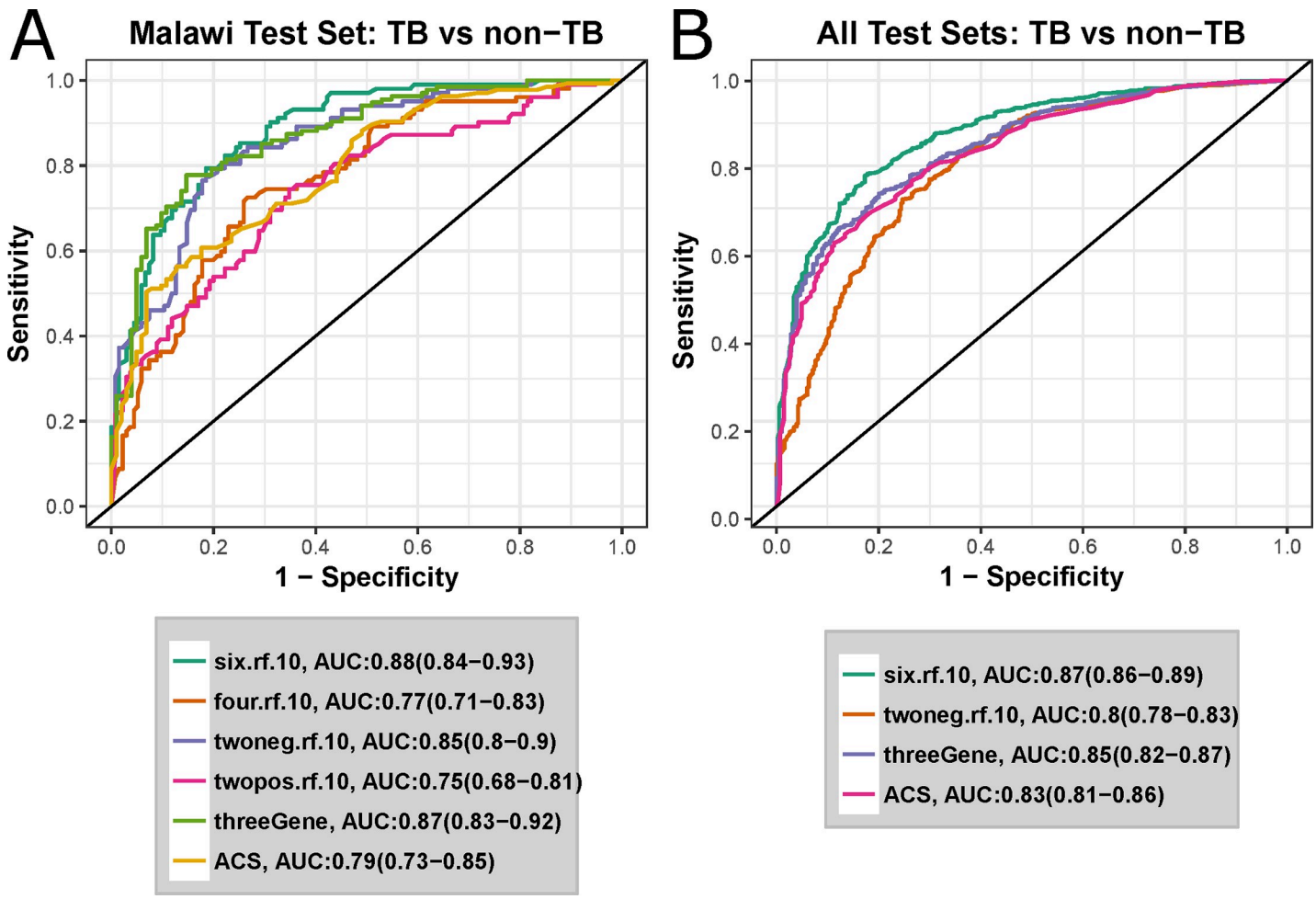

Fig 3. A six-class multinomial model optimally predicts 4 independent test sets. ROC curves for active TB vs non-TB classification of independent test sets. Legends shows the AUC for each model, with the 95% confidence intervals in parentheses. Models developed in this study are named in the form <number-of-classes>.<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 Sweeney et al [23], and ACS refers to the signature described by Zak et al [25]. 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 1.

https://doi.org/10.1371/journal.pone.0219322.g003

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] (Table 1). The two-class HIV- model was also included as a comparator for the six-class multinomial model. Fig 3(B) and S1 Fig show 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 was significantly better predictive performance than the top external model, the threeGene signature ($p = 0.006$ by a single tailed DeLong [29] test). Thus, multinomial modelling of TB disease states significantly improved the accuracy of discrimination of TB vs. non-TB samples.

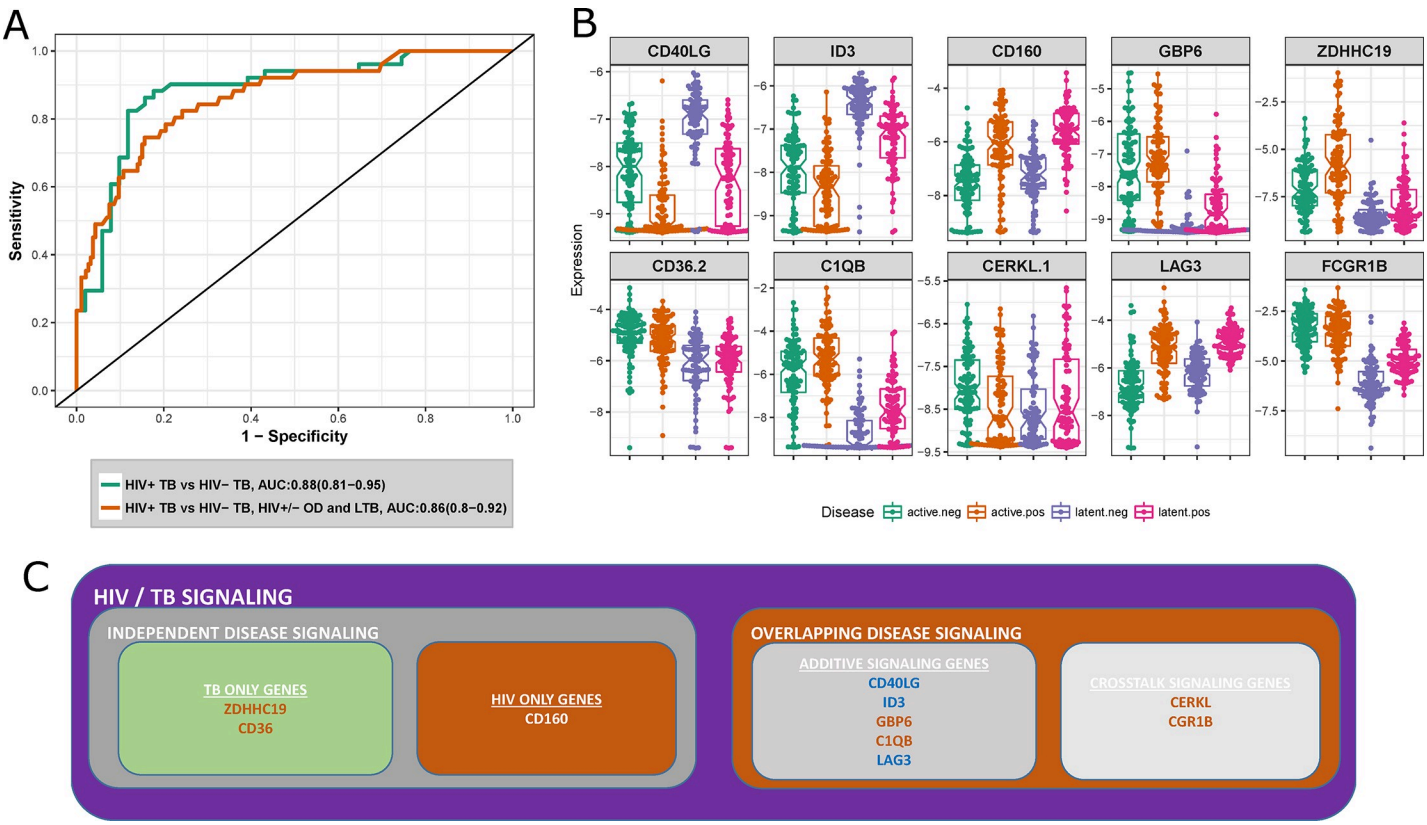

Fig 4. The six-class multinomial model identifies HIV+ TB as a distinct state. A ROC curves for the 10-gene six-class multinomial model discriminating HIV+ active TB samples from HIV- active TB samples, and HIV+ active TB samples from HIV- active TB and HIV+/- LTB and HIV+/- OD samples in the Malawi test set. **B** Dot- and boxplots of expression levels of six-class multinomial model genes in the entire Kaforou dataset. **C** Six-class multinomial genes classified by their TB/HIV behavior as determined by fitting linear models to gene expression as a function of disease state. TB upregulated genes are indicated in orange and downregulated genes shown in blue.

https://doi.org/10.1371/journal.pone.0219322.g004

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 (Fig 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 normalized 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 revealed a diverse pattern of transcriptional responses (Fig 4(B)).

Using HIV- latent TB samples as a baseline, 8 of the 10 signature genes were either downregulated in both active TB and HIV+ patients or upregulated in both active TB and HIV + patients. Two exceptions were 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 (Fig 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 exhibited 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 was upregulated in HIV+ individuals but downregulated in active TB. Two genes exhibited a significant interaction term: FCGR1B and CERKL. This interaction suggests crosstalk between the TB and HIV transcriptional response. In the case of FCGR1B, transcription reached 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 [30], exhibited 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 hypothesized 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 (S2 Table, Fig 5). Mapping these genes to blood transcriptional genes sets [27,28] revealed significant enrichment for cytotoxic T-cell and NK-cell pathways (S3 Table), 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 did not exhibit similarly strong correlations (Spearman $\rho > 0.8$) with any individual gene. At a more permissive correlation threshold ($\rho > 0.6$), CERKL was correlated with 9 genes (S4 Table), but this set of genes was not significantly enriched for any gene set. Genes most strongly correlated with CERKL were 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 Fig 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

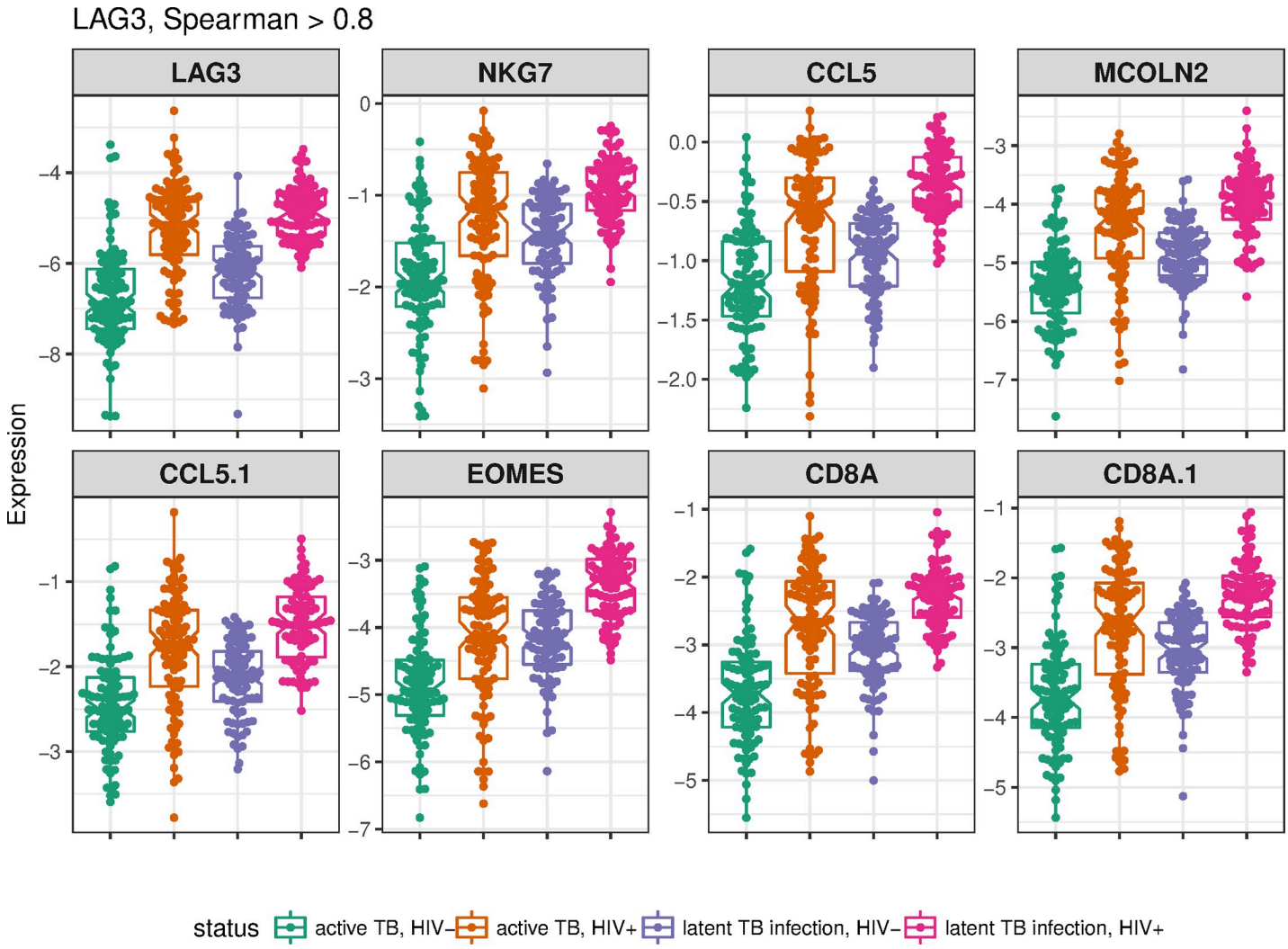

Fig 5. LAG3-correlated genes. Dot and boxplots for each microarray primer, named as the corresponding gene, strongly correlated with LAG3 (spearman correlation $\rho > 0.8$) for latent and active TB samples from the Kaforou dataset.

https://doi.org/10.1371/journal.pone.0219322.g005

cohort. By combining cohorts, signatures with a stable "global" operating score are revealed. Thus, it can be seen in Fig 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. Characterization 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 (Fig 4(B)). Upregulation of LAG3 is known to suppress T-cell activity in chronic HIV infection [31], 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 [32]. 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 was 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 [30]. 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 signaling 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.

Supporting information

S1 Fig. ROC curves for the 10-gene six-class multinomial model discriminating HIV + active TB samples from HIV- active TB samples, and HIV+ active TB samples from

HIV- active TB and HIV+/- LTB and HIV+/- OD samples in each independent test set listed in Table 1.
(PDF)

S1 File. Candidate probes. Illumina 12v4 microarray probe names selected for the initial data normalization and model cross-validation.
(XLSX)

S2 File. Guide to model naming structure.
(DOCX)

S1 Table. Genes in the six.rf.10 model. Gene names and descriptions for the best performing 10-probe 6-class multinomial random forest model.
(CSV)

S2 Table. LAG3 correlated genes. Illumina probe IDs with accompanying gene names strongly correlated with LAG3 in the combined South Africa and Malawi (GSE37250) data-sets.
(CSV)

S3 Table. Gene-set enrichment for LAG3 correlated genes. Gene-sets significantly enriched in LAG3 correlated genes, as determined by the hypergeometric test.
(CSV)

S4 Table. CERKL correlated genes. Illumina probe IDs with accompanying gene names strongly correlated with CERKL in the combined South Africa and Malawi (GSE37250) data-sets.
(CSV)

Author Contributions

Conceptualization: Fergal J. Duffy, Ethan G. Thompson, Thomas J. Scriba, Daniel E. Zak.

Data curation: Fergal J. Duffy.

Formal analysis: Fergal J. Duffy, Ethan G. Thompson, Daniel E. Zak.

Funding acquisition: Thomas J. Scriba, Daniel E. Zak.

Investigation: Fergal J. Duffy.

Methodology: Fergal J. Duffy, Thomas J. Scriba, Daniel E. Zak.

Software: Fergal J. Duffy.

Supervision: Ethan G. Thompson, Thomas J. Scriba, Daniel E. Zak.

Visualization: Fergal J. Duffy.

Writing – original draft: Fergal J. Duffy.

Writing – review & editing: Fergal J. Duffy, Ethan G. Thompson, Thomas J. Scriba, Daniel E. Zak.

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. https://doi.org/10.1371/journal.pmed.1002152 PMID: 27780211

2. World Health Organization. Tuberculosis Fact Sheet No 104 [Internet]. 2016. Available: http://www.who.int/mediacentre/factsheets/fs104/en/#
3. Gardiner JL, Karp CL. Transformative tools for tackling tuberculosis. J Exp Med. 2015; 212: 1759–1769. https://doi.org/10.1084/jem.20151468 PMID: 26458772
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. Elsevier Ltd; 2010; 10: 489–498. https://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: 191–194. https://doi.org/10.1016/0962-8479(93)90010-U PMID: 8369514
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. Elsevier Ltd; 2011; 377: 1495–1505. https://doi.org/10.1016/S0140-6736(11)60438-8 PMID: 21507477
7. Lawn SD, Mwaba P, Bates M, Piatek A, Alexander H, Marais BJ, et al. Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. Lancet Infect Dis. Elsevier Ltd; 2013; 13: 349–361. https://doi.org/10.1016/S1473-3099(13)70008-2 PMID: 23531388
8. Jensen P a, Lambert L a, Iademarco MF, Ridzon R. Guidelines for preventing the transmission of Mycobacterium tuberculosis in health-care settings, 2005. MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports / Centers for Disease Control. 2005. https://doi.org/10.2307/42000931
9. Kaforou M, Wright VJ, Oni T, French N, Anderson ST, Bangani N, et al. Detection of Tuberculosis in HIV-Infected and -Uninfected African Adults Using Whole Blood RNA Expression Signatures: A Case-Control Study. PLoS Med. 2013; 10. https://doi.org/10.1371/journal.pmed.1001538 PMID: 24167453
10. Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluka G, et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. N Engl J Med. 2014; 370: 1712–23. https://doi.org/10.1056/NEJMoa1303657 PMID: 24785206
11. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch S a a, Oni T, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature. Nature Publishing Group; 2010; 466: 973–977. https://doi.org/10.1038/nature09247 PMID: 20725040
12. Bloom CI, Graham CM, Berry MPR, Rozakeas F, Redford PS, Wang Y, et al. Transcriptional Blood Signatures Distinguish Pulmonary Tuberculosis, Pulmonary Sarcoidosis, Pneumonias and Lung Cancers. PLoS One. 2013; 8. https://doi.org/10.1371/journal.pone.0070630 PMID: 23940611
13. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available: https://www.r-project.org/
14. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008; 28: 1–26. https://doi.org/10.18637/jss.v028.i07
15. Liaw A, Wiener M. Classification and Regression by randomForest. R news. 2002; 2: 18–22. Available: http://cran.r-project.org/doc/Rnews/
16. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab–An S4 Package for Kernel Methods in R. J Stat Softw. 2004; 11: 1–20. https://doi.org/10.1016/j.csda.2009.09.023
17. Venables WN, Ripley BD. Modern Applied Statistics with S. Issues of Accuracy and Scale. 2002; 868. https://doi.org/10.1198/tech.2003.s33
18. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33: 1–22. https://doi.org/10.1359/JBMR.0301229 PMID: 20808728
19. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. xgboost: Extreme Gradient Boosting [Internet]. 2019. Available: https://cran.r-project.org/package=xgboost
20. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7: 3. https://doi.org/10.1186/1471-2105-7-3 PMID: 16398926
21. Garson DG. Interpreting neural-network connection weights. Artif Intell Expert. 1991; 6: 46–51.
22. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12: 77. https://doi.org/10.1186/1471-2105-12-77 PMID: 21414208
23. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. Lancet Respir Med. Elsevier Ltd; 2016; 2600: 1–12. https://doi.org/10.1016/S2213-2600(16)00048-5

24. Zak DE, Penn-Nicholson A, Scriba TJ, Thompson E, Suliman S, Amon LM, et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. Lancet. Elsevier Ltd; 2016; 6736: 1–11. https://doi.org/10.1016/S0140-6736(15)01316-1

25. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43: e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

26. Weiner J. tmod: Feature Set Enrichment Analysis for Metabolomics and Transcriptomics [Internet]. 2016. Available: https://cran.r-project.org/package=tmod

27. Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. Nat Immunol. 2014; 15: 195–204. https://doi.org/10.1038/ni.2789 PMID: 24336226

28. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. Immunity. 2008; 29: 150–164. https://doi.org/10.1016/j.immuni.2008.05.012 PMID: 18631455

29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44: 837–845. https://doi.org/10.2307/2531595 PMID: 3203132

30. Tuson M, Garanto A, Gonzàlez-Duarte R, Marfany G. Overexpression of CERKL, a gene responsible for retinitis pigmentosa in humans, protects cells from apoptosis induced by oxidative stress. Mol Vis. 2009; 15: 168–80. Available: http://www.ncbi.nlm.nih.gov/pubmed/19158957 PMID: 19158957

31. Tian X, Zhang A, Qiu C, Wang W, Yang Y, Qiu C, et al. The upregulation of LAG-3 on T cells defines a subpopulation with functional exhaustion and correlates with disease progression in HIV-infected subjects. J Immunol. 2015; 194: 3873–82. https://doi.org/10.4049/jimmunol.1402176 PMID: 25780040

32. Jayaraman P, Jacques MK, Zhu C, Steblenko KM, Stowell BL, Madi A, et al. TIM3 Mediates T Cell Exhaustion during Mycobacterium tuberculosis Infection. PLoS Pathog. 2016; 12: e1005490. https://doi.org/10.1371/journal.ppat.1005490 PMID: 26967901