Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

Toward a Unified Biosignature for Tuberculosis

Jeroen Maertzdorf, Stefan H.E. Kaufmann, and January Weiner III

Max Planck Institute for Infection Biology, Department of Immunology, 10117 Berlin, Germany

Correspondence: maertzdorf@mpiib-berlin.mpg.de

Accurate and rapid diagnosis of active tuberculosis (TB) disease is still hampered by inadequate tools. Although current assays relying on single-marker readouts mostly display inadequate sensitivity and/or specificity, host-related multimarker signatures are especially poorly developed. As a consequence, research programs have been initiated to search for combinations of markers—so-called biosignatures with superior performance. Many such investigations harness high-throughput platforms to analyze the host response during infection and disease. A major challenge for these activities is the analysis of vast amounts of data produced. Specialized bioinformatic tools are being applied to identify the most robust biosignatures for classification of exposed and diseased individuals and prognosis of risk of disease in endemic areas. Validation of the most promising biosignatures in ongoing multicohort studies will bring us a step closer to the identification of an accurate unified signature.

Whereas most current diagnostic assays identify the presence of the pathogen in the host, host-related biomarkers hold promise for use in prognostic research and vaccine trials and for monitoring treatment responses. Over the last decade, host-related biomarker research in tuberculosis (TB) has witnessed a clear shift from investigations on single markers to high-throughput studies revealing signatures consisting of multiple integrated markers (Weiner et al. 2013). On the one hand, such high-throughput profiling has become feasible by advanced modern technologies. On the other hand, researchers in the field have realized that single markers usually lack the sensitivity and/or specificity of a true correlate that can successfully be implemented in a clinical setting. Biosignatures are believed to hold higher promise for clinical application in areas such as pathology and immunology and to be suitable to discriminate etiologically different diseases with similar clinical features in a single assay (Weiner et al. 2013). These apparent improvements of multimarker biosignatures, however, suffer from the drawback of complex readout and interpretation. Whereas single-marker assays can be easily interpreted and evaluated, studies relying on high-throughput platforms are faced with the challenge of highly dimensional data sets, thereby requiring more sophisticated bioinformatics tools and more in-depth interpretation skills.

HIGH-THROUGHPUT BIOMARKER PLATFORMS

High-throughput platforms currently used in TB biomarker research comprise (1) proteo-

Editors: Stefan H.E. Kaufmann, Eric J. Rubin, and Alimuddin Zumla
Additional Perspectives on Tuberculosis available at www.perspectivesinmedicine.org

 doi: 10.1101/cshperspect.a018531
Cite this article as *Cold Spring Harb Perspect Med* 2015;5:a018531

mics, assessing the total composition of proteins and peptides in a given biological sample; (2) metabolomics, which measures small molecules involved in, or an end product of, basic metabolic processes taking place in biological systems; and (3) transcriptomics, which covers the expression of genes in relation to disease or in response to a pathogenic insult. Transcriptomics has become the most widely applied experimental readout in TB biomarker research. Microarray platforms have thus far been the method of choice to simultaneously quantify the transcriptional activity of all genes in host cells. This approach has revealed a multitude of potential biosignatures in TB, which can potentially be integrated into clinical settings to identify active TB disease in suspected cases (Maertzdorf et al. 2012a; Joosten et al. 2013). More recent studies also attempt to combine the transcriptional activity of both human cells and the pathogen to investigate host–pathogen interactions in more detail.

One drawback of microarray technologies is the lack of absolute and detailed evaluation of gene expression. Modern deep sequencing technologies provide quantitative and qualitative information on gene expression and genomic composition down to the single-nucleotide level (Normand and Yanai 2013). Because of the high cost and time-consuming process of such deep sequencing analyses, this technology is currently mostly restricted to small-scale exploratory studies, in which evaluation and validation of identified biomarkers are assessed by more standard methods.

Although high-throughput platforms currently in use have their own strengths and weaknesses, they share the inevitable need for complex analysis of the huge amount of data generated.

A prerequisite of successfully applying high-throughput platforms is the high quality of the biological samples. In the case of TB, most studies rely on peripheral blood as the source for transcriptional, metabolic, and proteomic profiling. Although the immunological battle against the invading pathogen is focused at the site of infection and pathogen persistence (i.e., the lung), this pathologic status is reflected in the peripheral blood by circulating immune cells (Weiner et al. 2013). For studies focused on proteomic and genomic analysis of blood cells, the complexity of such biological samples poses an additional challenge. The many different types of cells present in the blood and their varying composition between individuals add an additional level of complexity to the already staggering number of readout parameters being assessed (Jacobsen et al. 2006). This biological variation between individuals, and between different ethnic groups, renders validation studies in distinct cohorts an essential requirement to assess the potential clinical validity of biomarkers and biosignatures.

SINGLE BIOMARKERS IN TB DIAGNOSTICS—A MATTER OF SENSITIVITY AND SPECIFICITY

Among the major challenges in using single canonical biomarkers in TB diagnostics is the issue of sensitivity and specificity. Generally, current clinical diagnosis applies a combination of "old" and "new" biomarkers. One of the most broadly used assays is the sputum smear test, which detects the presence of acid-fast bacteria in sputum of (suspected) TB cases (Norbis et al. 2013). This sputum smear test has the benefits of being low cost and easy to handle and of the fact that a positive result provides certainty about the presence of mycobacteria. However, there are several important drawbacks to this assay. First and foremost of all, this test often results in false-negative results, thereby missing a substantial number of patients with active pulmonary TB (Harries et al. 1998). Second, the assay does not distinguish between live and dead bacteria, and thus it cannot detect active bacterial replication as a hallmark of active TB disease, which is of importance for monitoring treatment responses. Thus, although the sputum smear test has a high specificity, its overall sensitivity is low (Davies and Pai 2008).

A more sensitive and widely applied diagnostic test to detect the presence of live bacilli is culture from sputum. Although a positive culture result directly proves the presence of live and replicating mycobacteria in the lungs of a

suspected TB patient, the slow growth characteristics of *Mycobaterium tuberculosis* (*Mtb*) make the culture assay a lengthy procedure lasting up to several weeks. Although new technologies such as liquid culture have brought this down to several days, it still is a long time for rapid patient management (Norbis et al. 2013).

The more recently developed GeneXpert assay is designed to detect mycobacteria in sputum of suspected cases. Instead of staining, this assay uses the much more sensitive (quantitative) polymerase chain reaction (PCR) to detect specific sequences in the bacterial genome. Although much more sensitive than the smear test, it too does not distinguish between live and dead bacteria. The GeneXpert assay has very high specificity and reasonably good sensitivity (Steingart et al. 2013) and has been endorsed by the World Health Organization. It is relatively costly (making routine screening in clinical settings unaffordable for countries with low financial resources) and requires special equipment operating on a constant and reliable power supply.

Often performed in combination with sputum smear or GeneXpert assays is the examination of chest X rays. Principally, lung lesions and inflammatory tissue are readily recognized in patients with active TB (Walzl et al. 2011). However, X ray does not detect active bacterial replication and may produce false-negative results, especially in early stages of TB disease development. The specificity of chest X rays is restricted by the inability to identify the causative agent or processes underlying pulmonary pathology and lung damage. Therefore, chest X ray is inconclusive for diagnosis of TB (Al-Zamel 2009).

Another group of clinical assays, targeted at host responses, rely on biomarkers of immune activation and memory to mycobacterial antigen(s). The widely used tuberculin skin test (TST) assesses local immunological responses elicited by mycobacterial antigens (Bekmurzayeva et al. 2013) and is primarily used to show prior sensitization and latent infection. Subcutaneous injection of an extract of heat-killed *Mtb* proteins evokes a delayed-type hypersensitivity (DTH) reaction, causing transient swelling and reddening of the skin. Although of low cost, the test has profound limitations for detecting latent infections both with respect to sensitivity and specificity, which range widely depending on the population tested and can be affected by HIV infection and Bacillus Calmette–Guérin (BCG) vaccination (Bekmurzayeva et al. 2013).

The more recent approach in this group of diagnostic tests, the interferon-γ release assays (IGRAs), such as the Quantiferon test, rely on ex vivo activation of T lymphocytes secreting interferon-γ (IFN-γ) on stimulation with defined *Mtb*-specific proteins (Bekmurzayeva et al. 2013). Although a positive test result indicates previous exposure of a person's cellular immune response to *Mtb*, it is not a measure of bacterial activity or even persistence of the bacillus. As such, failure to discriminate latent *Mtb* infection (LTBI) from active TB is a major drawback of IGRAs (Bekmurzayeva et al. 2013; Norbis et al. 2013). IGRAs have a sensitivity equal to the TST but are more specific. However, particularly in highly TB-endemic regions, IGRAs are unsuitable for diagnosing active TB disease (Walzl et al. 2011). Some individual cases under clinical suspicion of having TB can also present with negative results, which in turn increases the risk of false exclusion of TB (Mazurek et al. 2007). The test may also be impaired by T-cell anergy in HIV^+ individuals (Barth et al. 2008).

The lack of sensitivity and specificity of current diagnostic tools illustrate that a single marker is unlikely to fulfill the requirements for reliable discrimination between active TB and other pulmonary or related infections. As a consequence, researchers have started searching for combinations of biomarkers, often referred to as biosignatures, to tackle the issue of sensitivity and specificity in TB diagnosis.

GENETIC BIOSIGNATURES IN TB DISEASE

One of the most pronounced genetic biosignatures present in the blood of TB patients is the elevated expression of transcripts involved in IFN signaling. First described in detail by Berry et al. (2010), this signature, consisting of both IFN type I and type II, was shown to be mainly driven by neutrophils. Albeit a prominent sig-

nature in TB disease, an IFN gene expression profile has also been observed in other chronic inflammatory conditions like systemic lupus erythematosus (SLE) and sarcoidosis (Koth et al. 2011; Maertzdorf et al. 2012b; Bloom et al. 2013). As such, an IFN signature in itself is insufficient for a definitive identification of TB, but it may be implemented for the identification of a more generic underlying inflammatory condition. Likely, it will also be of value for the definition of predictive biomarkers (e.g., for progression of LTBI toward active TB disease before actual clinical signs are observed).

Another pronounced transcript signature, which has been described in several studies, is the increased activity of the Fc γ receptor signaling pathways. First described as an increased expression of CD64 (another alias for FCGR1A) in TB patients (Jacobsen et al. 2007) and later confirmed by others (Maertzdorf et al. 2011a; Joosten et al. 2012), its expression was shown to be strongly correlated with other inflammatory markers with a central role for the JAK-STAT pathway (Maertzdorf et al. 2011b). This feature is also illustrated by its connection to IFN-inducible genes and its association with Toll-like receptors (TLRs), guanylate-binding proteins (GBPs), and components of the complement system (Maertzdorf et al. 2011a). Key regulators within the JAK-STAT pathway are cognates of the suppressor of cytokine signaling (SOCS) family. Dysregulated expression of SOCS3 in T lymphocytes from TB patients has also been suggested to play a role in susceptibility to TB (Jacobsen et al. 2011).

Activation of the complement system, as indicated by increased gene expression of multiple factors herein, may, on the one hand, reflect an innate defense mechanism against *Mtb*. On the other hand, its chronic activation may significantly contribute to inflammation-mediated tissue damage (Welsh et al. 2012), contributing to the pathogenesis of the disease.

Involvement of the complement system is also observed in autoimmune diseases like SLE (Ballanti et al. 2013), in which the disease is not triggered by a pathogen, but rather by a dysregulated chronic activation of the complement system. Complement activation in TB and similarities with SLE are also illustrated by the high enrichment of TB-induced genes in signatures associated with SLE (Maertzdorf et al. 2011a).

A recent publication was the first to combine information of all genetic biosignatures in blood of TB patients described so far (Joosten et al. 2012). As the diverse platforms used in transcriptome analysis for TB biomarkers by various authors are not directly comparable with canonical analytical methods, the authors have analyzed a "superset" of transcripts, which have been found in the previous experimental studies. This resulted in improved sensitivity, allowing identification of significant enrichment in pathways and groups of genes which were until now not linked to a TB signature. While confirming the presence of most described signatures in multiple studies, their study identified a hitherto underappreciated role of the B-cell compartment. The results also point to an involvement of TREM1 signaling in active TB disease, something that had not been described in any of the previous studies separately (see also Walzl et al. 2014). Shortly after, Cliff et al. (2013) also identified significant changes in expression of a network of B-cell-related genes during disease resolution.

The major biological pathways active in TB can be clearly grouped into three dominant clusters—that is, (1) signaling and inflammation pathways, (2) IFN and immune pathways, and (3) T- and B-cell pathways—based on a set of 1446 differentially expressed transcripts, as recently shown by Bloom et al. (2013). First described by Maertzdorf et al. (2012b), this study revealed similarities, but also significant differences in blood gene expression signatures between TB and sarcoidosis. Moreover, Bloom et al. also included pneumonia and lung cancer in their study, indicating that, whereas TB and sarcoidosis share the dominant IFN signatures, the other two pulmonary diseases show distinct signatures dominated by inflammatory genes.

Although the above-described biological signatures reflecting TB pathology have been identified in multiple studies, their unique specificity for TB is low and many signatures are shared with other inflammatory diseases. These

features provide important insights into the biology underlying TB pathology; yet they are of lesser value for clinical applications such as distinguishing TB patients from healthy individuals or from patients suffering from other diseases. Several studies have therefore tried to identify gene expression signatures that can specifically accomplish such distinctions, regardless of their biological relevance. The number of gene transcripts in such signatures range from three (Jacobsen et al. 2007) to hundreds (Berry et al. 2010; Bloom et al. 2013). Although the sensitivity and specificity of these signatures in distinguishing TB patients from healthy individuals are quite robust in most studies, they also seem to contain cohort-specific information at least to some degree. This is, for example, illustrated by the study by Bloom et al. (2013), applying signatures from two other studies (Koth et al. 2011; Maertzdorf et al. 2012b) onto their own data set led to a significant reduction in sensitivity in class predictions (Bloom et al. 2013). A global analysis combining genetic signatures in numerous cohorts of TB patients and controls from various TB-endemic areas around the world would allow the identification of a unified and general signature.

METABOLIC PROFILING BIOMARKERS IN TB

Although gene expression biomarkers from whole blood RNA have been widely applied to characterize differences between TB patients and control groups, other avenues have been taken as well. Assuming that RNA expression from whole blood predominantly reflects changes in peripheral blood cells, additional information can be gained by measuring molecules that are derived directly from the site of infection (Maertzdorf et al. 2012a). Weiner et al. (2012) analyzed metabolic profiles in serum samples of TB patients and healthy controls and showed that a biosignature based on metabolic profiles can achieve sensitivity and specificity, at least on par with transcription profiles (Weiner et al. 2012). Some of the predictive metabolites were directly derived from the site of infection. Kynurenine, which was found at significantly higher levels in the sera of TB patients than in controls, is synthesized from tryptophan by the enzyme IDO1, induced in macrophages and dendritic cells in contact with *Mtb*. It is likely that high kynurenine abundance in serum reflects release of this molecule from granulomas.

Among groups of metabolites that were found at different abundance in TB patients and healthy controls, distinct amino acids (e.g., histidine and cysteine), bile acids and uremic toxins (such as taurocholate and glycocholenate), *N*-acetylneuraminate, cortisol, fibrinopeptides, and lysophosphatidylcholines were most prevalent (Weiner et al. 2012). Intriguingly, several processes comprising changes in their metabolic components (including higher levels of kynurenine and cortisol and lower abundance of lysophosphatidilcholines, which may be caused by inhibition of phospholipase A2) can be linked to immunosuppressive mechanisms.

In a parallel approach, cytokines were determined in the same study groups. Distinct cytokines correlated with defined metabolites, showing the link between metabolite abundance in serum and the immune processes in response to TB disease. For example, the chemokine C-X-C motif chemokine 10 (CXCL10, IP-10), the cytokine interleukin 6, and the growth factor granulocyte colony-stimulating factor (GCSF) showed a negative correlation with abundances of several amino acids (for example, glutamine and tryptophan) and positive correlation with metabolic markers elevated in sera of TB patients, including *N*-acetylneuraminate and hypoxanthine.

VARIABILITY OF BIOMARKERS IN TB PATIENTS

Both transcriptomics and metabolomics revealed that the variance of the variables was higher in TB patients than in healthy subjects (J. Weiner, unpubl.). This becomes most apparent when visualization techniques such as Chernoff faces or spider plots are applied—the corresponding features are more profound in patients than in healthy individuals. A proximate explanation for this phenomenon is that

the disease induces higher expression of several genes, and expression strength may correlate with variance of the measured expression signal. However, differential variability of gene expression (Ho et al. 2008) has been found in immune-modulated, stress-induced, and hormonally regulated genes in mice (Pritchard et al. 2001).

Some other chronic infectious diseases are characterized by a spectrum of clinical manifestations. Notably, in leprosy, a disease caused by *Mycobacterium leprae*, clinical manifestations range from tuberculoid leprosy to lepromatous leprosy. Furthermore, despite the fact that TB is often depicted as having two stages (latent and active TB), the reality is more complex (Barry et al. 2009), and the spectrum of clinical TB disease includes a diversity of manifestations and stages that falls between these two extremes and depends on a balance between pro- and anti-inflammatory responses (Lin and Flynn 2010). Moreover, even within an individual host, the granulomatous lesions show a remarkable diversity both in pathology and in response to treatment (Lin et al. 2014). Thus, the host response to TB is variable, and consequently individual variability is reflected by gene expression and metabolic profiles. This merits a more detailed analysis.

BIOSIGNATURES AND MACHINE LEARNING

Biosignatures are often presented as lists of genes that distinguish between groups of interest (e.g., between TB patients, patients with other diseases, and healthy controls). Generally, these genes are differentially expressed between groups. However, the reverse need not be true: Generating a biomarker set or a multivariate biosignature is not limited to selecting genes that differ most profoundly between conditions. A biosignature must be capable of distinguishing not only the samples it was derived from (the *training set*), but more importantly, any novel sample set that it will be confronted with. For example, if one has selected a number of genes that differentiate between the conditions in a given sample set, one needs to test whether these genes distinguish between conditions in another, independent set (the *validation set*). In essence, this is an application of supervised machine learning (ML).

An ML model is defined not only by a list of genes, metabolites, or other predictive variables, but also by the specific values of these variables and model parameters, such as weights which determine how each variable influences the model. Kaforou et al. (2013) analyzed a substantial number of samples from 12 different groups, from two cohorts (Malawi and South Africa, see Table 1), and three disease states (healthy control subjects, TB patients, and patients from a disease other than TB, which includes TB in differential diagnosis) (Kaforou et al. 2013). Furthermore, the study included HIV^- and HIV^+ individuals. The authors developed a simple, yet effective ML algorithm. Based on the training set, a list of less than 50 genes was compiled to differentiate between distinct disease states, and each gene was assigned a weight of $+1$ or -1 depending on whether it was more highly expressed in TB ($+1$) or in the control group (-1). The expression values from the selected list were multiplied by the weights and then summed and compared with a threshold calculated based on the levels of gene expression in the training set. Thus, the actual biosignature consisted of (1) a list of genes, (2) the assigned weights, and (3) a parameter (disease risk score [DRS] threshold) based on expression levels from the training set. In more complex ML algorithms the number of parameters required can be even higher. Notably, this same signature may not necessarily be successful if applied to another study based on an unrelated cohort and an independent readout platform.

MACHINE LEARNING ALGORITHMS

Several supervised ML algorithms exist, of which only a few have been exploited for TB data set analyses thus far (Table 1, see Method column).

k-*nearest neighbors* (KNN) (related to an unsupervised clustering method, *k*-means clustering) has been successfully applied by Berry et al. (2010). In KNN, the variates are treated as coordinates in an *N*-dimensional space,

Table 1. List of studies of human whole blood transcriptome in the context of active TB

Cohort	Study groups (sample size)	Classification error measures and error rates	Method	Study
United Kingdom (training set)	TB (13), LTBI (17), CTRL (12)	All results TB versus non-TB: sensitivity, 91.7%; specificity, 96.6%	*k*-nearest neighbor	Berry et al. 2010[a]
United Kingdom (test set)	TB (21), LTBI (21), CTRL (12)	Sensitivity, 61.7%; specificity, 93.8%; indeterminate, 1.9%	*k*-nearest neighbor	Berry et al. 2010
South Africa (validation set)	TB (20), LTBI (31)	Sensitivity, 94.1%; specificity, 96.7%; indeterminate, 7.8%	*k*-nearest neighbor	Berry et al. 2010
South Africa	TB (33), LTBI (34), CTRL (9)	TB versus LTBI using a four-gene signature: sensitivity, 94%; specificity, 97%	Random forest with LOO cross-validation	Maertzdorf et al. 2011b
Gambia	TB (46), LTBI (25), CTRL (37)	TB versus non-TB[e]: accuracy, 82%; sensitivity, 74%; specificity, 87%	Random forest with LOO cross-validation	Maertzdorf et al. 2011a
United States	CTRL (12)	Accuracy, 100%	PAM-R	Lesho et al. 2011
Brazil	LTBI (6), TB (5)	Accuracy, 100%	PAM-R	
Germany	TB(8), sarcoidosis (18), CTRL (18)	Three-way classification error rates: sarcoidosis, 5.6%; TB, 12.5%; CTRL, 0% (overall error rate, 4.55%)	Random forest with LOO cross-validation	Maertzdorf et al. 2012b
United Kingdom; France (training set)	TB (16), sarcoidosis (25), pneumonia (8), lung cancer (8), CTRL (38)	TB versus non-TB: sensitivity, 88%; specificity, 94%	Support vector machines	Bloom et al. 2013
United Kingdom; France (test set)	TB(11), sarcoidosis (25), pneumonia (6), lung cancer (8), CTRL (52)	Sensitivity, 82%; specificity, 91%	Support vector machines	Bloom et al. 2013
United Kingdom; France (validation set)	TB (8), sarcoidosis (11), CTRL (23)	Sensitivity, 88%; specificity, 92%	Support vector machines	Bloom et al. 2013
South Africa	TB/HIV$^+$ (49), TB (47), LTBI HIV$^+$ (48), LTBI/ (50), OD/HIV$^+$ (68), OD/(49)	TB versus LTBI: sensitivity, 95%; specificity, 90%. TB versus OD: sensitivity, 93%; specificity, 88%	DRS	Kaforou et al. 2013
Northern Malawi	TB/HIV$^+$ (60), TB/ (59), LTBI HIV$^+$ (41), LTBI/ (36), OD/HIV$^+$ (38), OD/(39)	TB versus LTBI: sensitivity, 95%; specificity, 90%. TB versus OD: sensitivity, 93%; specificity, 88%	DRS	Kaforou et al. 2013
South Africa	TB (10), cured (10)[c], recurrent(10)[c], LTBI (10)	Accuracy, 93.75%	Stepwise linear discriminant analysis	Mistry et al. 2007
Indonesia	TB (23)[d], CTRL (23)	Not used for classification task		Ottenhoff et al. 2012
South Africa	LTBI (38), TB (29)	Not used for classification task		Bloom et al. 2012
United Kingdom	TB (8)[b]	Not used for classification task		

Continued

Table 1. *Continued*

Cohort	Study groups (sample size)	Classification error measures and error rates	Method	Study
South Africa (training set)	TB (18) at time of diagnosis, and 1, 2, 4, and 26 weeks postdiagnosis	Overall accuracy, 92.6%	Neuronal network	Cliff et al. 2013
South Africa (test set)	TB (9)			

Donors were HIV^- unless indicated otherwise.

CTRL, healthy controls; DRS, disease risk score; HIV^+, human immunodeficiency virus positive; LOO, leave one out; LTBI, latent tuberculosis infection; OD, other diseases; PAM-R, prediction analysis of microarrays with R; TB, tuberculosis.

[a] Data for the 393-transcript signature.

[b] Eight patients were followed up during TB treatment.

[c] Cured, cured patients who underwent one episode of TB; recurrent, cured patients who underwent two or three recurring episodes of TB; six individuals from each group were used as a training set, and the remaining four individuals were used as a test set.

[d] Samples from 23 patients were collected before, during, and after successful completion of the treatment.

[e] Data provided by the authors (J Maertzdorf, pers. comm.).

where *N* is the number of variables (e.g., genes, for which expression has been measured). For each group, its center is determined by averaging the position of all samples that belong to the group. Classification of new samples relies on determining the distance from each of the centers and choosing the center closest to the sample in question. A major advantage of this method is that decisions can be easily interpreted. The DSR algorithm described above and used by Kaforou et al. (2013) can be considered as a variant of this method, using Manhattan distances (in which the expression levels of single genes are summed) instead of Euclidean distances (in which the distance is measured as a square root of sum of squared differences in gene expression).

Random Forests (RF) ML algorithm is an ensemble-based algorithm closely related to another class of ML algorithms, decision trees. RF generates a large number of decision trees; when applied to a sample of the validation set, the decision how to classify this sample is based on the majority vote from the trees. RF has been successfully applied in several settings; in gene expression analysis, it compares favorably with other methods. A major advantage of RF is the fact that high accuracy can be obtained with only a few predictors (Diaz-Uriarte et al. 2006) and that accurate cross-validated error rates are provided "out of the box" (so-called OOB classification error rates).

Support vector machines (SVMs) are a type of generalized linear classifiers. Data points (samples) are mapped onto a multidimensional space in such a way that it is possible to calculate a hyperplane (a multidimensional plane), which separates the classes of samples. Whereas originally suitable only for two-class data, it can be extended to analyze also multiclass problems.

PERFORMANCE OF BIOSIGNATURES

Some 500 gene expression profiles have been collected from TB patients in different studies and the majority of these have been used in an ML context to generate biosignatures of TB (Table 1). Several groups have attempted to compare the signatures generated based on their data with the signatures based on lists of genes from other publications, with the overall outcome of achieving lower error rates when the signature was defined from the same data set than when a signature derived from another data set and a different algorithm were applied. Analyzing disease-specific biosignatures, Bloom et al. (2013) applied lists of genes corresponding to the biosignatures published by Maertz-

dorf et al. (2012b) and Koth et al. (2011). They found that the derived SVM models had a much lower performance when applied to their data set. Similarly, Kaforou et al. (2013) used the 393-gene list derived from Berry et al. (2010) and found that when applied to their method and their data, its performance was much lower than the models corresponding to the list of genes originally derived by Kaforou et al. (2013).

Some sets of genes can (and do) perform objectively better than others when used in a classification task *ceteris paribus*. Yet, lists of genes derived from a biosignature comprising specific parameters and applied to a specific data set generated with a specific microarray or RNAseq platform need not be compatible with other data sets and other algorithms. Gene lists produced by variable selection will differ even for the same data set when different algorithms are applied to the same data set. For a meaningful comparison, complete biosignatures, including the specific algorithms and parameters, such as weights or decision trees, should be directly applied to the same data set. Such a comparison has not yet been performed. Consequently, one of the major challenges awaiting the field of TB biomarkers is the unification of diverse platforms and data sets to achieve a platform-independent, global signature of TB.

A supervised ML algorithm is trained using a set of samples with known associations to groups (e.g., healthy controls and TB patients). Applied to a new set of samples, the algorithm should correctly classify these samples. In a simple setting with only two classifications (e.g., disease vs. healthy, or one selected class against all other classes), the errors of the algorithm correspond to type I and type II errors in a hypothesis-testing framework. Type I errors (false positives) arise when a "negative" sample (e.g., from a healthy patient) is classified as "positive" (e.g., TB disease). Type II errors (false negatives) arise in the opposite situation: A positive sample is classified as negative. Therefore, it is logical to describe the performance of an algorithm by its overall error rate (also termed accuracy) (Table 2).

For a single sample, the probabilities of type I and type II errors can be given. However, the actual number of false-positive and -negative predictions (e.g., in a clinical setting) also depends on the population demography. The actual numbers of incorrect predictions depend on the relative frequency of each group. An algorithm with sensitivity and specificity of 99% each will produce 10 times more false positives than false negatives if applied for detection of a disease with a 10% incidence in a population. This is known as the "accuracy paradox": If the prevalence of a group is sufficiently low, then an algorithm that uniformly predicts that a sample belongs to a majority group (e.g., healthy rather than TB disease) can have a higher accuracy (lower overall error rate).

More suitable measures of performance of an algorithm are available. Precision rate (also called positive predictive value, PPV) is the proportion of true positives among all samples that have been classified as positive. Another group of measures is derived from the receiver–operator characteristics (ROC) curves. If an ML algorithm is applied to a test set, the pre-

Table 2. Measures of machine learning algorithm performance

		Prediction		
		Negative	Positive	Accuracy measure
Reality	**Negative**	TN	FP, type I error	Specificity = TN/(TN + FP)
	Positive	FN, type II error	TP	Sensitivity = TP/(FN + TP)
	Accuracy measure	NPV = TN/(TN + FN)	PPV = TP/(FP + TP)	Accuracy = (TP + TN)/N =1 − overall error

FP, false positive; FN, false negative; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive.

dictions of that algorithm are sorted by their confidence to derive an ROC curve. The area under the curve is a test statistic with known characteristics; a set of random predictions will form a diagonal, and the greater the area, the better the predictions.

Attempts to estimate the predictive power of a whole set of transcript biomarkers in TB have reported strikingly similar results (Fig. 1; Joosten et al. 2013). The accuracies, specificities, and sensitivities frequently range between 85% and 95% (Table 1). Although two studies obtained no misclassifications in some comparisons, this would be attributable to the small sample size used: Given an error rate of 5% and a total sample size of 20, one expects on average only one incorrect classification. Intriguingly, a study focusing on metabolic profiles rather than transcriptomes also revealed 3%–5% overall errors (Weiner et al. 2012). This could indicate an upper limit for the performance of biomarkers derived from blood samples.

Note that the performance of biomarkers can only be as good as the quality of the training set—in this case, the accuracy of clinical diagnosis of TB. If the training set includes misclassified individuals, then the performance of the biosignature will inadvertently suffer. This is particularly important when considering comparisons of TB patients with clinically healthy, but latently infected individuals. Most attempts to harness biomarkers for the segregation of healthy uninfected and individuals with LTBI have produced relatively large error rates (often >30%), which could be the result of low sensitivity and low specificity of the tests used to determine the infection status. Considering that both groups are free of active disease, their blood signatures are more similar, resulting in a lower discriminative power to distinguish the two groups. Moreover, subclinical TB may already progress to clinical disease in some of these individuals and manifest as an apparent TB-like signature. However, if this were the case, it could in fact be harnessed for diagnosis TB in early stages using a biomarker-based diagnostic approach.

Another important issue has been raised by Kaforou et al. (2013). These authors compared the performance of biomarkers depending on whether the biosignature was based on a training set which included, in addition to samples from HIV^- donors, also HIV^+ TB patients and HIV^+ controls. They observed that performance substantially increased when HIV^+ TB patients were included; biomarkers derived from such a test set could be used to accurately classify HIV^- patients, but not vice versa: if the model construction was based on HIV^- TB patients only and controls only, then its performance in the HIV^+ population was poor.

PREDICTIVE BIOSIGNATURES FOR TB

A large body of evidence shows that diagnostic biosignatures of active TB disease can be defined. However, a high-throughput biosignature capable of predicting TB disease at a subclinical stage remains to be identified. This could be used for predicting (1) TB reactivation (progression from latent to active TB), (2) protective vaccine efficacy, and (3) treatment outcome.

It is tempting to assume that a host response signature can be defined at the molecular level, which precedes clinical diagnosis. This issue is being addressed by two ongoing efforts, a biomarker study based on an adolescent cohort (Mahomed et al. 2013) and GC6-74 (http://www.biomarkers-for-tb.net/consortium). In these studies, household contacts of active TB patients are being monitored over years to define signatures that can predict progression from infection to active TB disease. In the GC6 study, for example, ~3% of ~4500 household contacts developed active TB disease during the observation period of 2 years. Currently, metabolic and transcriptomic signatures are being defined that predict active TB disease at an early stage (i.e., before clinical TB diagnosis). It is hoped that it will be possible to capture early changes in gene expression and metabolic composition predictive of clinical TB onset thereafter.

The effect of drug treatment of TB on transcriptomic profiles has been investigated in independent studies by Bloom et al. (2012) and Cliff et al. (2013), who both detected rapid changes in IFN signaling and in innate immu-

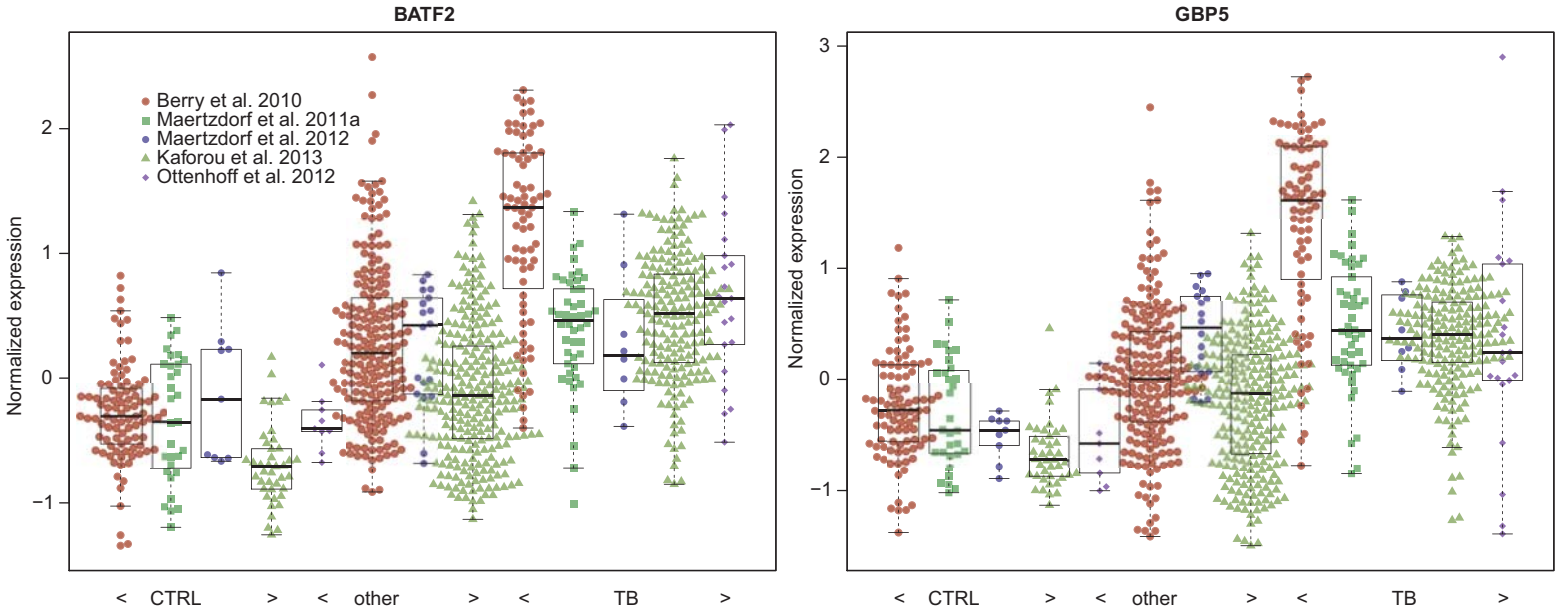

Figure 1. Normalized expression profiles of two selected genes (BATF2, basic leucine zipper transcription factor ATF-Like 2; GBP5, guanylate-binding protein 5) across different studies and study groups show a consistent landscape of TB responses on transcriptional level. CTRL, healthy donors (uninfected or with latent TB); other, other diseases; TB, clinical TB diagnosed.

nity, as early as 1 to 2 weeks after initiation of treatment. This finding suggests that it would be possible to quantitatively measure the response to TB treatment at an early stage of chemotherapy, although larger studies are needed for validation.

CONCLUSIONS

Most current diagnostic assays for TB have their own specific drawbacks and often rely on canonical single biomarkers with insufficient sensitivity and specificity. These limitations have prompted intensive search for multimarker host signatures with better predictive accuracy and improved diagnostic validity. High-throughput platforms, notably transcriptomics and metabolomics, have revealed numerous TB-related features, the majority of which have been confirmed in multiple studies. Focusing on the host response in TB, results are to a large extent compatible and reveal that the overall accuracy of a transcript biosignature can be as high as 95%. Bioinformatic analyses of the signatures identified are currently under validation in multicohort studies, bringing us a step closer to the identification of a unified signature of general applicability. Future efforts will be focused on validating the existing findings and, most importantly, on derivation of predictive signatures, useful for early diagnosis of risk of disease in infected healthy individuals.

ACKNOWLEDGMENTS

We thank Mary Louise Grossman for help preparing the manuscript. The authors are supported by the European Union's Seventh Framework Programme (EU FP7) ADITEC (HEALTH-F4-2011-280873) and NEWTBVAC (HEALTH-F3-2009-241745); the Bill and Melinda Gates Foundation Grand Challenges in Global Health Program (BMGF GC6-74, #37772 and GC12, #37885), and the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (BioVacSafe) Grant No. 115308.

REFERENCES

*Reference is also in this subject collection.

Al-Zamel FA. 2009. Detection and diagnosis of *Mycobacterium tuberculosis*. *Expert Rev Anti Infect Ther* **7:** 1099–1108.

Ballanti E, Perricone C, Greco E, Ballanti M, Di MG, Chimenti MS, Perricone R. 2013. Complement and autoimmunity. *Immunol Res* **56:** 477–491.

Barry CE III, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ, Young D. 2009. The spectrum of latent tuberculosis: Rethinking the biology and intervention strategies. *Nat Rev Microbiol* **7:** 845–855.

Barth RE, Mudrikova T, Hoepelman AI. 2008. Interferon-γ release assays (IGRAs) in high-endemic settings: Could they play a role in optimizing global TB diagnostics? Evaluating the possibilities of using IGRAs to diagnose active TB in a rural African setting. *Int J Infect Dis* **12:** e1–e6.

Bekmurzayeva A, Sypabekova M, Kanayeva D. 2013. Tuberculosis diagnosis using immunodominant, secreted antigens of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* **93:** 381–388.

Berry MP, Graham CM, McNab FW, Xu Z, Bloch SA, Oni T, Wilkinson KA, Banchereau R, Skinner J, Wilkinson RJ, et al. 2010. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466:** 973–977.

Bloom CI, Graham CM, Berry MP, Wilkinson KA, Oni T, Rozakeas F, Xu Z, Rossello-Urgell J, Chaussabel D, Banchereau J, et al. 2012. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. *PLoS ONE* **7:** e46191.

Bloom CI, Graham CM, Berry MP, Rozakeas F, Redford PS, Wang Y, Xu Z, Wilkinson KA, Wilkinson RJ, Kendrick Y, et al. 2013. Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PLoS ONE* **8:** e70630.

Cliff JM, Lee JS, Constantinou N, Cho JE, Clark TG, Ronacher K, King EC, Lukey PT, Duncan K, van Helden PD, et al. 2013. Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *J Infect Dis* **207:** 18–29.

Davies PD, Pai M. 2008. The diagnosis and misdiagnosis of tuberculosis. *Int J Tuberc Lung Dis* **12:** 1226–1234.

Diaz-Uriarte R, Alvarez de AS. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7:** 3.

Harries AD, Maher D, Nunn P. 1998. An approach to the problems of diagnosing and treating adult smear-negative pulmonary tuberculosis in high-HIV-prevalence settings in sub-Saharan Africa. *Bull World Health Organ* **76:** 651–662.

Ho JW, Stefani M, dos Remedios CG, Charleston MA. 2008. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics* **24:** i390–i398.

Jacobsen M, Repsilber D, Gutschmidt A, Neher A, Feldmann K, Mollenkopf HJ, Kaufmann SH, Ziegler A. 2006. Deconfounding microarray analysis—Independent mea-

surements of cell type proportions used in a regression model to resolve tissue heterogeneity bias. *Methods Inf Med* **45:** 557–563.

Jacobsen M, Repsilber D, Gutschmidt A, Neher A, Feldmann K, Mollenkopf HJ, Ziegler A, Kaufmann SH. 2007. Candidate biomarkers for discrimination between infection and disease caused by *Mycobacterium tuberculosis*. *J Mol Med (Berl)* **85:** 613–621.

Jacobsen M, Repsilber D, Kleinsteuber K, Gutschmidt A, Schommer-Leitner S, Black G, Walzl G, Kaufmann SH. 2011. Suppressor of cytokine signaling-3 is affected in T-cells from tuberculosis TB patients. *Clin Microbiol Infect* **17:** 1323–1331.

Joosten SA, Goeman JJ, Sutherland JS, Opmeer L, de Boer KG, Jacobsen M, Kaufmann SH, Finos L, Magis-Escurra C, Ota MO, et al. 2012. Identification of biomarkers for tuberculosis disease using a novel dual-color RT-MLPA assay. *Genes Immun* **13:** 71–82.

Joosten SA, Fletcher HA, Ottenhoff TH. 2013. A helicopter perspective on TB biomarkers: Pathway and process based analysis of gene expression data provides new insight into TB pathogenesis. *PLoS ONE* **8:** e73230.

Kaforou M, Wright VJ, Oni T, French N, Andersen ST, Bangani N, Banwell CM, Brent AJ, Crampin AC, Dockrell HM, et al. 2013. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: A case-control study. *PLoS Med* **10:** e1001538.

Koth LL, Solberg OD, Peng JC, Bhakta NR, Nguyen CP, Woodruff PG. 2011. Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis. *Am J Respir Crit Care Med* **184:** 1153–1163.

Lesho E, Forestiero FJ, Hirata MH, Hirata RD, Cecon L, Melo FF, Paik SH, Murata Y, Ferguson EW, Wang Z, et al. 2011. Transcriptional responses of host peripheral blood cells to tuberculosis infection. *Tuberculosis (Edinb)* **91:** 390–399.

Lin PL, Flynn JL. 2010. Understanding latent tuberculosis: A moving target. *J Immunol* **185:** 15–22.

Lin PL, Ford CB, Coleman MT, Myers AJ, Gawande R, Ioerger T, Sacchettini J, Fortune SM, Flynn JL. 2014. Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. *Nat Med* **20:** 75–79.

Maertzdorf J, Ota M, Repsilber D, Mollenkopf HJ, Weiner J, Hill PC, Kaufmann SH. 2011a. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. *PLoS ONE* **6:** e26938.

Maertzdorf J, Repsilber D, Parida SK, Stanley K, Roberts T, Black G, Walzl G, Kaufmann SH. 2011b. Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes Immun* **12:** 15–22.

Maertzdorf J, Weiner J III, Kaufmann SH. 2012a. Enabling biomarkers for tuberculosis control. *Int J Tuberc Lung Dis* **16:** 1140–1148.

Maertzdorf J, Weiner J III, Mollenkopf HJ, Bauer T, Prasse A, Muller-Quernheim J, Kaufmann SH. 2012b. Common patterns and disease-related signatures in tuberculosis and sarcoidosis. *Proc Natl Acad Sci* **109:** 7853–7858.

Mahomed H, Ehrlich R, Hawkridge T, Hatherill M, Geiter L, Kafaar F, Abrahams DA, Mulenga H, Tameris M, Geldenhuys H, et al. 2013. TB incidence in an adolescent cohort in South Africa. *PLoS ONE* **8:** e59652.

Mazurek GH, Weis SE, Moonan PK, Daley CL, Bernardo J, Lardizabal AA, Reves RR, Toney SR, Daniels LJ, LoBue PA. 2007. Prospective comparison of the tuberculin skin test and 2 whole-blood interferon-γ release assays in persons with suspected tuberculosis. *Clin Infect Dis* **45:** 837–845.

Mistry R, Cliff JM, Clayton CL, Beyers N, Mohamed YS, Wilson PA, Dockrell HM, Wallace DM, van Helden PD, Duncan K, et al. 2007. Gene-expression patterns in whole blood identify subjects at risk for recurrent tuberculosis. *J Infect Dis* **195:** 357–365.

Norbis L, Miotto P, Alagna R, Cirillo DM. 2013. Tuberculosis: Lights and shadows in the current diagnostic landscape. *New Microbiol* **36:** 111–120.

Normand R, Yanai I. 2013. An introduction to high-throughput sequencing experiments: Design and bioinformatics analysis. *Methods Mol Biol* **1038:** 1–26.

Ottenhoff TH, Dass RH, Yang N, Zhang MM, Wong HE, Sahiratmadja E, Khor CC, Alisjahbana B, van CR, Marzuki S, et al. 2012. Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. *PLoS ONE* **7:** e45839.

Pritchard CC, Hsu L, Delrow J, Nelson PS. 2001. Project normal: Defining normal variance in mouse gene expression. *Proc Natl Acad Sci* **98:** 13266–13271.

Steingart KR, Sohn H, Schiller I, Kloda LA, Boehme CC, Pai M, Dendukuri N. 2013. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database Syst Rev* **1:** CD009593.

Walzl G, Ronacher K, Hanekom W, Scriba TJ, Zumla A. 2011. Immunological biomarkers of tuberculosis. *Nat Rev Immunol* **11:** 343–354.

* Walzl G, Haks MC, Joosten SA, Kleynhans L, Ronacher K, Ottenhoff THM. 2014. Clinical immunology and multiplex biomarkers of human tuberculosis. *Cold Spring Harb Perspect Med* doi: 10.1101/cshperspect.a018515.

Weiner J III, Parida SK, Maertzdorf J, Black GF, Repsilber D, Telaar A, Mohney RP, Arndt-Sullivan C, Ganoza CA, Fae KC, et al. 2012. Biomarkers of inflammation, immunosuppression and stress with active disease are revealed by metabolomic profiling of tuberculosis patients. *PLoS ONE* **7:** e40221.

Weiner J, Maertzdorf J, Kaufmann SH. 2013. The dual role of biomarkers for understanding basic principles and devising novel intervention strategies in tuberculosis. *Ann NY Acad Sci* **1283:** 22–29.

Welsh KJ, Lewis CT, Boyd S, Braun MC, Actor JK. 2012. Complement factor C7 contributes to lung immunopathology caused by *Mycobacterium tuberculosis*. *Clin Dev Immunol* **2012:** 429675.