INT J TUBERC LUNG DIS 16(9):1140–1148
©2012 The Union
http://dx.doi.org/10.5588/ijtld.12.0246

STATE OF THE ART SERIES
New Tools, *Series editor:* Wing-Wai Yew
NUMBER 3 IN THE SERIES

Enabling biomarkers for tuberculosis control

J. Maertzdorf, J. Weiner III, S. H. E. Kaufmann

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

SUMMARY

Accelerated control of tuberculosis (TB) requires better control measures. Biomarkers, which reliably diagnose active TB or even predict risk of disease progression in individuals, could facilitate rapid diagnosis and treatment of TB patients and allow preventive measures for latently infected individuals with a high risk of TB. Moreover, biomarkers could speed up clinical trials with novel drug and vaccine candidates. Three platforms of global biomarker profiling will be described, with an emphasis on the most recent achievements: transcriptomics, proteomics and metabolomics. Moreover, we will discuss the need for computational analyses to make the best use of the plethora of data generated by biomarker research. Aside from their potential prognostic and diagnostic value, biomarkers could provide deeper insight into pathological processes underlying disease, and hence form the basis for novel intervention measures that target host molecules and pathways. We propose that biosignatures, which discriminate active TB from both latent infection and uninfected status, as well as from other diseases, will become available within the next decade. However, simple, low-cost biomarker-based point-of-care diagnosis will probably not be achieved in the next few years.

KEY WORDS: metabolomics; proteomics; TB diagnosis; transcriptomics; computational biology

BIOMARKERS are under the spotlight for the diagnosis of malignant and chronic inflammatory and infectious diseases.[1–6] Ideally, biomarkers provide 1) novel diagnostic measures that can predict risk of developing disease; 2) guidelines for personalised medicine, i.e., for individually tailored therapy of patients; and 3) guidance for diagnosis of difficult-to-define diseases or disease stages. It goes without saying that this does not come without a price. Can biomarkers also be applicable for a disease of inequality such as TB?

In regions with limited resources, TB is typically diagnosed by the identification of the causative agent, *Mycobacterium tuberculosis*, as acid-fast bacilli in sputum smears. Although this diagnostic test comes at low cost, it is also of low sensitivity, missing up to 40% of all TB cases, notably in regions with high HIV/AIDS (human immunodeficiency virus/acquired immune-deficiency syndrome) incidence.[7] More sensitive, but also more costly, are cultures of *M. tuberculosis* that, in addition, take several weeks. A breakthrough in TB diagnosis is the GeneXpert® MTB/RIF assay (Cepheid, Sunnyvale, CA, USA), which combines rapidity with high sensitivity and specificity.[8,9] The current cost of >US$16 per assay, however, is prohibitive for regions with limited resources.

In addition to diagnostic tests based on the identification and characterisation of *M. tuberculosis*, the host response is being exploited for TB diagnosis. The most widely used test, the tuberculin skin test (TST), measures the immune response against soluble antigens of *M. tuberculosis* as a delayed-type hypersensitivity reaction at the site of injection. It has proven its value for epidemiological studies. In a more sophisticated form, immune responses are measured as interferon-gamma (IFN-γ) release assays (IGRAs) using peripheral blood cells. Neither the TST nor IGRAs, however, can distinguish between latent TB infection (LTBI) and active TB. Rather, these assays measure immunity in response to infection independent from active disease. In regions with high incidence of *M. tuberculosis* infection, approaching 100%, TST and IGRA are hence of limited value.[10]

Previous articles in this series **Editorial:** Yew W-W, Borgdorff M W, Enarson D A. New tools for global tuberculosis control: are we any closer? Int J Tuberc Lung Dis 2012; 16(7): 855–856. **No 1:** Drobniewski F, Nikolayevskyy V, Balabanova Y, Bang D, Papaventsis D. Diagnosis of tuberculosis and drug resistance: what can new tools bring us? Int J Tuberc Lung Dis 2012; 16(7): 860–870. **No 2:** Grosset J H, Singer T G, Bishai W R. New drugs for the treatment of tuberculosis: hope and reality. Int J Tuberc Lung Dis 2012; 16(8): 1005–1014.

Correspondence to: Stefan H E Kaufmann, Max Planck Institute for Infection Biology, Charitéplatz 1, 10117 Berlin, Germany. Tel: (+49) 30 28460 500/502. Fax: (+49) 30 28460 501. e-mail: Kaufmann@mpiib-berlin.mpg.de

Article submitted 29 March 2012. Final version accepted 26 April 2012.

Biomarker studies typically measure different parameters of the host response, although they can also include markers from the pathogen in the infected individual. Biomarkers could be exploited for differential diagnosis of patients with active TB vs. LTBI. This is of particular value for cases in the process of converting from LTBI to active TB, particularly those who are not at the extremes of the health and illness spectrum. Such biomarkers should be highly sensitive and specific to achieve better performance than current measures. Moreover, they should eventually become simple and cost-effective so that one day they can be used as a point-of-care test. Most of all, however, biomarkers with prognostic value are needed in TB. Such biomarkers could predict the risk of developing active TB in individuals with LTBI and the outcome of drug treatment, i.e., cure vs. relapse. Such biomarkers would also be valuable for monitoring in clinical trials with new vaccines and new drugs by serving as surrogate markers of clinical endpoints. In TB vaccine trials, surrogate markers could predict vaccine efficacy in preventing TB before disease becomes clinically manifest; in drug trials, these markers would predict the efficacy of treatment before cure or relapse. Finally, biomarkers should be harnessed for stratification of study participants recruited into clinical trials. By identifying individuals with a high risk of active TB, they could significantly shorten the duration of, and reduce group sizes (and hence costs) in, clinical trials.

WHAT ARE BIOMARKERS AND BIOSIGNATURES?

A biomarker is defined as a biological characteristic that is objectively measured and evaluated as an indicator of physiological or pathological processes, for example as a response to a therapeutic (e.g., drug) or preventive (e.g., vaccine) measure.[11] Ultimately, biosignatures could provide correlates of resistance or susceptibility to developing active TB disease.

Biomarkers based on global or targeted measurements by transcriptomic, metabolomic and proteomic platforms can be tailored to form a biosignature composed of several markers, possibly provided by different platforms. In addition, antigen-induced immune responses are being exploited for biomarkers in TB, with an emphasis on multiplex cytokine measurements of antigen-stimulated whole blood cells.[12,13]

In the following, we will describe state-of-the-art platforms for host biosignatures in different biological samples and consider their future scope and feasibility for TB (Table 1, Figure). We will discuss the importance of statistical and bioinformatical approaches for analysing the plethora of data generated by biomarker studies and the need to develop simpler test systems for TB diagnosis. Although it is likely that biomic marker research will develop platforms for better TB diagnosis and prognosis in high-income countries within the next decade, a simple, low-cost biomarker-based point-of-care test for use in low-income regions will likely not become available in the near future.

Table 1 Major platforms for host biosignatures in TB[15]

- The transcriptome comprises RNA molecules, which first serve as messengers from the DNA blueprint to the synthesis of proteins, and second, as noncoding small-RNA, perform regulatory functions. In humans, an estimated 50 000–100 000 transcripts are targets of transcriptomic measurements. For these analyses, peripheral blood leukocytes are typically used.
- The proteome encompasses up to 1 000 000 different proteins. Typically, serum or plasma depleted of highly abundant compounds of low informative value, such as albumin, are used.
- The metabolome includes small biochemicals in the order of a few thousand molecules. These analyses are typically performed with serum or urine samples. A subgroup of metabolites, volatile compounds, can be measured in the breath.

TRANSCRIPTOMICS

Gene expression and regulation are highly dynamic processes that vary considerably in different circumstances. In response to its requirements or dedication, each cell constantly regulates the activation state of its genes. At any time, particular sets of genes will be active or silent, providing a kind of genetic blueprint of that cell under a given condition. The study of the complete set of RNA transcripts at a given time, transcriptomics, has developed into a valuable field of research to investigate different disease conditions and to identify disease-related biomarkers. The best established approach is the high-throughput measurement of gene expression by microarray chips, targeting and quantifying the vast majority of cellular gene transcripts.

Comparisons of gene expression profiles in TB patients and LTBI in blood cells by several groups has provided novel insights into pathological conditions in TB.[16–22] Combined, a general but consistent picture of altered gene expression patterns in TB patients has emerged from these studies. Identified profiles are indicative of chronic activation of the immune system in TB patients, with a marked activation of IFN signalling, pro-inflammatory signalling through the Janus kinase signal transducer and activator of transcription (JAK-STAT) pathway, elevated expression of Fc-gamma receptors and their downstream response elements, as well as regulation of apoptosis. So far, no published study has focused on cohort-specific gene expression and differences between geographic regions. Despite the robust microarray analyses, combining data from several groups and cohorts may not allow for adequate inter-cohort comparisons due to differences in sample treatment, array platforms and data analysis.

A central issue in the interpretation of transcriptomic data is the disease specificity of the profile. Definition of a unique biosignature for TB demands that

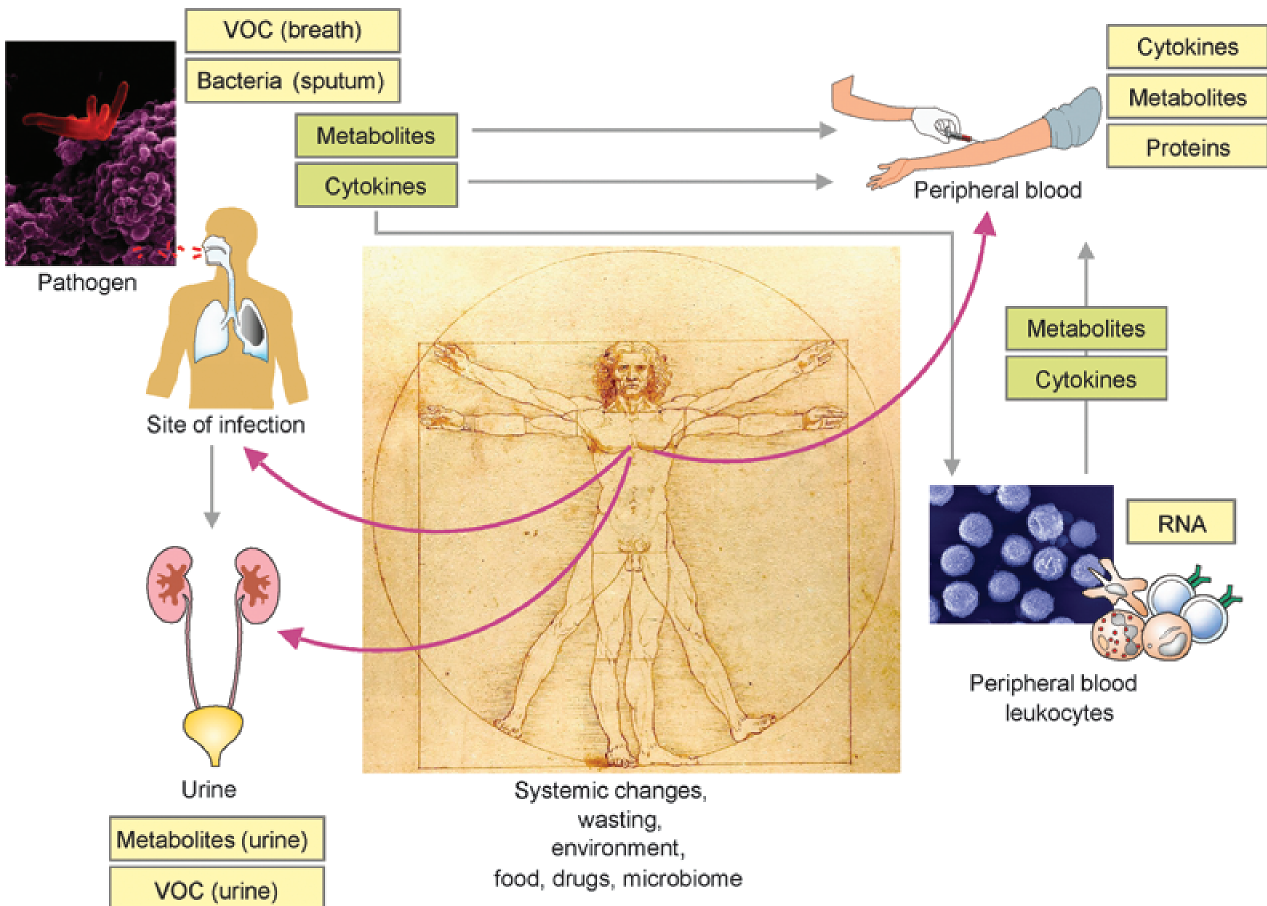

Figure Simplified scheme describing different options for biomarker measurement in TB. Most biomarkers are determined in peripheral blood, either in serum/plasma or in leukocytes. Other, less invasive sources for biomarkers are breath, sputum and urine. However, measurements of biomarkers in body fluids or blood cells need to take into account potentially incomplete reflection of activities at the organ of bacterial residence—in the case of TB, typically the lung. Moreover, biomarker measurements need to consider the impact of exogenous factors, including food, drugs and microbiome. Note that antibodies are not included in the figure because current antibody determinations are of doubtful diagnostic value (see also http://www.who.int/tdr/news/2011/tb-test-alert/en/).[14] VOC = volatile organic compounds; TB = tuberculosis.

identified profiles are disease-specific. Computational approaches making a comparison with known disease pathways and expression data from other diseases indicate remarkable similarity with other diseases.[19,23] Changes in gene expression thus partially reflect a more common state of inflammation, while disease-specific patterns are needed for differential diagnosis. This need for disease specificity is illustrated by the high degree of similarity recently identified in transcriptomic profiles between pulmonary TB and sarcoidosis, a disease with similar pathology but distinct aetiology.[24] Nevertheless, several expression patterns unique to each disease could also be identified. The commonalities in blood profiles between these diseases reflect shared inflammatory processes, which are critical for understanding the underlying mechanisms of pathology and may be used as building blocks for the identification of clusters of diseases, while others allow for differential diagnosis of distinct types of diseases, e.g., of infectious or non-infectious aetiologies. This consideration could greatly benefit the definition of biosignatures for TB and other diseases, and the development of new array-based diagnostic tools.

The transcriptome also includes small, noncoding RNAs, which have emerged as key regulators of gene expression, genome stability and chromatin modification.[25,26] As post-transcriptional regulators of gene expression, microRNAs (miRNAs) likely play an important role in modulating biological pathways relevant to inflammation. As such, miRNAs are potential therapeutic targets for the treatment of chronic inflammatory lung diseases.[27] Due to the regulatory nature of these small RNAs, targeting multiple gene transcripts, their potential role in directing immune processes is currently under investigation. The first studies describing the relevance of miRNA as biomarkers in TB have recently appeared in the literature,[28–31] although their precise roles in the development or regulation of disease processes still remain to be elucidated.

Recent technological advances also allow for affordable, highly detailed transcriptome analysis by RNA sequencing (RNA-seq).[32] This next-generation

sequencing provides a more comprehensive and unbiased analysis of all types of RNA molecules (i.e., messenger RNAs, miRNAs and other noncoding small RNAs), allowing both quantitative and qualitative measurements. Due to the detailed analysis down to the single nucleotide level, RNA-seq enables identification of point mutations and changes in splice variants of expressed genes. Moreover, the digital readout allows a highly dynamic range of precise quantification of gene expression.[32] Hybridisation-based approaches such as microarrays, on the other hand, are less costly, allowing for larger sample sizes; the main drawbacks include high background noise, bias imposed by the need for normalisation of the signal, and limitation to transcripts known to be expressed a priori. Early studies have revealed fairly robust correlations in quantitative gene expression measurements between RNA-seq and microarray analysis.[32,33] An open source database has recently been launched, providing data mining tools and access to RNA-seq expression profiles.[34]

PROTEOMICS

The field of proteomics comprises the large-scale study of the entire content of proteins in a particular biological sample, including their structures and functions. Proteomics can thus provide a deep understanding of the actual biological functions of gene products in an organism, as proteins are the actual effector molecules within a biological system. Disease-related high-throughput proteomics has attracted increasing interest since the unravelling of the genome sequences of humans and numerous pathogens.[35] Proteomic technologies can play an important role in drug discovery, diagnostics and molecular medicine, where protein biomarkers represent a promising target of clinical proteomics. The application of proteomics technology in biomarker discovery has recently been extensively reviewed.[36,37] Although proteomics may deliver a biosignature that is more directly related to TB stages than gene transcripts, it is also much more complex than genomics and transcriptomics, with the proteome comprising many more analytes that are dynamic in structure and function. Moreover, many proteins undergo post-translational modifications and often form multimeric complexes as the actual functional entities within an organism.

Many approaches aim to identify all proteins in a complex sample, but identifying clinically relevant biomarkers using such shotgun applications has mostly failed.[38] The search for protein biomarkers has been dominated by high-throughput mass spectrometry (MS), allowing the identification of thousands of proteins within complex biological samples. Selected reaction monitoring experiments, on the other hand, focus on selected sets of proteins.[39] Such targeted approaches, which are today more widely used, focus on defined sets of proteins, e.g., based on their size or homology with selected sets of synthetic peptides.

The development of antibody-based arrays offers another targeted approach of high-throughput screening of the proteome in biological samples.[40] Using antibody-based proteomics, the Human Protein Atlas project was initiated in 2003 with the aim of creating a map of protein expression patterns in a variety of human cells and tissues. These data are now publicly available.[41] Other platforms that are less high-throughput-oriented, such as bead assays, are primarily exploited for the determination of particular subsets of proteins (e.g., cytokines in serum).

Current proteomic technologies applied in microbiology and the impact of proteomics on understanding *M. tuberculosis* biology have been the subject of numerous reviews.[42–46] On the host side, studies focusing on the host response to *M. tuberculosis* have revealed considerable proteomic variations, disclosing functional changes in macrophages upon stimulation with *M. tuberculosis* lipids,[47] or modulation of the phagosomal proteome by mycobacterial infection.[48]

Most widely used for the identification of protein biomarkers are peripheral blood serum samples (Figure). MS profiling and statistical modelling has yielded potential serum biomarkers for TB, able to distinguish pulmonary from extra-pulmonary TB and other infectious diseases.[49,50] Serum proteome profiles have also revealed differential changes in smear-positive and -negative TB patients and controls, which may shed light on the proteins involved in underlying disease processes.[51] Another study has described the dynamic antibody responses to *M. tuberculosis*, revealing responses to a relatively small proportion of the *M. tuberculosis* proteome and variations in target preference amongst patients.[52]

Thus far, experimental approaches have mostly relied on relative quantification of proteins, but there is a growing need for absolute quantification to enable predictive systems biology modelling.[53] The proteome is a complex mosaic composed of thousands of proteins, ranging in abundance from tens to hundreds of thousands of copies. Notably, the identification and quantification of proteins of low abundance in this respect remains a significant challenge.[54]

Using proteomics for the identification of disease-related biomarkers still faces several serious obstacles. Despite numerous potential proteomic biosignatures published over the years, validation of such markers in clinical settings has proven challenging.[55] The difficulty of defining disease-related proteomic markers is illustrated by a recently published metastudy analysing biomarker studies in several diseases. The results revealed significant inconsistency in identified biomarkers between the studies in most diseases.[56] Finally, proteomics research faces several analytical challenges. These include the use of algorithms for protein identification, differences in MS spectra resulting

from different fragmentation methods, and the relative low precision and accuracy of the most widely used instruments.[39]

METABOLIC PROFILING

Over the last decades, metabolic profiling has been increasingly applied to classify diseases and diagnose onset of disease.[57] Metabolic profiles have been defined for different types of cancer,[58] diabetes[59–61] and hereditary diseases. A number of technologies are being employed for metabolic profiling, comprising two steps in the wet laboratory: first, separation of molecules (e.g., by gas or liquid chromatography) and second, detection of separated molecules (e.g., by nuclear magnetic resonance [NMR] or MS[62]). Each combination of methodologies will result in a different profile. This is in stark contrast to the analyses of the transcriptome, which normally captures the full extent of known expressed protein-coding genes or even, in the case of next-generation sequencing, of the whole transcriptome.

The term 'metabolomics' is somewhat loosely defined.[62] Principally, it concerns the sum of small molecules that are involved in the metabolism of an organism, such as sugars, nucleotides, amino acids, lipids and fatty acids. However, when a high-throughput method, such as MS or NMR, is used for metabolic biomarkers from body fluids, whole tissues or organs (Figure), the term 'metabolic profiles' or 'profiling' should be employed for the sake of precision, as this latter term is established in the field of spectrometry.[63]

Metabolic profiling has been successfully applied to only a few infectious diseases. Laiakis et al. investigated matched plasma and urine samples for a biomarker signature distinguishing pneumonia patients from community controls.[64] Despite the relatively small sample size, the metabolic signature, including uric acid, hypoxanthine, glutamic acid, tryptophan and adenosine-5′-diphosphate, was sufficiently well-defined. Shin et al. applied metabolomic profiling to the serum, lung, spleen and liver of mice infected with *M. tuberculosis*.[65] Although the study did not derive a strict biosignature, the relative abundances of several compounds differed significantly, indicating distinct biological processes, including carbohydrate, lipid, fatty acid, amino acid and nucleotide metabolism, and indications of an anti-oxidative stress response.

Analysis of lung granulomas from *M. tuberculosis*-infected guinea pigs[66] revealed increased tissue lactate levels (which could be attributed to hypoxia and necrosis in granulomas), increased concentrations of glutathione (suggesting an involvement of the redox system and possibly oxidative stress), increased abundance of betaine and trimethylamine-N-oxide, as well as a signature for lipolysis. We compared metabolic profiles of healthy controls (both LTBI and uninfected) and untreated patients with active TB.[67] This analysis provides clues on the activity of different biological pathways during TB.

A particular feature of metabolomics is the broad spectrum of analytes, which originate not only from the host and the pathogen but also from the microbiome and the environment. Thus, it is not unusual to detect elevated levels of drugs or drug-derived degradation products in patient groups. While this plethora can provide interesting information on the complex milieu of a patient, it can complicate the definition of biomarkers specific for a distinct disease stage. Another group of metabolic biomarkers, volatile organic compounds (VOC; Figure), have been exploited for the diagnosis of TB patients. Banday et al. identified isopropyl acetate, o-xylene, cymol, 2,6-dimethystyrene and 3-pentanol in urine for the differential diagnosis of TB patients,[68] and Phillips et al. defined a biosignature based on VOC in the breath of patients with symptoms of TB.[69] These were identified by electronic noses.[70]

Alternatively, African pouched rats (*Cricetomys gambianus*) have been exploited as biological sensors for VOC. These rats have a highly sensitive olfactory system and can be trained to detect a composition of compounds indicative of *M. tuberculosis*. Proof of principle for TB diagnosis by *C. gambianus* rats has been documented.[71,72] Typical compounds identified by these rats are generic VOC, including methyl-4-anisate, methyl nicotinate, 2-phenylanisole, 4-methyl anisole, methyl-2-amino benzoate, and fatty acid derivatives, such as 3-methyl-4-pentanolide and 4-methyl-5-hexanolide.[71,72]

BIOMARKERS AND CLASSIFICATION ALGORITHMS

Definition of a set of biomarkers derived from a high-throughput platform requires that the output data provide answers to the following questions: 1) Can the data predict the infection/disease status of an individual in a sensitive and specific way? and 2) What is the smallest subset of data required for robust classification? The data set from a high-throughput platform is best viewed as a large matrix of numerous biological samples, grouped in defined classes whereby for each sample a large number of variables exists (predictors, e.g., genes in transcriptome analysis). A common first approach to these questions is the use of a so-called unsupervised algorithm, such as principal component analysis (PCA).[73] The data are clustered or visualised without any information about their classification, this being the main advantage of such an approach. The classification results are subsequently explored to account for possible common effects of grouping the data. However, statistical significance of the findings or calculation of the overall specificity and sensitivity are impossible at this stage. As a next step, a supervised machine learning

algorithm is applied. Usually, the data are divided into classes (e.g., disease and control), and a randomly sampled subset of the data (called the 'training set') is used to generate a mathematical model that can be applied to a new set of samples to infer the class these samples belong to, called the 'test set'. This allows the model to be tested against the remaining subset of the original data.[74] The correctness of the model can be verified by a final test against an independent data set (called the 'validation set').

Splitting the data into training and test sets significantly reduces the power of the analysis, as a substantial proportion of samples is required for the test set. Therefore, the whole sequence—generating the sets, model construction and testing—is repeated several times, and the results are summarised over all iterations, for example, using leave-one-out (LOO) randomisation or bootstrapping.[74] This is repeated for all samples, maximising the predictive power, while at the same time ensuring the independence of the test set. After splitting the data into a training set and a test set, external cross-validation of the model against the test set is essential. If the model were tested against the data used to construct it, the results would become meaningless, as the models would be prone to over-fitting.

Several families of machine learning algorithms have been developed that can be applied to biological data sets; most of them have been implemented within the Bioconductor framework in the widely used statistical analysis language *R*.[75] Note that the method needs to allow for easy backtracking of the variables used to construct the model, ruling out algorithms such as neural networks that function as black boxes. Moreover, the method should allow classification into more than one study group. Examples of algorithms frequently applied in biomarker research include random forests, linear discriminant analysis (LDA) and its variants, k-nearest neighbour and support vector machines.[76]

Analysing large data sets such as those from gene expression profiling raises a particular issue. The sheer number of variables involved—exceeding 30 000 in the case of common microarray platforms—may be prohibitive due to the computational complexity of the algorithms applied. The variables used (e.g., set of transcripts used for further analysis) must be filtered to reduce the size of the data. A common method of filtering is by choosing components with the highest variability, as determined by interquartile ranges (IQR).[74] The variables (e.g., transcripts) are sorted according to their IQR, and a limited number is selected for classification. While such an approach seems unbiased with respect to the classification of the samples (this information is not directly used in the selection of variables), it relies on both test and training set samples. Selection of variables for biosignatures may also employ projections to latent structures (PLS) or, alternatively, partial least squares regression. PLS can be understood as an expansion of PCA.[73] While PCA reduces the numbers of variables of a data set by creating composite variables (principal components), PLS does so on two data sets: one containing predictors (such as transcripts or metabolites), and another one containing the response variables (such as several variables relating to disease severity in a patient or response to drug treatment or vaccination). A particular variation of PLS, PLS discriminant analysis (PLS–DA), can select a categorical variable as response, and translate it into the equivalent of a machine learning algorithm.

FINDING CONNECTIONS BETWEEN BIOMARKERS AND INFERRING HIGHER ORDER FUNCTIONS

Given a sufficient sample size, data from high-throughput biomarker studies can be used to mine functional connections between the variables in question. Mutual information index and correlation coefficients are frequently employed.[77] One assumes that within a group of individuals, variability in functionally linked variables (e.g., expression of genes, abundance of metabolites, etc.) is correlated. Genes that participate in the same pathway or are regulated by the same transcription factor will be highly correlated. The resulting network of functional connections between variables allows extraction of relevant biological pathways and mechanisms, which can then be employed for a more targeted approach.

Functional relations inferred from co-expression can be used to create clusters or modules of functionally related genes, which can provide deeper biological insights and thus form the basis for new hypotheses as well as data exploration and visualisation. Berry et al.[19] utilised a previously defined set of modules of co-expressed genes related to immune functions[78] to explore the biosignature of the TB-specific transcriptome. Extensive genomic co-expression networks are available for eukaryotic genomes,[79–81] but for other biomarker platforms comparably extensive resources do not exist. The most useful publicly available resources for functional analyses of metabolites include the KEGG (Kyoto Encyclopedia of Genes and Genomes) atlas,[82] the Human Metabolome Database (HMDB),[83] and Recon 1.[84]

MULTIPLATFORM APPROACHES

The various biomarker platforms record the state of an organism from different perspectives: for example, whole blood cell transcriptomics reveals expression patterns of cells from the circulatory system, which partially reflects local reactions but does not provide direct insights into immune cell activities at sites of disease manifestation. However, the very nature of

the information—recording the activation of single genes, gene groups and pathways—allows inference of particular responses, such as the IFN-γ pathway in the case of TB.

Metabolic profiling of serum or urine samples provides a broader perspective with lower resolution. Some of the identified molecules—numerous amino acids or lipids in the case of TB—will change according to the general state of the host, for example, resulting from TB-related cachexia. On the other hand, changes in abundance of other substances may be caused more directly by TB pathogenesis. Evidence has been presented that serum concentrations of kynurenine in TB patients result from increased indoleamine 2, 3-dioxygenase (IDO) expression in *M. tuberculosis*-infected antigen-presenting cells.[67] Although these metabolites can be directly correlated to gene expression in whole blood cells, they likely result from ongoing pathogenesis in the lungs (Figure).

Given that different platforms reveal distinct aspects of the host response, a greater explanatory power and, consequently, deeper biological insights can be expected from combining variables from them. Multiplatform approaches are not yet widespread. Recently, Blanchet et al. described how to combine predictions from metabolomic and proteomic data sets for experimental autoimmune encephalomyelitis.[85] They proposed a two-stage statistical method to join the data sets, based on calculation of canonical variates in the first step, followed by PCA reduction of the resulting joined global scores. The method, based on the joined prediction from metabolite and protein signatures, resulted in higher robustness of predictions and allowed coherent biological exploration of the host response.[85] For TB, we investigated a link between metabolite and cytokine profiles, and found that relative abundances of several small metabolites were correlated with distinct cytokines present at higher serum abundance in TB patients than in controls.[67] This indicates consistency in the information provided by the two profiles.

CONCLUSIONS

Biomarker research has appeared on the radar screen of TB research. High expectations have been raised that host biomarkers will benefit the diagnosis and clinical surveillance of TB patients, as well as vaccine and drug development. In particular, biomarkers are foreseen to 1) predict risk of active TB in healthy individuals with LTBI, 2) more precisely diagnose progression from LTBI to active TB disease, which often oscillates between health and illness, 3) discriminate between different types of TB disease, including extrapulmonary TB, and 4) provide a surrogate marker for disease severity. Biomarkers are already well advanced for the diagnosis and prognosis of cancer, metabolic disorders and diabetes, and are gaining increasing importance for infectious diseases.[1,6,13,64]

Table 2 Challenges in biomarker development

Disease
- Continuous spectrum of disease severity rather than two clear-cut classes (LTBI vs. active TB)
- Poorly defined TB case definition
- Disease relevance of candidate biomarkers
- Predictive value of biosignatures on an individual level

Study design
- Sample numbers too small
- Variations within and between study populations
- Validation and clinical utility of candidate markers

Platform
- Variations in sample collection and preparation
- Lack of a standard technology platform
- Reproducibility of platforms

Computational
- Complexity in statistical evaluation of heterogeneous data sets
- Computational challenges in the analysis of large data sets

LTBI = latent TB infection; TB = tuberculosis.

One drawback of previous biomarker studies using unbiased, high-throughput biomics platforms is that they have largely focused on differential abundance of individual molecules. However, differential abundance of particular markers does not provide sufficient information on their relevance in a specific disease or disease stage. Rather, high-throughput biomarker scans followed by pathway analyses and other downstream approaches[86] may be more suited to gain biological insights into the processes involved in the pathology of a particular disease and allow comparison with other diseases. This in turn would allow the rational design of pathognomonic biosignatures (Table 2).

An essential prerequisite for a qualified biosignature is its validation in different clinical settings—probably the most challenging aspect of all biomarker research.[87] Without doubt, meta-analyses of TB biomarkers will be performed once sufficient data have been made publically available. However, extensive multiregional studies involving various disease cohorts will still be needed. A second major challenge in translational biomarker research in TB is the heterogeneous interplay between *M. tuberculosis* and the human host. Similar to the diverse pathological representations of active TB disease, LTBI encompasses a diverse range of biological states. A clinically useful biosignature should thus not only diagnose active TB, but also predict risk of disease. In this respect, the concept of a TB spectrum implies that biomarkers need to define a certain position in the spectrum, so that the risk for reactivation and disease progression can be assessed.[88]

Acknowledgements

The authors thank D Schad for help with graphics and M L Grossman for help with preparing the manuscript. SHEK is an advisor at Cellestis Ltd. This work is supported by EU FP7 projects NEWTBVAC (HEALTH-F3-2009-241745), TRANSVAC (FP7-INFRASTRUCTURES-2008-228403), ADITEC (HEALTH-F4-2011-280873) by the European Union, the European and Developing Countries Clinical Trials Partnership project 'Collaboration and integration of tuberculosis vaccine trials in Europe and Africa'

(TBTEA, MS.20120.10800.002), the Bill & Melinda Gates Foundation Grand Challenges in Global Health Program (BMGF GC6-74, #37772) and the Innovative Medicines Initiative Joint Undertaking 'Biomarkers for Enhanced Vaccine Safety' (IMI JU Grant No. 115308).

Reference

1 Bryant P A, Venter D, Robins-Browne R, Curtis N. Chips with everything: DNA microarrays in infectious diseases. Lancet Infect Dis 2004; 4: 100–111.
2 Gerszten R E, Wang T J. The search for new cardiovascular biomarkers. Nature 2008; 451: 949–952.
3 Heidecker B, Kasper E K, Wittstein I S, et al. Transcriptomic biomarkers for individual risk assessment in new-onset heart failure. Circulation 2008; 118: 238–246.
4 Makarov D V, Loeb S, Getzenberg R H, Partin A W. Biomarkers for prostate cancer. Annu Rev Med 2009; 60: 139–151.
5 Spratlin J L, Serkova N J, Eckhardt S G. Clinical applications of metabolomics in oncology: a review. Clin Cancer Res 2009; 15: 431–440.
6 Chaussabel D, Pascual V, Banchereau J. Assessing the human immune system through blood transcriptomics. BMC Biol 2010; 8: 84.
7 Swai H F, Mugusi F M, Mbwambo J K. Sputum smear-negative pulmonary tuberculosis: sensitivity and specificity of diagnostic algorithm. BMC Res Notes 2011; 4: 475.
8 Ioannidis P, Papaventsis D, Karabela S, et al. Cepheid GeneXpert *M. tuberculosis*/RIF assay for *Mycobacterium tuberculosis* detection and rifampin resistance identification in patients with substantial clinical indications of tuberculosis and smear-negative microscopy results. J Clin Microbiol 2011; 49: 3068–3070.
9 Boehme C C, Nicol M P, Nabeta P, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert *M. tuberculosis*/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. Lancet 2011; 377: 1495–1505.
10 Lönnroth K, Castro K G, Chakaya J M, et al. Tuberculosis control and elimination 2010–50: cure, care, and social development. Lancet 2010; 375: 1814–1829.
11 Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001; 69: 89–95.
12 Chegou N N, Black G F, Kidd M, van Helden P D, Walzl G. Host markers in QuantiFERON supernatants differentiate active TB from latent TB infection: preliminary report. BMC Pulm Med 2009; 9: 21.
13 Walzl G, Ronacher K, Hanekom W, Scriba T J, Zumla A. Immunological biomarkers of tuberculosis. Nat Rev Immunol 2011; 11: 343–354.
14 Morris K. WHO recommends against inaccurate tuberculosis tests. Lancet 2011; 377: 113–114.
15 Zhu J, He F, Wang J, Yu J. Modeling transcriptome based on transcript-sampling data. PLoS ONE 2008; 3: e1659.
16 Mistry R, Cliff J M, Clayton C L, et al. Gene-expression patterns in whole blood identify subjects at risk for recurrent tuberculosis. J Infect Dis 2007; 195: 357–365.
17 Jacobsen M, Repsilber D, Gutschmidt A, et al. Candidate biomarkers for discrimination between infection and disease caused by *Mycobacterium tuberculosis*. J Mol Med 2007; 85: 613–621.
18 Jacobsen M, Mattow J, Repsilber D, Kaufmann S H. Novel strategies to identify biomarkers in tuberculosis. Biol Chem 2008; 389: 487–495.
19 Berry M P, Graham C M, McNab F W, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 2010; 466: 973–977.
20 Maertzdorf J, Ota M O, Repsilber D, et al. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. PLoS ONE 2011; 6: e26938.
21 Maertzdorf J, Repsilber D, Parida S K, et al. Human gene expression profiles of susceptibility and resistance in tuberculosis. Genes Immun 2011; 12: 15–22.
22 Lesho E, Forestiero F J, Hirata M H, et al. Transcriptional responses of host peripheral blood cells to tuberculosis infection. Tuberculosis (Edinb) 2011; 91: 390–399.
23 Koth L L, Solberg O D, Peng J C, Bhakta N R, Nguyen C P, Woodruff P G. Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis. Am J Respir Crit Care Med 2011; 184: 1153–1163.
24 Maertzdorf J, Weiner J, Mollenkopf H-J, et al. Common patterns and disease-related signatures in tuberculosis and sarcoidosis. Proc Natl Acad Sci USA 2012; 109: 7853–7858.
25 Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature 2009; 457: 413–420.
26 Wery M, Kwapisz M, Morillon A. Noncoding RNAs in gene regulation. Wiley Interdiscip Rev Syst Biol Med 2011; 3: 728–738.
27 Hassan T, McKiernan P J, McElvaney N G, Cryan S A, Greene C M. Therapeutic modulation of miRNA for the treatment of proinflammatory lung diseases. Expert Rev Anti Infect Ther 2012; 10: 359–368.
28 Wang C, Yang S, Sun G, et al. Comparative miRNA expression profiles in individuals with latent and active tuberculosis. PLoS ONE 2011; 6: e25832.
29 Fu Y, Yi Z, Wu X, Li J, Xu F. Circulating microRNAs in patients with active pulmonary tuberculosis. J Clin Microbiol 2011; 49: 4246–4251.
30 Liu Y, Wang X, Jiang J, Cao Z, Yang B, Cheng X. Modulation of T-cell cytokine production by miR-144* with elevated expression in patients with pulmonary tuberculosis. Mol Immunol 2011; 48: 1084–1090.
31 Wu J, Lu C, Diao N, et al. Analysis of microRNA expression profiling identifies miR-155 and miR-155* as potential diagnostic markers for active tuberculosis: a preliminary study. Hum Immunol 2012; 73: 31–37.
32 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10: 57–63.
33 Subramaniam S, Hsiao G. Gene-expression measurement: variance-modeling considerations for robust data analysis. Nat Immunol 2012; 13: 199–203.
34 Krupp M, Marquardt J U, Sahin U, Galle P R, Castle J, Teufel A. RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next generation sequencing. Bioinformatics 2012; 28: 1184–1185.
35 Hanash S. Disease proteomics. Nature 2003; 422: 226–232.
36 Calligaris D, Villard C, Lafitte D. Advances in top-down proteomics for disease biomarker discovery. J Proteomics 2011; 74: 920–934.
37 Ray S, Reddy P J, Jain R, Gollapalli K, Moiyadi A, Srivastava S. Proteomic technologies for the identification of disease biomarkers in serum: advances and challenges ahead. Proteomics 2011; 11: 2139–2161.
38 Ohtsuki S, Uchida Y, Kubo Y, Terasaki T. Quantitative targeted absolute proteomics-based ADME research as a new path to drug discovery and development: methodology, advantages, strategy, and prospects. J Pharm Sci 2011; 100: 3547–3559.
39 Noble W S, MacCoss M J. Computational and statistical analysis of protein mass spectrometry data. PLoS Comput Biol 2012; 8: e1002296.
40 Borrebaeck C A, Wingren C. Recombinant antibodies for the generation of antibody arrays. Methods Mol Biol 2011; 785: 247–262.
41 Ponten F, Schwenk J M, Asplund A, Edqvist P H. The Human Protein Atlas as a proteomic resource for biomarker discovery. J Intern Med 2011; 270: 428–446.
42 Jungblut P R, Schaible U E, Mollenkopf H-J, et al. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. Mol Microbiol 1999; 33: 1103–1117.

43 Ehebauer M T, Wilmanns M. The progress made in determining the *Mycobacterium tuberculosis* structural proteome. Proteomics 2011; 11: 3128–3133.
44 Graham C, McMullan G, Graham R L. Proteomics in the microbial sciences. Bioeng Bugs 2011; 2: 17–30.
45 de Souza G A, Wiker H G. A proteomic view of mycobacteria. Proteomics 2011; 11: 3118–3127.
46 Mehaffy M C, Kruh-Garcia N A, Dobos K M. Prospective on *Mycobacterium tuberculosis* proteomics. J Proteome Res 2012; 11: 17–25.
47 Shui W, Gilmore S A, Sheu L, Liu J, Keasling J D, Bertozzi C R. Quantitative proteomic profiling of host-pathogen interactions: the macrophage response to *Mycobacterium tuberculosis* lipids. J Proteome Res 2009; 8: 282–289.
48 Rao P K, Singh C R, Jagannath C, Li Q. A systems biology approach to study the phagosomal proteome modulated by mycobacterial infections. Int J Clin Exp Med 2009; 2: 233–247.
49 Agranoff D, Fernandez-Reyes D, Papadopoulos M C, et al. Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. Lancet 2006; 368: 1012–1021.
50 Deng C, Lin M, Hu C, et al. Establishing a serologic decision tree model of extrapulmonary tuberculosis by MALDI-TOF MS analysis. Diagn Microbiol Infect Dis 2011; 71: 144–150.
51 Liu Q, Chen X, Hu C, et al. Serum protein profiling of smear-positive and smear-negative pulmonary tuberculosis using SELDI-TOF mass spectrometry. Lung 2010; 188: 15–23.
52 Kunnath-Velayudhan S, Salamon H, Wang H Y, et al. Dynamic antibody responses to the *Mycobacterium tuberculosis* proteome. Proc Natl Acad Sci USA 2010; 107: 14703–14708.
53 Brownridge P, Holman S W, Gaskell S J, et al. Global absolute quantification of a proteome: challenges in the deployment of a QconCAT strategy. Proteomics 2011; 11: 2957–2970.
54 Brewis I A, Brennan P. Proteomics technologies for the global identification and quantification of proteins. Adv Protein Chem Struct Biol 2010; 80: 1–44.
55 Borrebaeck C A, Wingren C. Transferring proteomic discoveries into clinical practice. Expert Rev Proteomics 2009; 6: 11–13.
56 Al-Tarawneh S K, Border M B, Dibble C F, Bencharit S. Defining salivary biomarkers using mass spectrometry-based proteomics: a systematic review. OMICS 2011; 15: 353–361.
57 Goodacre R, Vaidyanathan S, Dunn W B, Harrigan G G, Kell D B. Metabolomics by numbers: acquiring and understanding global metabolite data. TRENDS in Biotechnology 2004; 22: 245–252.
58 Yang J, Xu G, Zheng Y, Kong H, Pang T, Lv S, et al. Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. J Chromatogr B Analyt Technol Biomed Life Sci 2004; 813: 59–65.
59 Yuan K, Kong H, Guan Y, Yang J, Xu G. A GC-based metabonomics investigation of type 2 diabetes by organic acids metabolic profile. J Chromatogr B Analyt Technol Biomed Life Sci 2007; 850: 236–240.
60 Wang T J, Larson M G, Vasan R S, et al. Metabolite profiles and the risk of developing diabetes. Nat Med 2011; 17: 448–453.
61 Salek R M, Maguire M L, Bentley E, et al. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat and human. Physiol Genomics 2007; 29: 99–108.
62 Nanda T, Das M, Tripathy K, Ravi Teja Y. Metabolomics: the future of systems biology. J Comput Sci Syst Biol 2011: S13.
63 Horning E C, Horning M G. Metabolic profiles: gas-phase methods for analysis of metabolites. Clin Chem 1971; 17: 802–809.
64 Laiakis E C, Morris G A, Fornace A J, Howie S R. Metabolomic analysis in severe childhood pneumonia in the Gambia, West Africa: findings from a pilot study. PLoS ONE 2010; 5: e12655.
65 Shin J H, Yang J Y, Jeon B Y, et al. (1)H NMR-based metabolomic profiling in mice infected with *Mycobacterium tuberculosis*. J Proteome Res 2011; 10: 2238–2247.
66 Somashekar B S, Amin A G, Rithner C D, et al. Metabolic profiling of lung granuloma in *Mycobacterium tuberculosis* infected guinea pigs: ex vivo 1H magic angle spinning NMR studies. J Proteome Res 2011; 10: 4186–4195.
67 Weiner J III, Parida S K, Maertzdorf J, et al. Discrimination between latent infection and active tuberculosis by metabolomic serum signatures. PLoS ONE 2012. [In press].
68 Banday K M, Pasikanti K K, Chan E C, et al. Use of urine volatile organic compounds to discriminate tuberculosis patients from healthy subjects. Anal Chem 2011; 83: 5526–5534.
69 Phillips M, Basa-Dalay V, Bothamley G, et al. Breath biomarkers of active pulmonary tuberculosis. Tuberculosis (Edinb) 2010; 90: 145–151.
70 Turner A P, Magan N. Electronic noses and disease diagnostics. Nat Rev Microbiol 2004; 2: 161–166.
71 Mgode G F, Weetjens B J, Nawrath T, et al. Diagnosis of tuberculosis by trained African giant pouched rats and confounding impact of pathogens and microflora of the respiratory tract. J Clin Microbiol 2012; 50: 274–280.
72 Nawrath T, Mgode G F, Weetjens B J, Kaufmann S H E, Schulz S. The volatiles of pathogenic and nonpathogenic mycobacteria. Beilstein J Org Chem 2012; 8: 290–299.
73 Le Cao K A, Gonzalez I, Dejean S. integrOmics: an *R* package to unravel relationships between two omics datasets. Bioinformatics 2009; 25: 2855–2856.
74 Hahne F, Huber W, Gentleman R, Falcon S. Bioconductor case studies. Berlin, Germany: Springer Verlag, 2008.
75 Gentleman R C, Carey V J, Bates D M, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5: R80.
76 Diaz-Uriarte R, Alvarez de A S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006; 7: 3.
77 Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol 2007; 3: 78.
78 Chaussabel D, Quinn C, Shen J, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity 2008; 29: 150–164.
79 Stuart J M, Segal E, Koller D, Kim S K. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003; 302: 249–255.
80 Nayak R R, Kearns M, Spielman R S, Cheung V G. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. Genome Res 2009; 19: 1953–1962.
81 Obayashi T, Kinoshita K. COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res 2011; 39 (Database issue): D1016–D1022.
82 Okuda S, Yamada T, Hamajima M, et al. KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res 2008; 36 (Web Server issue): W423–W426.
83 Wishart D S, Knox C, Guo A C, et al. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 2009; 37 (Database issue): D603–D610.
84 Duarte N C, Becker S A, Jamshidi N, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 2007; 104: 1777–1782.
85 Blanchet L, Smolinska A, Attali A, et al. Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. BMC Bioinformatics 2011; 12: 254.
86 Khatri P, Sirota M, Butte A J. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 2012; 8: e1002375.
87 Ioannidis J P, Khoury M J. Improving validation practices in 'omics' research. Science 2011; 334: 1230–1232.
88 Gideon H P, Flynn J L. Latent tuberculosis: what the host 'sees'? Immunol Res 2011; 50: 202–212.

RÉSUMÉ

L'accélération de la lutte contre la tuberculose (TB) exige une amélioration des méthodes de lutte. Les marqueurs biologiques qui diagnostiquent en toute confiance la TB active et même prédisent le risque de progression de la maladie chez les individus pourraient faciliter un diagnostic et un traitement rapide des patients TB et permettre des mesures préventives chez les individus en état d'infection latente qui comportent un risque élevé de TB. De plus, les marqueurs biologiques pourraient accélérer les essais cliniques au moyen de médicaments novateurs et de candidats-vaccin. On décrira trois plateformes de profil des marqueurs biologiques globaux en insistant sur les réalisations les plus récentes : transcriptomiques, protéomiques et métabolomiques. De plus, nous discuterons de la nécessité d'analyses computées pour faire le meilleur usage de la pléthore de données générées par la recherche sur les marqueurs bioloqiques. A côté de leur valeur potentielle en matière de pronostic et de diagnostic, les marqueurs biologiques pourraient fournir des vues plus approfondies sur les processus pathologiques qui sous-tendent la maladie et, dès lors, être à la base de mesures d'interventions novatrices visant les molécules et les voies de l'hôte. Nous proposons que les signatures biologiques qui discriminent une TB active à la fois du statut d'infection latente et du statut de non-infection, ainsi que d'autres maladies, deviennent disponibles au cours de la prochaine décennie. Toutefois, un diagnostic simple et peu couteux par les marqueurs biologiques sur les sites de soins ne sera probablement pas réalisé dans les quelques prochaines années.

RESUMEN

Un control acelerado de la tuberculosis (TB) exige mejores medidas de control. El estudio de los biomarcadores que diagnostiquen de manera fiable la TB activa o que incluso pronostiquen el riesgo de progresión de la enfermedad podrían facilitar un diagnóstico rápido y el tratamiento oportuno de los pacientes y facilitarían además las medidas preventivas dirigidas a las personas con infección tuberculosa latente reciente, que presentan un alto riesgo de progresión hacia la enfermedad. Los biomarcadores también podrían acelerar los ensayos clínicos con nuevos medicamentos y vacunas experimentales. En el presente artículo se describen tres plataformas de caracterización de biomarcadores generales, con hincapié en los progresos más recientes en la transcriptómica, la proteinómica y la metabolómica. Asimismo, se analiza la necesidad de análisis computacionales que favorezcan el uso óptimo de la plétora de datos generados por la investigación de los biomarcadores. Además de sus posibilidades pronósticas y su valor diagnóstico, los biomarcadores podrían permitir una comprensión integral de los mecanismos patológicos fundamentales de la enfermedad y constituir la base de nuevas medidas de intervención dirigidas a las moléculas y a las vías de respuesta del huésped. Se propone que en el próximo decenio se contará con bioseñales que distingan la TB activa de la infección latente y la ausencia de infección y también de otras enfermedades. Sin embargo, es probable que en los próximos años no se logre el desarrollo de métodos diagnósticos utilizables en el lugar de la consulta, basados en biomarcadores sencillos y de bajo costo.