International Journal of Infectious Diseases 56 (2017) 258–262

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Review

High-throughput and computational approaches for diagnostic and prognostic host tuberculosis biomarkers

January Weiner *, Stefan H.E. Kaufmann

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

ARTICLE INFO

Article history:
Received 20 October 2016
Accepted 20 October 2016

Corresponding Editor: Eskild Petersen, Aarhus, Denmark

Keywords:
Tuberculosis
Biomarker
High-throughput
Gene set enrichment analysis

SUMMARY

High-throughput techniques strive to identify new biomarkers that will be useful for the diagnosis, treatment, and prevention of tuberculosis (TB). However, their analysis and interpretation pose considerable challenges. Recent developments in the high-throughput detection of host biomarkers in TB are reported in this review.

1. Introduction

There is a widely acknowledged need for novel biomarkers of tuberculosis (TB), for all levels of TB diagnosis, treatment, and prevention.[1–3] Much effort in this direction has been devoted to host biomarkers, because progression towards clinical disease can be detected by specific changes that are evoked by the pathological processes in the host organism. In TB, this is a particular advantage, as the diagnosis of direct symptoms of TB (e.g., by auscultation or chest X-ray (CXR), through detection of the causative agent, acid-fast bacteria in sputum by microscopy, or positive bacterial cultures) may sometimes be problematic. Sputum samples are difficult to obtain from neonates, who moreover frequently suffer from extrapulmonary TB.

However, there are further reasons to focus on the host response. The onset of active TB disease is frequently delayed for years, and the time span between the first TB symptoms and diagnosis has been estimated to range from 5 days to as long as 162 days.[4] Thus, TB may exist without apparent symptoms, although the molecular processes underlying TB pathology have already commenced. Likewise, TB may persist in a subclinical stage after drug treatment and may later relapse. Positron emission computed tomography (PET/CT) has revealed hallmarks of active TB in patients who have been treated successfully.[5] Host biomarkers may provide a sensitive and specific approach to detect subclinical manifestations of clinical or subclinical TB.

The early detection of TB is another important area for biomarker research. Of two billion *Mycobacterium tuberculosis*-infected individuals, most remain healthy but infected (latent TB infection, LTBI) and only a fraction of 5–7% will develop clinical TB during their lifetime. Although *M. tuberculosis* infection can be determined reliably by interferon gamma release assay (IGRA), this test cannot be used to diagnose or determine the prognosis of active TB.[6] Thus, the identification of biomarkers of TB risk and early stage of progression to active TB would allow screening for individuals at risk. This would allow preventive drug therapy, and also interruption of transmission, with a marked influence on treatment success. Practically, the treatment outcome cannot be assessed in a point-of-care setting. Although PET/CT has predictive value for the treatment outcome,[7] simpler and more accessible tests have thus far failed. For example, although CXR allows a reliable diagnosis of TB, it has limited predictive value for the treatment outcome.[8] Early and personalized treatment adjustment, as well as prediction of the treatment outcome in new drug trials is a major concern in the face of increasing incidences of drug-resistant TB.

2. Computational approaches to high-throughput biomarkers

High-throughput techniques such as transcriptomics allow the inspection of tens of thousands of variables (such as gene expression, protein or metabolite levels) in one step (A glossary

* Corresponding author.
E-mail address: january@mpiib-berlin.mpg.de (J. Weiner).

http://dx.doi.org/10.1016/j.ijid.2016.10.017

Table 1
Glossary

Biomarker	A measurable indicator of the organism state.
Signature	A set of individual biomarkers, corresponding values, and specific machine learning models, which act together as an indicator of the state of an organism.
Predictive vs. prospective	Biomarkers that allow the prediction of the likely natural course of the untreated disease in the individual are termed 'prospective biomarkers'. Biomarkers that allow the prediction of the outcome of treatment are termed 'predictive biomarkers'.
Machine learning (ML)	Methods in computer science that allow the construction of a model of reality based on automatic inspection of data. In 'supervised ML', a model of reality is first derived from a training data set, and subsequently validated by application to a test data set. For example, a model can be trained on gene expression data from TB patients and healthy controls. Its performance will then be evaluated by applying the model to a separate validation set.
ROC curve	A curve describing the predictive ability of a supervised ML model, showing all possible combinations of specificity and sensitivity that can be obtained from that model.
Random forests	A type of supervised ML in which a large number of partially randomized decision trees is generated. When applied to a sample, each tree casts a vote, and the model then decides on the classification of the sample by majority rule.
Variable importance	A measure that determines the relative importance of different variables for correctly classifying a sample by a machine learning model.
Gene set enrichment analysis	Genes (or other variables) can be grouped into functional categories such as gene ontology sets, co-expression modules, or sets of genes that are up- or down-regulated in a particular condition or are specific for a given cell subtype. Gene set enrichment analysis can take advantage of such a classification by testing whether a particular category of genes (e.g., interferon inducible genes or monocyte surface proteins) are enriched in genes that are strongly regulated in a given comparison (e.g., TB vs. healthy controls).

of the terms used in this article is given in Table 1). However, the large number of variables (compared to the number of samples analysed) is a two-edged sword. The obvious advantage of such an approach is the comparatively unbiased acquisition of a large number of potential candidates. On the other hand, if the number of variables is much larger than the number of samples utilized, sophisticated and careful statistical analyses are necessary. Most importantly, the statistical power for detecting a single or a few suitable biomarkers amongst the thousands of variables analysed decreases profoundly, thus correct signals are often hidden in a deluge of false-positives. Moreover, given that the number of functionally characterized protein-coding genes remains insufficient, and only a few microRNAs have been functionally characterized, the interpretation of results may pose an additional obstacle.

Data mining tools such as supervised and unsupervised learning have been employed successfully in a number of biomarker studies.[2,9] Supervised machine learning algorithms include both established methods (such as linear discrimination analysis,[10] k-nearest neighbour algorithm,[11] and random forests[12,13]) and novel, unique approaches. Zak et al.[14] constructed a new classification method by combining k-top-scoring pairs[15] with support vector machines (SVMs), taking advantage of relatively simple interpretability of the k-top-scoring pairs approach with the flexibility of SVMs. Kaforou et al. defined a new metrics termed the 'disease risk score' (DSR), defined as the sum of signed absolute intensities of discriminatory biomarkers, combined with a TB/no TB threshold.[16] Despite the computational simplicity of DSR, it was shown to perform well in discriminating TB patients both from healthy individuals and from patients suffering from other diseases.

A disease signature is only superficially a compilation of variables (e.g., genes) that differ between two conditions. Firstly, as a minimum these variables are linked to particular values (e.g., gene expression in healthy individuals and in TB patients, as in the k-nearest neighbour algorithm) or more complex structures (e.g., decision trees). Secondly, most machine learning algorithms provide a score, which subsequently is compared to an arbitrarily chosen threshold. This latter step, however, depends on a given context, because modifying the threshold optimizes either specificity or sensitivity. As a solution, results of such biomarker analyses are frequently shown as so-called receiver operating characteristic (ROC) curves—all possible sensitivity/specificity combinations for a given signature (Figure 1A).

The interpretation of signatures is increasingly confounded by the size and complexity of the model. While biological functions to which a four-gene signature is related may be glimpsed with relative ease, it is much harder to gain an overview in more complex cases. However, machine learning algorithms often allow the calculation of a 'variable importance' (VI) measure. VI can be used to rank genes according to their contribution to the model, which in turn can be used by adapting a gene set enrichment analysis framework such as GSEA[21], piano,[22] or tmod.[20] In the case of a shrunken model based on a subset of genes, the subset itself can be tested for enrichment in relevant classes of genes.

Note that all statistical approaches are based on assumptions, which incompletely fit the biological reality. Moreover, the large number of variables tested in a high-throughput setting increases the risk of false-positives, even when strictly adhering to standards in statistical methodology, e.g. by using a suitable method for family-wise error correction. It has been estimated that at $p < 0.05$, as many as 30% of the rejected hypotheses may be false-positives,[23] irrespective of using a correction for multiple testing, which may be one of the reasons for the much debated 'reproducibility crisis' in science. The point here is that high-throughput analyses are especially vulnerable to these problems.

Three not mutually exclusive approaches are suggested here, which do not require additional statistical assumptions or novel techniques. Firstly, because unblinded studies overestimate the actual observed effect size,[24] any biomarker study in future should consider separating ('locking') a randomly chosen subset of samples for a blinded, post-hoc validation of the findings, and studies should be evaluated by adherence to this rule. Secondly, an independent analysis by several statisticians (both as study authors and reviewers) would greatly increase confidence in the findings. Thirdly, biomarker studies need to be validated in various settings and cohorts, and using independent experimental approaches. This would facilitate the process of translating the high-throughput to practical clinical applications.

3. High-throughput biomarkers in TB

3.1. Transcriptomic profiling

High-throughput-derived transcriptomic biomarkers have been studied for almost a decade in TB, with the first studies appearing in 2007.[2,10,25] The broadly studied differences in gene expression between TB patients and healthy (infected or uninfected) controls thus far have been investigated in a total of over a thousand individuals on four continents. Kaforou et al. included over 500 individuals in two cohorts, not only TB patients and healthy controls (both HIV-negative and HIV-positive), but also