EBioMedicine xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Unbiased Identification of Blood-based Biomarkers for Pulmonary Tuberculosis by Modeling and Mining Molecular Interaction Networks

Awanti Sambarey [a], Abhinandan Devaprasad [a,1], Abhilash Mohan [a,1], Asma Ahmed [b], Soumya Nayak [b], Soumya Swaminathan [c], George D'Souza [d], Anto Jesuraj [d], Chirag Dhar [d], Subash Babu [e], Annapurna Vyakarnam [b,f], Nagasuma Chandra [a,*]

[a] Department of Biochemistry, IISc, Bangalore 560012, India
[b] Centre for Infectious Disease Research (CIDR), IISc, Bangalore 560012, India
[c] National Institute for Research in Tuberculosis, Mayor Sathiyamoorthy Road, Chetpet, Chennai 600031, India
[d] St John's Research Institute, St. John's National Academy of Health Sciences, 560034 Bangalore, India
[e] NIH-NIRT-ICER, Mayor Sathiyamoorthy Road, Chetpet, Chennai 600031, India
[f] Department of Infectious Diseases, King's College London School of Medicine, Guy's Hospital, Great Maze Pond, London, UK

ARTICLE INFO

Article history:
Received 23 September 2016
Received in revised form 16 December 2016
Accepted 16 December 2016
Available online xxxx

Keywords:
Tuberculosis
Biomarkers
Network biology
Computational medicine
Diagnostics

ABSTRACT

Efficient diagnosis of tuberculosis (TB) is met with multiple challenges, calling for a shift of focus from pathogen-centric diagnostics towards identification of host-based multi-marker signatures. Transcriptomics offer a list of differentially expressed genes, but cannot by itself identify the most influential contributors to the disease phenotype. Here, we describe a computational pipeline that adopts an unbiased approach to identify a biomarker signature. Data from RNA sequencing from whole blood samples of TB patients were integrated with a curated genome-wide molecular interaction network, from which we obtain a comprehensive perspective of variations that occur in the host due to TB. We then implement a sensitive network mining method to shortlist gene candidates that are most central to the disease alterations. We then apply a series of filters that include applicability to multiple publicly available datasets as well as additional validation on independent patient samples, and identify a signature comprising 10 genes — *FCGR1A*, *HK3*, *RAB13*, *RBBP8*, *IFI44L*, *TIMM10*, *BCL6*, *SMARCD3*, *CYP4F3* and *SLPI*, that can discriminate between TB and healthy controls as well as distinguish TB from latent tuberculosis and HIV in most cases. The signature has the potential to serve as a diagnostic marker of TB.

1. Introduction

Tuberculosis (TB) now ranks along with HIV as the leading cause of death due to an infectious agent worldwide, with approximately 10.4 million people estimated to have acquired TB in 2015, resulting in 1.4 million deaths (World Health Organization, 2016). These deaths are largely preventable by early and efficient diagnosis of the disease. Unfortunately, diagnosis is often delayed due to insensitive and time-consuming methods. Present diagnostic measures rely largely on the detection of *Mtb* in patient samples together with radiological assessments, and they have several shortcomings. Sputum cultures are the current standard for detecting *Mtb*, but while sensitive, they take 3–6 weeks to provide conclusive results, thereby delaying the initiation of treatment. Host-based diagnostic methods provide an alternative for early detection of TB onset and enable the monitoring of symptomatic changes. IFN-γ release assays (IGRAs) such as the T-SPOT.TB (Richeldi, 2006; Pai et al., 2014) or the QuantiFERON test (Sultan et al., 2010) measure IFN-γ+ production in response to stimulation with *Mtb*-specific antigens ESAT6 and CFP10 (Mazurek and Villarino, 2003; Ravn et al., 2005). However, IGRAs cannot discriminate between active and latent *Mtb* infection, and are thus inadequate for marking the disease status. In the clinic, IGRAs are used more often to detect latent tuberculosis than for diagnosis of active disease (Herrera et al., 2011). Existing assays that rely on single-marker readouts, such as that of serum deaminase levels (Gui and Xiao, 2014), also suffer from inadequate sensitivity and/or specificity, calling for more effective host-related multi-marker signatures that hold promise for applications in prognostic research and vaccine trials as well as in monitoring treatment responses. There is thus a current need for a shift from investigations on single markers to high-coverage studies that will reveal signatures consisting of multiple integrated markers (Maertzdorf et al., 2014). Recent years have witnessed an increase in host *omics* data

* Corresponding author.
E-mail address: nchandra@biochem.iisc.ernet.in (N. Chandra).
[1] Joint second authors.

http://dx.doi.org/10.1016/j.ebiom.2016.12.009

to identify specific gene variations upon infection with *Mtb*, including genetic polymorphisms identified by GWAS and linkage and association studies that ascribe host susceptibility to infection (Azad et al., 2012), genome-wide expression variations in patient cohorts as compared to healthy controls, as well as variations over the course of treatment in the same patient.

Transcriptomics provide global coverage into host responses, and are widely used in TB biomarker research (Maertzdorf et al., 2011a; Joosten et al., 2013). 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). RNA sequencing (RNA-Seq) is fast gaining foothold, and provides more accurate measurements of transcript levels and their isoforms with greater sensitivity than microarrays, as it overcomes probe-dependency (Wang et al., 2009). RNA-seq has been applied to study host variations due to mycobacterial infections and has led to rich insights, an example being dual RNA sequencing of host and pathogen in *Mtb* infected Thp-1 cells that indicated a simultaneous induction of *Mycobacterium bovis* BCG cholesterol degradation genes and a compensatory upregulation in the host de novo cholesterol biosynthesis genes (Rienksma et al., 2015). Recently, a whole blood signature that could predict the risk of developing active tuberculosis in patients with latent infection was identified by RNA-seq data (Zak et al., 2016).

Although the immunological response against *Mtb* will be primarily focused in the lung, its pathologic status is reflected in the peripheral blood by circulating immune cells (Weiner et al., 2013). Whole blood transcriptomic profiles provide global insights into host immune responses in tuberculosis and serve as essential tools in determining underlying molecular players of infection.

A multi-marker set of gene classifiers determined from blood transcriptomic data with sufficient discriminatory prowess would thus support current diagnostic measures to enhance early detection of TB in the clinic (Cliff et al., 2015). Gene expression values highlight differentially expressed genes (DEGs), which by themselves are indicative of the variations in disease, but further selection is required to identify a small biomarker set that is characteristic of the disease. Such a selection has been achieved using machine learning methods for a number of diseases including tuberculosis (Blankley et al., 2014). Use of networks, however, provides a different perspective to identify DEGs that may be functionally linked to other differentially regulated genes, either directly or indirectly through other bridging nodes. A systems approach integrating transcriptomic data and genome-wide molecular interaction networks is necessary to provide mechanistic insights into the nature of dynamic responses to infection and help identify the most significant contributors to the disease phenotype. Biological network analysis involves the construction of a pair-wise assembly of molecular interactions among cellular components that will yield a connected network of interactions. The network can be compared to a street-map of a city and provides an overview of the interconnected routes or in other words the topological architecture of the molecular interactions in a cell. Mapping genome-wide expression profiles into molecular networks to construct condition-specific response networks provides an unbiased systematic approach to enable the identification of combinations of host components that can serve as markers for tuberculosis and aid early diagnosis.

India currently leads the world's burden of tuberculosis, accounting for about 2.8 million cases out of the global incidence of 10 million (World Health Organization, 2016). *Omics* studies on the Indian population have been few and far between. In this study, with an aim to differentiate pulmonary tuberculosis from other conditions, we use a new network-based pipeline for biomarker discovery. We obtain RNAseq data from an Indian cohort and map them onto interaction networks, from which we identify the most influential genes in the host whole blood response network to tuberculosis. We then apply a series of filters to finally discover a 10-gene validated signature that can discriminate TB samples and healthy controls in multiple cohorts from different geographical locations and also discriminate between latent and active tuberculosis. In addition, the signature distinguishes between TB and HIV and has a potential to be used for diagnosis in the clinical setting.

2. Materials and Methods

2.1. Study Participants

Clinical samples were obtained from participants enrolled at the National Institute for Research in Tuberculosis (NIRT), Chennai: active TB (BL), IGRA − ve/healthy control (HC), and IGRA + ve/latent TB (LTB); St. John's Research Institute, Bengaluru (IGRA − ve/healthy controls and HIV +) and Arogyavaram Medical Centre, Madanapalle (IGRA + -ve/latent TB). Patients attending the outpatient clinics of NIRT and community controls were enrolled for this study. This was a prospective case control study and we enrolled consecutive patients and controls. The diagnosis of pulmonary tuberculosis (TB) was based on smear and culture positivity. Chest X-rays were used to define cavitary disease as well as unilateral vs bilateral involvement. Smear grades were used to determine bacterial burdens and classified as 1 +, 2 + and 3 +. At the time of enrolment, all active TB cases had no record of prior TB disease or anti-tuberculosis treatment (ATT). Latent tuberculosis (LTB) diagnosis was based on tuberculin skin test (TST) and QuantiFERON TB-Gold in Tube ELISA positivity, absence of chest radiograph abnormalities or pulmonary symptoms. A positive TST result was defined as an induration at the site of tuberculin inoculation of at least 12 mm in diameter to minimize false positivity due to exposure to environmental mycobacteria. NTB individuals were asymptomatic with normal chest X-rays, negative TST (indurations < 5 mm in diameter) and QuantiFERON ELISA results. All participants were BCG vaccinated, HIV negative, non-diabetic and had normal body mass index. All participants did not exhibit signs or symptoms of any associated lung or systemic disease. Standard anti-TB treatment (ATT) was administered to TB individuals using the directly observed treatment, short course (DOTS) strategy. At 6 months following ATT initiation, fresh plasma samples were obtained. All TB individuals were culture negative at the end of ATT. All individuals were examined as part of a study protocol approved by the 'Internal Ethics Committee' of NIRT and written informed consent was obtained from all participants (approval number NIRTIEC2010002). Table 1 describes the breakdown of different patient classes. Clinical details of all enrolled participants are provided in Additional File 2. Samples for RNA sequencing were exclusively from NIRT, Chennai. A total of three samples from the IGRA − ve and IGRA + ve categories and 4 from the active TB category were used for RNA sequencing. An additional 13 active TB, 10 IGRA − ve, 9 IGRA + ve and 7 HIV + samples were used for validation of gene expression by qRT-PCR. Classification of IGRA − ve and IGRA + ve individuals was done on the basis of a QuantiFERON assay.

2.2. RNA Isolation

Blood (3 ml) from each participant was collected in a Tempus tube, vigorously shaken and transported to IISc, Bangalore, where it was stored at −80 °C till use. For RNA isolation, frozen Tempus tubes were thawed and RNA was extracted using a Tempus Spin RNA isolation kit (Applied Biosystems) following the manufacturer's instructions. Briefly, blood from the Tempus tube was centrifuged at 3000 *g* for 30 min at 4 °C to pellet down the RNA which was then re-suspended and loaded onto a spin column for purification. RNA bound to the column was eluted and aliquoted RNA was quantified and either subjected to RNA sequencing or converted to cDNA for gene expression studies by qRT-PCR.

RNA sequencing RNA isolated from Tempus tubes was quantified and subjected to quality control analysis. RNA samples with a RIN > 5 were taken further for RNA sequencing. Library preparation was performed at Genotypic Technology's Genomics facility at Bangalore. Five

Please cite this article as: Sambarey, A., et al., Unbiased Identification of Blood-based Biomarkers for Pulmonary Tuberculosis by Modeling and Mining Molecular Interaction Networks, EBioMedicine (2016), http://dx.doi.org/10.1016/j.ebiom.2016.12.009