International Journal of Infectious Diseases 56 (2017) 253–257

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

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

Review

A tuberculosis biomarker database: the key to novel TB diagnostics

Seda Yerlikaya[a,1], Tobias Broger[a,*,1], Emily MacLean[b], Madhukar Pai[b,c], Claudia M. Denkinger[a]

[a] FIND, Chemin des Mines 9, CH-1202 Geneva, Switzerland
[b] McGill International TB Centre, Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[c] McGill Global Health Programs, McGill University, Montreal, QC, Canada

ARTICLE INFO

Article history:
Received 10 October 2016
Received in revised form 18 January 2017
Accepted 22 January 2017
Corresponding Editor: Eskild Petersen, Aarhus, Denmark

Keywords:
Tuberculosis
Biomarkers
Database
Pipeline

SUMMARY

New diagnostic innovations for tuberculosis (TB), including point-of-care solutions, are critical to reach the goals of the End TB Strategy. However, despite decades of research, numerous reports on new biomarker candidates, and significant investment, no well-performing, simple and rapid TB diagnostic test is yet available on the market, and the search for accurate, non-DNA biomarkers remains a priority. To help overcome this 'biomarker pipeline problem', FIND and partners are working on the development of a well-curated and user-friendly TB biomarker database. The web-based database will enable the dynamic tracking of evidence surrounding biomarker candidates in relation to target product profiles (TPPs) for needed TB diagnostics. It will be able to accommodate raw datasets and facilitate the verification of promising biomarker candidates and the identification of novel biomarker combinations. As such, the database will simplify data and knowledge sharing, empower collaboration, help in the coordination of efforts and allocation of resources, streamline the verification and validation of biomarker candidates, and ultimately lead to an accelerated translation into clinically useful tools.

Contents

Introduction

"Scarcely four years have elapsed since the important discovery of the tubercle-bacillus by Koch was announced. Many then thought that the key to the various problems of pulmonary consumption was close at hand, if not in our actual possession" W N, 1886.[1]

More than a century after Koch's discovery and the hopes that answers would quickly follow,[1] tuberculosis (TB) continues to kill 4000 people per day,[2] and we are still searching for 'the key' as envisioned in the 1880s. An array of diagnostic and treatment solutions is required to control and prevent the complex medical and socio-economic problems caused by TB. In countries experiencing the worst TB epidemics, i.e., high-burden, low- and middle income settings, the continuing dependence on slow diagnostic tools with limited performance allows epidemics to persist. Despite the advent of molecular diagnostic tests such as Xpert MTB/RIF,[3] limited access and affordability prevent people with TB symptoms from accessing services. Within this context,

* Corresponding author.
E-mail address: tobias.broger@finddx.org (T. Broger).
[1] Seda Yerlikaya and Tobias Broger contributed equally to this work.

http://dx.doi.org/10.1016/j.ijid.2017.01.025

intensified research and innovation towards the discovery, development, and rapid implementation of new diagnostic tools have been identified as important components of the End TB Strategy of the World Health Organization (WHO).[4,5]

Recent efforts to identify the highest priorities in the field of TB diagnostics have revealed the urgent need for biomarker-based assays that will enable more efficient, affordable, and accessible diagnosis for those in need.[6,7] Table 1 lists the target product profiles (TPPs) that will likely rely on non-DNA biomarkers. The highest priority is a rapid biomarker-based, non-sputum test for detecting active TB with the purpose of initiating treatment.[6,7] The second highest priority is a triage test, also probably biomarker-based and with high sensitivity, that can rule out disease and be used to refer patients to the more expensive and accurate molecular testing for confirmation.

Unfortunately, despite decades of research, significant investment, and numerous reports on new biomarker candidates, few biomarkers have been independently validated for specific use cases and translated into new diagnostic tests.[8,9] This problem is not unique to TB; it is true for biomarker research in general, with very few of the biomarkers discovered having advanced to the clinic in the form of approved diagnostic tests.[10–12] Preliminary data from our ongoing systematic review of biomarker studies reporting on the detection of active TB confirm this lack of validation: for the majority of biomarkers ($n = 399$), diagnostic performance is not reported (161 biomarkers), or is based on testing of a non-blinded, usually retrospective set of conveniently obtained samples (170 biomarkers), or on blinded testing in a single study (68 biomarkers) (Figure 1). Only 12 biomarkers have been confirmed in prospectively designed studies and, to date, only one urine biomarker-based test has been endorsed by the WHO (Determine LAM; Alere, Waltham, MA, USA);[13] however, none of the biomarkers identified has so far led to a diagnostic test that meets the performance requirements of any TPP.

Key issues that limit the impact and translation of biomarker research include: (1) a lack of coordination of similar research activities and limited knowledge-sharing between researchers; (2) an often limited assessment of a biomarker to one or two exploratory studies and a lack of well-designed validation studies; (3) the lack of standards and frameworks for biomarker validation, as well as generally low reporting quality; (4) the failure of many studies to clearly articulate the intended use case and benchmark a biomarker towards it; and (5) optimism and publication bias, which result in a lack of confirmation of initially promising findings. Concerns over intellectual property (IP) rights are common, incentives to share data in the current publishing system

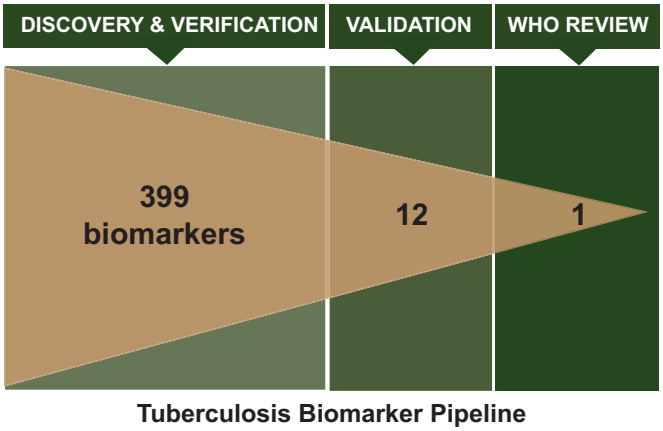

Figure 1. TB biomarker pipeline based on the evidence from 763 studies on TB biomarkers published between 2010 and 2015. A systematic search was employed to find studies in PubMed, Embase, and Web of Science reporting either statistical significance or diagnostic performance of TB biomarkers for the detection of active TB. In total, the 763 included studies reported on 413 (non-DNA) biomarkers or biomarker signatures. The 'discovery and verification' category includes 399 biomarkers for which only statistical significance was reported ($n = 161$), or data were based on testing of a non-blinded sample (usually retrospective set of conveniently obtained samples) ($n = 170$), or performance was based on blinded testing in an initial study ($n = 68$). Only 12 biomarkers were validated in a prospectively designed study in a blinded manner (category 'validation') and only one biomarker-based test has been reviewed by the World Health Organization (lipoarabinomannan in urine, category 'WHO review').

are limited, and easy-to-use tools for in-depth analysis of datasets are unavailable. Moreover, independent validation studies are laborious and costly, as they require larger sample sizes than discovery studies to ensure sufficient statistical power. This discouraging (and expensive) reality prevents researchers from moving from discovery to further stages of development. Additionally, some biomarkers are repeatedly 'discovered' or probed with retrospective, discovery-level studies based on whatever haphazard specimens can be obtained conveniently. This represents an avoidable waste of financial resources as well as patient and researcher time.

Combining resources and evaluating multiple biomarkers side-by-side could be a way to surpass these hurdles, but the lack of communication and coordination among scientists and funding bodies often impedes such possibilities. The end result is the poor translation of biomarkers into urgently needed, fit-for-purpose diagnostic solutions. Poste proposed replacing "this dismal

Table 1
World Health Organization endorsed priority target product profiles (TPP) for the detection of active TB for which non-DNA biomarkers may play a key role.[7]

Priority TPP	Description	Diagnostic sensitivity	Diagnostic specificity
Rapid biomarker-based non-sputum-based test for detecting TB	The majority of pulmonary TB cases are diagnosed by sputum smear microscopy. However, smear microscopy has suboptimal sensitivity, and children and HIV-infected individuals often have difficulties providing a good quality sputum sample. The unmet need is a rapid point-of-care test detecting characteristic biomarkers or biosignatures in non-sputum samples. The requirement is a very high specificity and moderate to high sensitivity for the purpose of initiating treatment.	Minimal: ≥65% overall Optimal: ≥80% overall (≥98% sputum smear-positive and ≥68% in sputum smear-negative patients)	≥98% ≥98%
Community-based triage or referral test for identifying people suspected of having TB	Two weeks of cough is a widely used symptomatic indicator to identify individuals with presumed active pulmonary TB who require diagnostic testing. Since most individuals with suspected TB do not have TB, a triage test can help to narrow down the population that needs more costly and complex confirmatory testing. The needed point-of-care test has a high overall sensitivity and moderate specificity.	Minimal: ≥90% overall Optimal: ≥95% overall	≥70% ≥80%

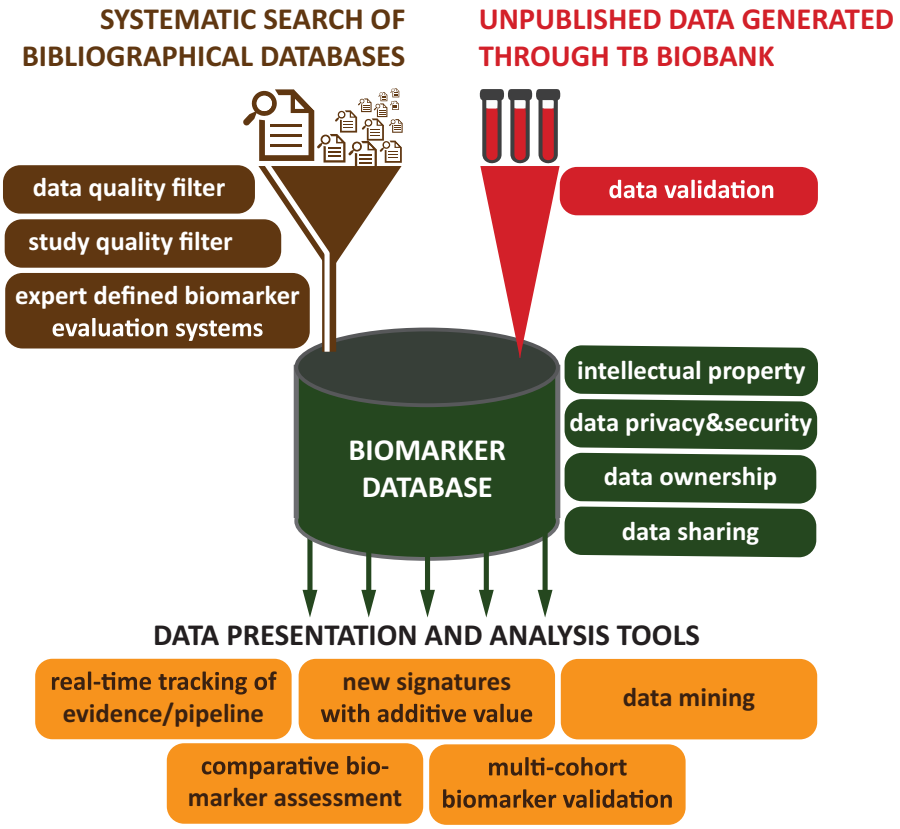

Figure 2. Concept of the biomarker database.

patchwork of fragmented research on disease-associated biomarkers with a coordinated 'big science' approach", which envisions "biomarker discovery and validation as a component of larger research networks", involving industry and research experts.[10]

To support this transition in the field of TB, we have embarked on the development of a curated TB biomarker database for the dynamic tracking of evidence towards clinical needs, the identification of promising biomarkers and biomarker signatures through standardized analysis, and the secured, easily manageable data sharing among researchers (Figure 2). On the data input side, the database will support a systematic search for evidence from the published literature that meets well-defined criteria for inclusion, data and study quality. A manual data curation workflow will ensure sufficient quality. In addition, the database will include results from research and development using banked patient specimens. This link to the FIND biobank will streamline the blinded validation of biomarkers with well-characterized specimens from relevant patient populations and different geographic regions. Ideally, the database will store raw data (biomarker measurements linked to sample meta-data), bibliographical data, and the results of a study. The database will address all relevant aspects of IP, data privacy, data security, data ownership, data use, and data sharing to ensure that the interests of data contributors are respected and, at the same time, facilitate the envisioned data presentation and analysis use-cases.

Dynamic tracking of biomarker evidence in a standardized format

Today, a large number of systematic reviews provide snapshots of biomarker evidence at specific moments in time.[14] However, a more dynamic, real-time method of tracking and synthesizing the rapidly accumulating data is critical to accelerate knowledge synthesis and decision-making. For example, in the case of antibodies as biomarkers for active TB diagnosis, the evidence base was summarized in 2007[15] and last updated in 2011 by Steingart et al.[16] If antibody validation data were continuously updated as new studies were published, updated data synthesis efforts could be made much easier. Additionally, cross-study comparisons, be they between combinations of existing evidence or new evidence, are difficult due to today's static data presentation methods, the reluctance to follow the guidelines for standardized reporting,[17] and the limited availability of raw datasets.

A biomarker database, synthesized and continually updated, could overcome these issues. Classifying biomarkers with standardized definitions and presenting them in a format respecting the widely accepted and recommended reporting guidelines for diagnostic studies (such as STARD[18] or QUADAS-2[19]) would help to pinpoint the differences between diagnostic biomarker studies in a systematic manner, allow for flexible sub-group comparisons, and enable meaningful and significant evaluation of the current status of biomarker candidates towards TPPs. Combined with the efforts of journals that mandate adherence to standard reporting guidelines as a requirement (as in the case of *PLOS Medicine* since 2014[20]), such an initiative would greatly help to boost data quality and trustworthiness.

Increased data and knowledge sharing

Increasing raw data availability is important to better appraise and increase the transparency of evidence on a biomarker candidate. Moreover, data transparency aids communication and understanding between stakeholders. When researchers have access to complete data, they can address new questions, explore different lines of analysis, and conduct large-scale analysis across

studies more efficiently.[21] In order to facilitate data sharing, the appropriate tools need to be in place, especially because it is now well-appreciated that multi-biomarker signatures hold greater promise than single biomarkers for improved diagnostic performance and are less susceptible to confounding factors.[22–24]

Since 2007, FIND has been providing high-quality specimens to researchers worldwide through its TB specimen bank (http://www.finddx.org/specimen-banks/), putting FIND in a unique position to oversee a compendium of datasets. Based on prior experience when different investigators use the same set of samples from a biorepository, FIND has observed that the combination of datasets from two or more investigators hold a lot of promise for the discovery of more predictive signatures. Obviously clear data sharing rules have to be in place to allow for such combined analyses. In addition, the availability of raw datasets for the same biomarker (or biomarker signature) will allow for in silico multi-cohort analyses and biomarker validation without the need for expensive, long-term studies; this concept was recently demonstrated for the validation of global gene expression signatures.[25,26]

At this point, we are empty-handed with respect to user-friendly tools devoted to TB biomarker research that would enable easy access to raw data and facilitate collaboration for the identification of novel biomarker signatures. The development of new platforms to capitalize on the huge potential of biobanks for biomarker identifications is imperative.

This initiative comes at a time when, across disciplines, efforts towards extended and standardized data archiving and sharing are rapidly growing, and journals such as *PLOS* and *eLIFE* are strongly recommending the deposition of all data and related metadata in appropriate public repositories[27] (e.g., Dryad (https://datadryad.org/) or figshare (https://figshare.com/)). These platforms support the standardized archiving and sharing of large (raw) datasets under Creative Commons licenses (https://creativecommons.org/licenses/). However – importantly – these 'generalist' repositories do not provide data synthesis and analysis tools specific to particular fields of research. On the other hand, there are a number of 'specialized' TB databases with a focus on genomic sequences, including mutations associated with drug resistance and gene expression such as The Tuberculosis Database (http://www.tbdb.org/),[28] Tuberculist (http://tuberculist.epfl.ch/),[29] webTB (http://www.webtb.org/), ReSeqTB (https://platform.reseqtb.org/), and TBDReaMDB (https://tbdreamdb.ki.se).[30] The main focus of these databases is molecular testing or TB research, but not non-DNA biomarkers with the potential to address high-priority TPPs.

A data repository integrated within a TB biomarker database would be the first to present raw data from a broad range of published as well as unpublished TB biomarker studies. The reporting and presentation of data would be tailored for the TB field with a clear focus on TPPs. The accessibility of the data would be defined by the researcher, e.g. IP rights are preserved by de-identifying biomarkers and grouping them into classes, data ownership is clearly defined, and access to data is secured in accordance with the permission of the data contributors. Existing platforms such as ReSeqTB exemplify the feasibility of respecting issues such as data access and transfer, privacy, sharing and use, as well as data validation and intellectual property. These points should be addressed and managed carefully to ensure that the interests of data contributors and, even more importantly, patients, are respected. The housing of such a database in a not-for-profit organization with a strong commitment to global public health and experience in managing a clinical database linked to a biobank is paramount to address these critical issues for the long-term credibility and sustainability of such a database.

Finally, the recognition and uptake of a database by its target users is key to its long-term sustainability. Regular manual curation and maintenance of the database to keep it up to date will be crucial for its appeal and utility, but is a demanding task that requires resources. The success of community-curated databases, such as Wikipedia, with added quality checks is a possible avenue to ensure a self-sufficient and sustainable database. Possible funding mechanisms include but are not limited to: (1) voluntary membership fees (such as Dryad users); (2) paid additional services for institutions, publishers, or companies (as is the case with sophisticated analytical tools and statistics such as figshare); (3) donations or sponsorships from industry (such as the Stanford University HIV database); or (4) a tiered user fee-based access model that secures free access for academic researchers and data contributors. To ensure a sufficient user base, it will be important to set up the database in a scalable way so that the 'biomarker pipeline problem' can be tackled for other diseases.

Conclusions

We make the case for a comprehensive, standardized biomarker database to facilitate collaboration and enable in-depth analysis of datasets between multiple studies and cohorts. Access to curated and validated data in an easily accessible format would ease the path from biomarker discovery research to clinical assay development by highlighting the missing links in the development path towards clinical products. Standardized reporting would support and reinforce the communication and coordination among stakeholders. Further, access to large amounts of data would facilitate secondary analysis, paving the way to the discovery of novel multi-biomarker signatures and simplifying biomarker validation. The real-time synthesis and tracking of evidence would enable data-driven decision-making, leading to targeted investments and eventually to a higher translational success rate.

As Hey and Kesselheim accurately state, centralized, publicly accessible biomarker databases are indispensable at this time to move the biomarker diagnostics field forward.[31] FIND and partners started the development of the database and plan to launch a beta version for initial testing mid-2017. We are confident that the TB community will take the lead in supporting such a database dedicated to TB biomarkers. Joint efforts will help to overcome the development pipeline problem and will be critical in delivering robust diagnostic solutions helping to eliminate TB.

Funding

This work was supported by the Ministry of Foreign Affairs, Government of the Netherlands (grant 11255).

Conflict of interest

SY, TB, and CMD are employed by FIND (Geneva, Switzerland), a non-profit organization that collaborates with industry partners. EML has no competing interests. MP serves as a consultant to the Bill & Melinda Gates Foundation, and on the Governing Council of IPAQT. He also serves on the Scientific Advisory Committee of FIND. He has no industry relations or conflicts.

References

1. W.N. The causation of pulmonary consumption. *Science* 1886;**7**(155S):86–8.
2. World Health Organization. *Global tuberculosisreport 2015*. Geneva: WHO; 2015.
3. Albert H, Nathavitharana RR, Isaacs C, Pai M, Denkinger CM, Boehme CC. Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: what lessons have we learnt and how can we do better? *Eur Respir J* 2016;**48**(2):516–25.
4. Stop TB Partnership. *The Global Plan to Stop TB 2016–2020*. Geneva: WHO; 2015 Available at: http://stoptbplan2020.org/wp-content/uploads/2015/06/Global-Plan-to-Stop-TB-2016-2020_Draft-9-June-2015_.pdf (accessed 19.9.16).

5. World Health Organization. *The End TB Strategy*. Geneva: WHO; 2015 Available at: http://www.who.int/tb/strategy/End_TB_Strategy.pdf (accessed 19.9.16).
6. Kik SV, Denkinger CM, Casenghi M, Vadnais C, Pai M. Tuberculosis diagnostics: which target product profiles should be prioritised? *Eur Respir J*. 2014;**44**:537–40.
7. World Health Organization. *High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting*. Geneva: WHO; 2014 Available at: http://apps.who.int/iris/bitstream/10665/135617/1/WHO_HTM_TB_2014.18_eng.pdf (accessed 19.9.16)..
8. Goletti D, Petruccioli E, Joosten SA, Ottenhoff M. Tuberculosis biomarkers: from diagnosis to protection. *Infect Dis Rep* 2016;**8**(2):6568.
9. Gardiner JL, Karp CL. Transformative tools for tackling tuberculosis. *J Exp Med* 2015;**212**(11):1759–69.
10. Poste G. Bring on the biomarkers. *Nature* 2011;**469**:156–7.
11. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* 2006;**24**:971–83.
12. Anderson NL. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002;**1**:845–67.
13. World Health Organization. *The use of lateral flow urine lipoarabinomannan assay (LF-LAM) for the diagnosis and screening of active tuberculosis in people living with HIV*. Geneva: WHO; 2015 Available at: http://www.who.int/tb/publications/use-of-lf-lam-tb-hiv/en/ (accessed 02.10.16).
14. Nicolau I, Ling D, Tian L, Lienhardt C, Pai M. Research questions and priorities for tuberculosis: a survey of published systematic reviews and meta-analyses. *PLoS One* 2012;**7**:e42479.
15. Steingart KR, Henry M, Laal S, Hopewell PC, Ramsay A, Menzies D, et al. Commercial serological antibody detection tests for the diagnosis of pulmonary tuberculosis: a systematic review. *PLoS Med* 2007;**4**(6):e202.
16. Steingart KR, Flores LL, Dendukuri N, Schiller I, Laal S, Ramsay A, et al. Commercial serological tests for the diagnosis of active pulmonary and extrapulmonary tuberculosis: an updated systematic review and meta-analysis. *PLoS Med* 2011;**8**(8):e1001062.
17. Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. *PloS One* 2009;**4**(11):e7753.
18. Bossuyt PM. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;**138**:W1.
19. Whiting PF. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;**155**:529.
20. PLOS Medicine Editors. Observational studies: getting clear about transparency. *PLoS Med* 2014;**11**:e1001711.
21. Warren E. Strengthening research through data sharing. *N Engl J Med* 2016;**375**:401–3.
22. Zak DE, Penn-Nicholson A, Scriba TJ, Thompson E, Suliman S, Amon LM, et al. A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *Lancet Lond Engl* 2016;**387**(10035):2312–22.
23. Bloom CI, Graham CM, Berry MPR, Rozakeas F, Redford PS, Wang Y, et al. Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PloS One* 2013;**8**(8):e70630.
24. Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluka G, et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. *N Engl J Med* 2014;**370**(18):1712–23.
25. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. *Lancet Respir Med* 2016;**4**:213–24.
26. Joosten SA, Fletcher HA, Ottenhoff TH. A helicopter perspective on TB biomarkers: pathway and process based analysis of gene expression data provides new insight into TB pathogenesis. *PLoS One* 2013;**8**:e73230.
27. PLOS data availability policy. *PLoS*. Available at: http://journals.plos.org/plosone/s/data-availability (accessed October 3, 2016).
28. Galagan JE, Sisk P, Stolte C, Weiner B, Koehrsen M, Wymore F, et al. TB database 2010: overview and update. *Tuberc Edinb Scotl* 2010;**90**(4):225–35.
29. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList—10 years after. *Tuberculosis (Edinb)* 2011;**91**:1–7.
30. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. *PLoS Med.* 2009;**6**(2):e2.
31. Hey SP, Kesselheim AS. Countering imprecision in precision medicine. *Science* 2016;**353**:448–9.