Stop TB Partnership
New Diagnostics Working Group

Critical Path to TB Drug Regimens

Meeting Report

2nd Workshop

Data sharing to improve management of drug-resistant TB

28 October 2014, Barcelona, Spain

2nd Workshop
Data sharing to improve management of drug-resistant TB

Meeting Report
NDWG, November 2014

Contact: ndwg@finddiagnostics.org

Overview

Event

2nd workshop on enabling standards and sharing of data on the molecular basis of drug resistance
"Data sharing to improve management of drug-resistant TB"

Date

28 October 2014, 17:30-20:30

Venue

Hotel AC Forum, Barcelona, Spain

Organisers

A workshop jointly organized by the Stop TB Partnership's New Diagnostics Working Group (NDWG) and the Critical Path to TB Drug Regimens (CPTR) with support from the Bill & Melinda Gates Foundation (BMGF).

Chairpersons

Daniela Cirillo, San Raffaele Research Institute, NDWG Co-Chair
Jim Gallarda, Bill & Melinda Gates Foundation

Attendees

54 invited experts from different backgrounds were assembled including representation from tuberculosis (TB) sequencing groups, TB reference laboratories, TB clinicians, software developers, test developers, diagnostics industry, the World Health Organization (WHO), US Centers for Disease Control and Prevention (CDC), European Centre for Disease Prevention and Control (ECDC), and non-governmental organizations (full list on p. 15).

Workshop objectives

The workshop was organized as a follow up of the meeting held in London on 3-4 February 2014 and aimed to:

1. Inform the group about recent plans to build a TB sequence database
2. Create consensus about the initiative and advocate for data sharing and integration
3. Initiate discussions to develop an action plan to share data

Meeting report

Paolo Miotto, Fondazione Centro San Raffaele, Italy
Alessandra Varga, FIND, NDWG Secretariat

Presentations

All presentations from the workshop are available online on the NDWG website at:
http://www.stoptb.org/wg/new_diagnostics/datasharing.asp

Agenda

Chairpersons

Daniela Cirillo, San Raffaele Research Institute, NDWG Co-Chair
Jim Gallarda, Bill and Melinda Gates Foundation

Session 1 – Building the bridge from sequencing to the patient **Chairs: David Dolinger and Marco Schito**		
17:30	Welcome and meeting objectives	*Daniela Cirillo*
17:35-17:45	Plans to build a TB sequencing database	*Jim Gallarda*
17:45-17:55	The role of FIND and the NDWG	*Claudia Denkinger*
17:55-18:05	Plan for data evaluation and validation	*David Dolinger*
18:05-18:20	Plenary discussion The strengths of the proposal and why scientists should contribute	*Lead Angela Starks*
Session 2 – The scientists perspective **Chairs: Daniela Cirillo and Jim Gallarda**		
18:35-18:45	Sample size	*Derrick Crook*
18:45-18:55	NGS data complexity: prospects and problems	*Stefan Niemann*
18:55-19:05	Combined effort for sequencing in Russia	*Igor Mokrousov*
19:05-19:15	Involvement of Eastern countries in Broad Institute initiatives	*Maha Farhat*
19:15-19:35	Plenary discussion Identify the barriers to data sharing and possibility to overcome them	*Lead Session Chairs*
Session 3 – Strategy for advocacy for data sharing **Chairs: Rumina Hasan and Nazir Ismail**		
19:35-19:45	Relevance of the project for Countries and Clinicians	*Dalene von Delft*
19:45-19:55	The role of the WHO in the initiative	*Matteo Zignol*
19:55-20:05	Motivation for countries to contribute with data: The experience of Belarus	*Alena Skrahina*
20:05-20:25	Plenary discussion How to proceed with advocacy for data sharing	*Lead Session Chairs*
20:25-20:30	Wrap-up and close of meeting	*Jim Gallarda*
20:30	**Cocktail reception and networking**	

Session 1

Building the bridge from sequencing to the patient

The limits of current available drug susceptibility testing (DST) assays are well known: phenotypic testing requires long time, high skills and biosafety infrastructure; genotypic tests are available for a limited number of drugs and/or require high skills to be performed. Despite several molecular tests are currently in the development pipeline, point-of-care (PoC) assays will be ready in years and the diagnostic scenario is expected to be as summarized in the diagram (*Figure 1*): open polymerase chain reaction (PCR) platforms for a higher (but still limited) number of drugs will be available in the next future, with the possibility to achieve PoC features in the next five years. Simultaneously, sequencing data will provide a novel layer of testing useful for surveillance purposes and for developing new assays with different levels of application (from peripheral laboratories to PoC).

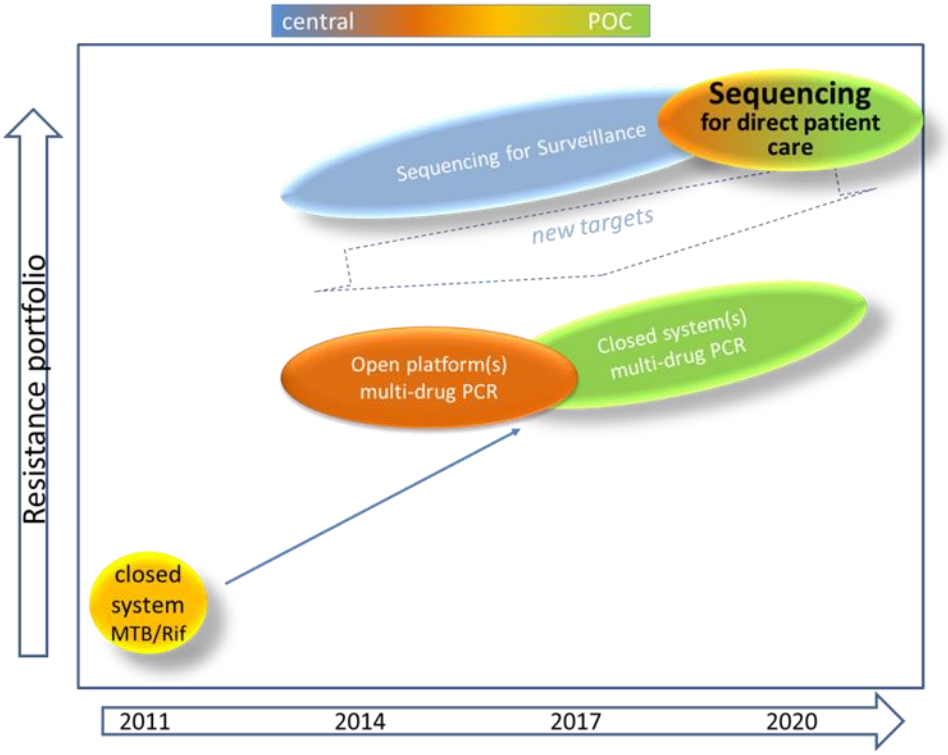

Figure 1

It is important to keep in mind that the development of molecular tests should reflect the progress of new treatment regimens and the discovery of new drugs. To achieve the development of shorter, more tolerable drug regimens and rapid drug susceptibility diagnostics there is the need to:

- Increase our understanding of resistance trends to support rapid DST development
- Support the development of iterative Target Product Profiles (TPP's)
- Develop models that inform TTP's and policies for use of DST
- Facilitate the development of rapid DST assays
- Share of expertise, information and data

It is clear that sharing of robust data represents a basic milestone to fulfill all the other objectives.

The proposed and agreed architecture of the project to develop a *TB drug resistance data sharing platform* is presented in the diagram (*Figure 2*).

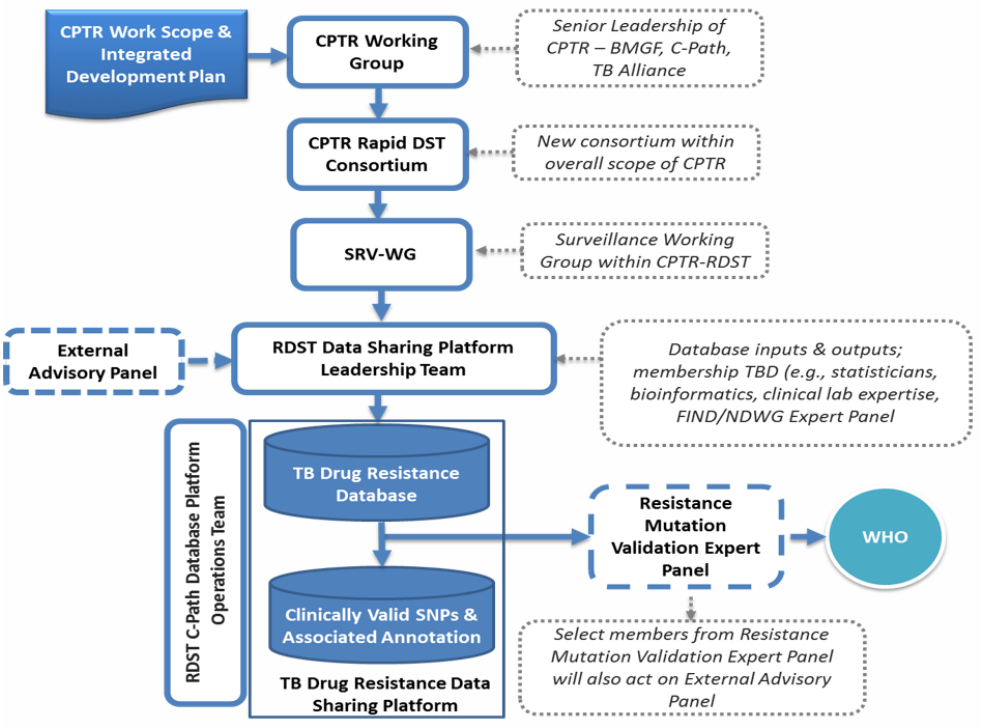

Figure 2

Key principles for a successful global database have been identified as follows:

- Open access: the database should be open access to provide maximum benefit
- Sustainable: the database should be maintained, updated, and easily expanded to include new data *i.e.* from new drugs
- High quality: the database should include robust, quality-controlled data
- Comprehensive: the database should collect data from different geographical regions, representative of the variety of clinical isolates circulating worldwide
- Collaborative: the database should be developed, implemented and maintained through the collaboration between research and clinical groups, policy makers and funders
- Valuable – public health: the database should provide public health benefit to all countries contributing to its development (including countries providing data)
- Valuable – industry: the database should be recognized as an useful tool for industry involved in the development of new diagnostics
- Fair: the database should be developed in agreement with transparency and social justice governance to ensure rights of participating countries (benefits clearly defined).

FIND proposed a two-step approach to achieve “convergence” between drugs and diagnostics. The first step would allow better understanding and elucidating the state-of-the-art on drug resistance-associated mutations. This information would then be used to develop TPPs, to identify high confidence markers of resistance and to inform developers of molecular tests. The second step would consist of enhanced sequencing surveillance programs and of development and strengthen the capacity to use sequencing information to drive patient care.

FIND, the NDWG and the CPTR are strongly committed to promote and develop a global data platform on DST. CPTR represents a strong partner with established capabilities in managing large amounts of data and strongly linked to drug development. The NDWG will represent the linkage between contributors and policy makers by managing the coordination and communication with stakeholders to (i) facilitate data collection and data contribution and (ii) communicate with test developers and policy makers, e.g. WHO.

In addition, FIND and NDWG will support lead experts in the field to:

- Ensure the high quality of the data included in the database
- Drive the development of criteria for the validation of mutations that are associated with resistance
- Create a 'living' list of relevant resistance mutations
- Define algorithms for the interpretation of genotypic data and their correlation with clinically relevant resistance in *M. tuberculosis* (MTB).

A global database for improving management of drug-resistant TB (DR-TB) and for developing new diagnostic tools claims for robust data. Therefore, a plan to drive data evaluation and validation is mandatory. In this process, mutations in MTB should be shown in a systematic and transparent manner to have adequate and objective evidence to either cause or be associated with resistance to a known and identified drug and/or drug class. The goals of this process are:

- the evaluation of existing and prospective data to determine clinically valid resistance-associated mutations
- the definition of consensus around drug susceptibility and genetic resistance determinants
- the roll-out of publications (process transparency, update of mutation panel, validity and acceptance criteria for a mutation to be associated with resistance, supporting data, statistical/quality scores).

The proposed process of validation/evaluation takes advantage of previous experiences in the field of HIV. First of all, the process should be based on consensus. For this reason, a *Resistance Mutation Validation Expert Panel* is needed (i) to develop quality metrics and requirements for data inclusion (genotype, phenotype, metadata), (ii) to develop a data weighting system and (iii) to validate algorithms associating genetic variants to specific (resistant) phenotypes (validity criteria, acceptance criteria). Finally, evidence-based validated genetic variants associated to drug resistance should be endorsed by the WHO.

The establishment of an Expert Panel has been proposed with five core members, and up to ten co-opted members covering representative areas of expertise. The Expert Panel is expected to meet four times per year and would be supervised by FIND and NDWG, which will be responsible for coordinating, and preparing meetings.

A tentative risk analysis has been also presented and three critical challenges (×) together with possible interventions (✓) have been identified:

× Unwillingness of researchers/countries to share data →
  - ✓ Leverage NDWG and the Expert Panel
  - ✓ Inclusion of researchers from geographically diverse regions in the Expert Panel
  - ✓ Utilize additional key opinion leaders to initiate discussion with MoHs
  - ✓ Provide additional tools and algorithms that can be utilized by researchers for data mining

× Diversity of data quality and variety →
  - ✓ Development of quality standards
  - ✓ Development of standard acquisition tools

× Limited data connecting mutations to phenotype and clinical outcomes →
  - ✓ Inform and drive further research investments
  - ✓ Establish a tiered, statistically-driven interpretation taking into account the certainty of knowledge.

Session 2
The scientist perspective

From a genome-based analysis we would expect to have an “all-in-one” analysis: species identification, drug resistance pattern, detection of virulence determinants, genotyping, evolution, population structure and phylogenetic data.

The implementation of whole-genome sequencing in clinical and public health laboratories is substantial, as widespread adoption would require incorporating the knowledge from more than a century of characterizing pathogens — currently delivered by a skilled workforce – into an entirely new framework of mainly computer driven genome processing. This would require a radical shift towards a new operational paradigm for routine laboratories (see *Figure 3*, adapted from Didelot et al. Nat Rev Genet 2012; 13(9):601-612).

Despite the increasing accessibility of next generation sequencing (NGS) bench top systems, interpretation of data is not immediate as needed. Several constraints concern each step of NGS:

- Data generation: different platforms, chemistry approaches and read lengths are available and each of these variables can affect sequencing results. In addition, data quality, coverage and reproducibility should not be underestimated variables.
- Mapping and single-nucleotide polymorphisms (SNPs) read out: mapping of reads and SNPs calling require analysis pipeline, computing power, data storage capacity, accurate reference genomes, filtering criteria and reporting systems. As for the data generation step, each of these variables contributes to the quality of NGS results.
- Data interpretation: at this stage, we are missing standard criteria and nomenclature to interpret NGS data. Often, each laboratory develops its own interpretation rules.

Given this scenario, **data exchange between different laboratories is currently not easy to achieve, and dedicated software tools for non-specialized users are not available.**

Data calculation to achieve sufficient statistical power in a study aiming at associating genetic variants to resistant phenotype should consider the following variables:

- The nature of genetic variants (what are the genetic determinants of resistance?)
- The number of genetic variants (how many different genetic determinants are there for each drug?)
- The frequency of occurrence (how many times should a determinant be phenotyped to gain confidence?)
- The clustering rate (observations clonal samples not very informative; homoplasy could be useful to assess the correlation between mutations and phenotypic resistance).

To test the power of a hypothetical “big” study and to determine “how big” it should be, preliminary data from three groups (labs: Crook, Niemann, Murray) performing whole genome sequencing (WGS) have been taken into account.

These three studies considered more than 9000 globally representative isolates and several drug resistances.

Challenges encountered can be summarized as follow:

- NGS initiatives are largely fragmented and poorly coordinated at national and international level (a clear example was also provided by Mokrousov who presented the MTB genome sequencing activities in Russia)
- Data generation and analysis is not standardized
- Genotype/phenotype correlation is still difficult
- No easy read out systems for relevant data

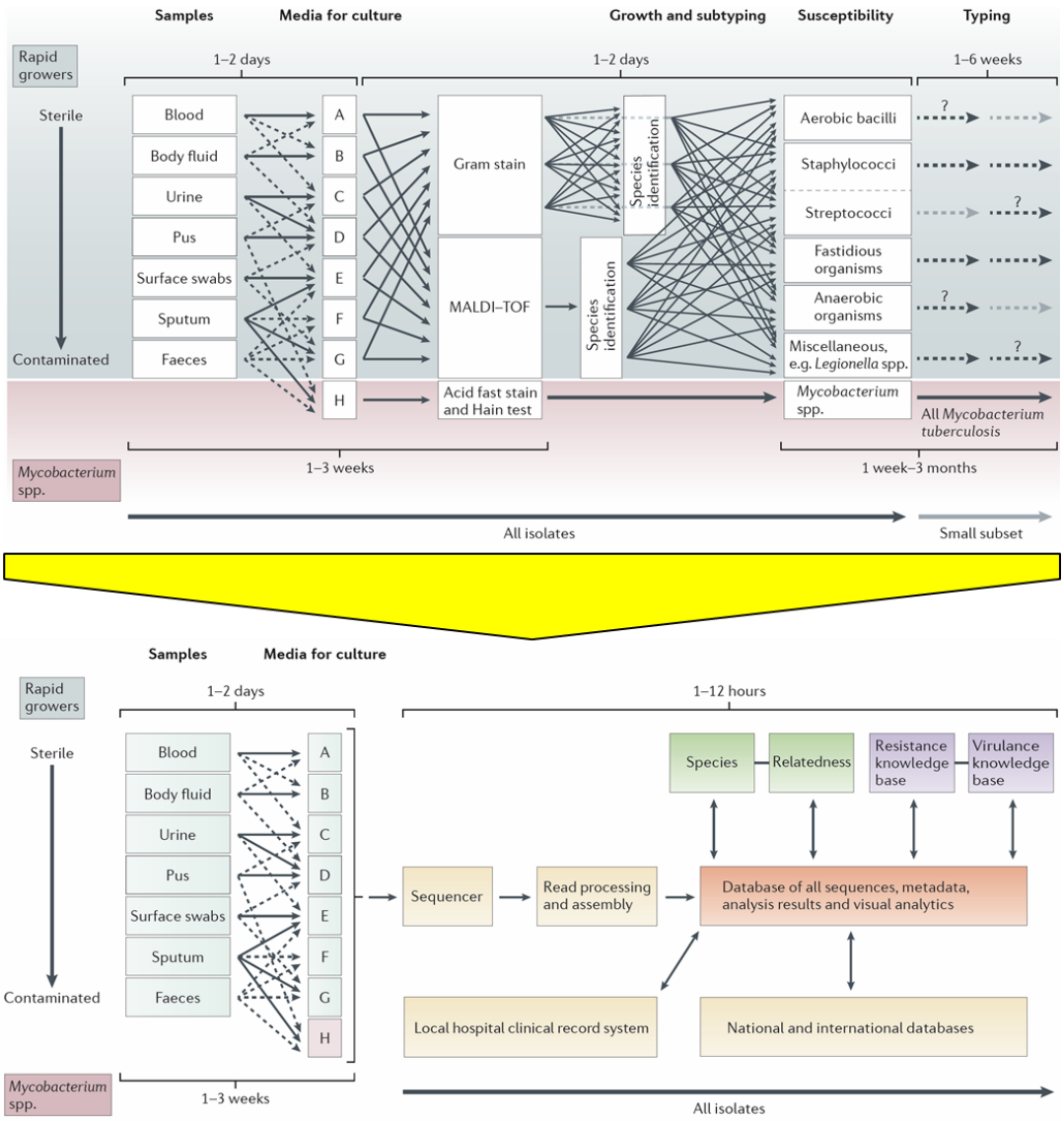

Figure 3

In terms of data collection, the following additional difficulties have been encountered:

- Combine large number of WGS and targeted sequencing data along with host and mycobacteriological data in one platform
- Streamline data flow (quality filters and data formats for easy merging of datasets)
- Provide fast, easy tools for consultation
- Embed prediction of drug resistance from genotype
- Allow the users to use and expand the dataset (in the easiest and quickest way possible)

Main outcomes from these three studies can be summarized as follows:

- Homoplasy: more than 150 homoplastic genetic variants were observed, but not all of them were associated with resistance. Independent mutations observed were often associated to high frequency.
- Variation: the rarefaction curve of variants (see *Figure 4*) shows how increasing the number of strains, the number of different variants observed also increases.
- Predicting resistance: analysis of >3000 strains allowed the identification of different determinants.
- Discrepancies: a lot of discrepancies between genotype/phenotype (especially for pyrazinamide (PZA) and ethambutol (EMB) drugs); many mutations found in both drug-resistant (DR) and drug-susceptible (DS) isolates. As an outcome, it is difficult to understand the role (if any) of these genetic variants in determining a drug resistant phenotype.

- Frequency of determinants: most of resistant determining variants only occurred once; only 25% occurred >2 times (see *Figure 4*).
- Others: mixed infections, additive effects, epistasis, phenotypic errors, and clerical errors.
- Some resistance only occurs between 0.1-1% of isolates.

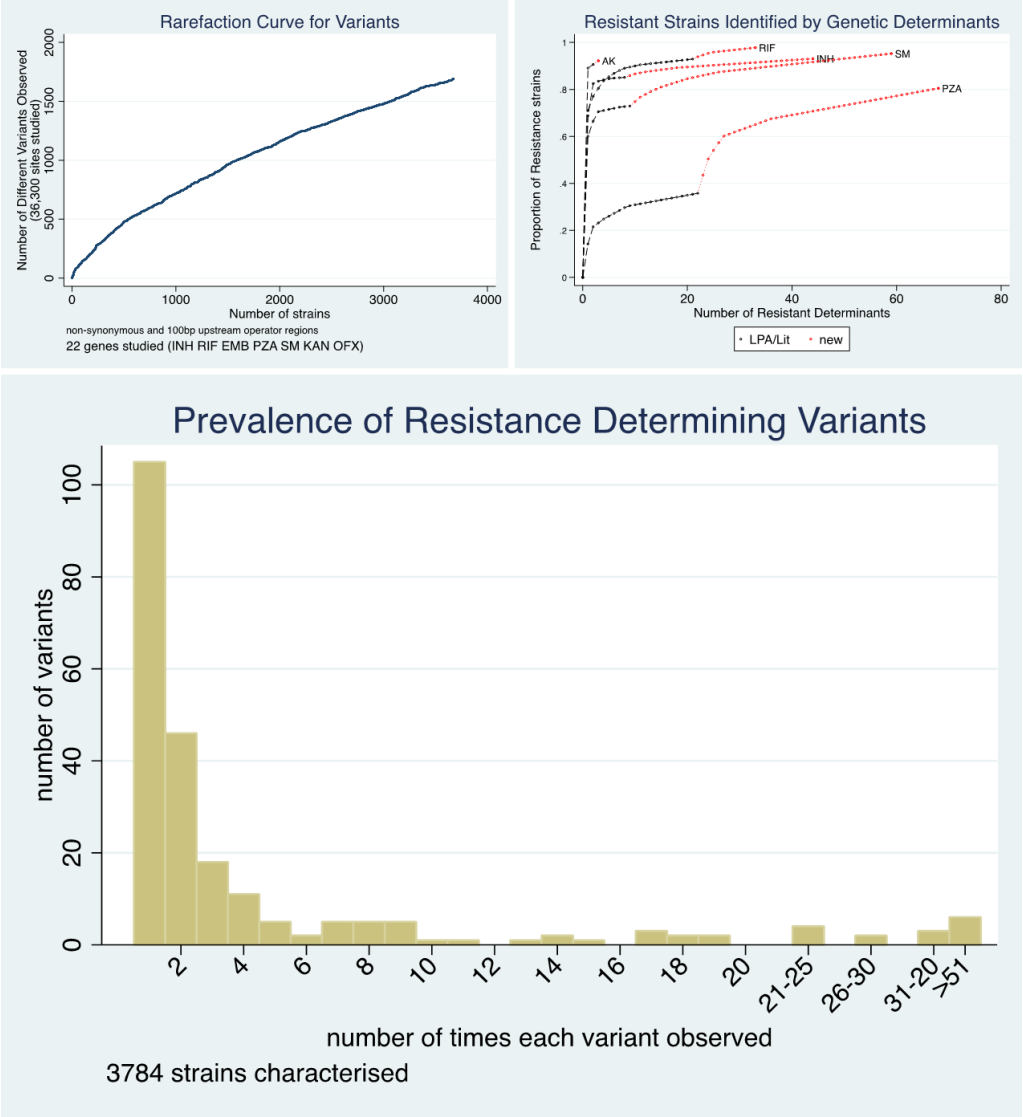

Figure 4

From this perspective, 10.000 strains are required to have at least 10 examples of drug resistance; 100.000 strains are required to have reasonable power to see "common" determinants of resistance to less common drugs in different phylogenetic lineages.

These data suggest that genetic validation is needed for some (at least those rare) mutations because DR-SNP association by statistics alone is not always sufficient.

The following mitigation strategies have been proposed:

- Join forces between researchers should help in providing sufficient data
- Fund the international online MTB encyclopedia initiative with relevant partners
- Develop easy interpretation tools for NGS data
- Build international consortia that work on genotype-phenotype correlation
- Contribute to the strain collection and characterization
- Contribute to the analysis

The group from the Harvard Medical School (Farhat, Murray) specifically worked on data collection and sharing options, drafting the path for developing a global, shared database. Two processes have been used:

the *Dataverse Network* and the *TwoRavens*. *Dataverse Network* can be described as a repository for research data that takes care of **long term preservation and good archival practices,** while researchers can **share, keep control** of and get **recognition** for their data. *Dataverse Network* also supports the sharing of research data with a persistent **data citation**, and enables **reproducible research**. *TwoRavens* allows data modeling using different variables to mine datasets. The two tools form the basis for the development of a new online platform under construction (*Translation Genomics of Tuberculosis – genTB, Figure 5*) to share data, run predictions, map results and explore shared data (a platform tool for both specialized and non-specialized users).

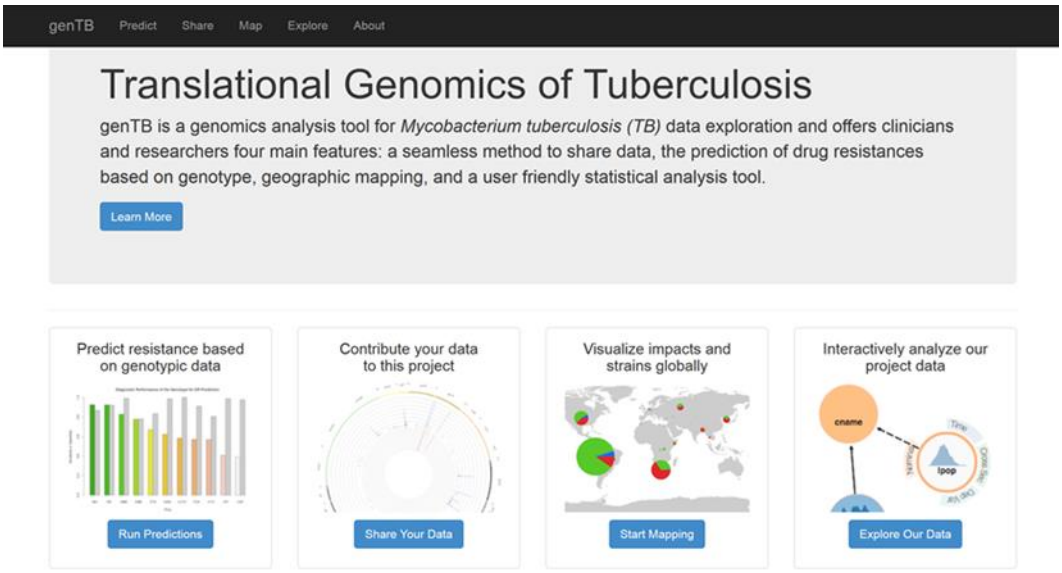

Figure 5

Session 3
Strategy for advocacy for data sharing

Clinicians claim for improved diagnostic tools for DR-TB: to achieve best management of patients they need rapid and accurate assays enabling the choice of the most effective treatment regimen, and this should also be available for drugs used to cure multidrug-resistant / extensively drug-resistant cases.

WHO is promoting the use of WGS data for public health purposes:

- Drug resistance surveillance in parallel with DST: 5,000-6,000 isolates (supported also by BMGF)
- Expert meeting planned at the end 2015 to evaluate the role of sequencing for surveillance and diagnosis of DR-TB
- Gene/WGS sequencing planned to replace DST at least in surveillance of DR-TB.

For these reasons, we need a global sequencing database and the WHO fully supports the initiative. The priority for **WHO is to ensure public health benefits from the database, which should have large geographical representation and should be accessible to all (with some restricted information). In addition, benefits should be clearly identified for countries/ministries of health contributing data to the data base (i.e. in terms of clinical management support).**

The role of WHO in the development of a global database would be:

- To liaise with countries contributing sequencing data (contact point with ministries of health)
- To avoid loss of data (especially from some settings)
- To ensure transparency of data management in the interest of countries and patients and in line with local legal requirements
- To ensure that evidence generated be rapidly translated into policy.

Countries could provide relevant metadata (social, clinical, radiology, history, risk factors…) along with clinical isolates. However, any contribution needs to be regulated and countries providing data should have benefits from this activity. For example, Belarus applies the following requirements for data sharing:

- Signed agreement between the National Tuberculosis Control Programme (NTP) and those receiving data
- NTP involved in publications
- NTP should receive support for interpretation of sequencing data and their clinical use
- NTP should have early access to new diagnostics which would eventually be derived.

Plenary discussion

Several points emerged from the open discussion:

- **This workshop is to join efforts and strengths; this is not to "steal" projects or data.**

- **Most of the efforts in WGS are intended to characterize drug-resistant isolates. We also need data from drug susceptible isolates.**

- **We need mutations with an evidence for their use.**

- **Sharing should allow higher productivity for researchers, policy makers and industries.**

- **Validation process for objective evidence for a mutation to be associated with resistance:**

  - Risks/hazards: genetic validation of mutations would be recommended. However, for understanding the contribution and the effect of multiple additive mutations the application of such approach could be a difficult task.

    Mitigations: the inclusion of data from thousands of clinical isolates and the use of appropriate statistical tools should also allow to provide reliable data. In several cases, this would avoid the need of genetic validation.

  - Risks/hazards: each WGS leads to mutations found only in one isolate. So homoplasy does not always represent the answer that helps with big numbers.

    Mitigations: genetic validation for some rare variants can be performed. These would not be expected to represent a high percentage of the "clinically relevant data".

  - Risks/hazards: genetic variants would be linked to phenotypic DST; however, a correlation between mutations and clinical outcome would also be appropriate. Collection and analysis of these data would represent a challenge: diverse informed consent rules across the countries would tangle the inclusion of data in the database and the use of different multi-drug regimens (including non-standard regimens) would reduce the statistical power of any association between genetic variants – single-drug resistance – clinical outcome. Finally, the inclusion of clinical data in the database requires the development of additional (i) privacy policies and guidelines and (ii) criteria for inclusion. Another important issue concerns the gold-standard: DST is often used as reference gold-standard, however its reliability depends upon the technique used and the quality of results. It is very difficult to assess data quality for phenotypic DST (external quality control, etc.).

    Mitigations: Standardization and specific guidelines could help mitigating drawbacks derived by the absence of an absolute phenotypic gold-standard (both in terms of phenotypic results and techniques used). The inclusion of data from thousands of clinical isolates and the use of appropriate statistical tools should also allow to provide a lot of reliable data: mistakes from phenotypic DST could be revised thanks to the replicates available in a big database.

- **Data sharing:**

  - Risks/hazards: lack of standardization is a real problem. Data-sharing poses also quality-related challenges.

    Mitigations: previous experience on HIV and other fields could help in solving problems. *Dataverse Network* represents a real tool for data collection; for this reason, it could be used as a central database. All data in *Dataverse Network* will be public and open access. Data upload can be done within the restricted area (accessible by credentials); this should allow to include quality control checkpoints for sequencing and DST data. In addition, the need of data sharing is common between researchers, clinicians and policy makers. This should drive efforts to find the best way to achieve the

establishment of a global database. The development of guidelines should help in regulating data sharing.

- Risks/hazards: how to share data before publication?

  Mitigations: the National Institutes of Health (NIH) seems to be working on developing regulations for data sharing.

- Risks/hazards: clinical/patient data.

  Mitigations: the development of guidelines and rules should help in regulating data sharing.

- Risks/hazards: Unwillingness of researchers/countries to share data. For some countries capacity should be built up.

  Mitigations: access to new tools/diagnostics could really drive countries to provide data (i.e. support in terms of clinical management, because introducing diagnostic prototypes could not always be feasible for incidence/prevalence and/or clustering reasons). The development of guidelines should help in regulating data sharing for countries. Every test going by BMGF/FIND is then available at controlled price; this should also act as an incentive for data sharing. Also in terms of new software/platforms developed, the benefit for those providing data would be the availability of new diagnostic algorithms.

- **Efforts:**

  - Risks/hazards: NGS initiatives are fragmented. How to bring together the efforts? How to bring together all the platforms developed/under development?

    Mitigations: BMGF is available to provide support for building this large shared database. CPTR, NDWG, FIND, WHO are committed to provide support and to facilitate partnerships and agreements between researchers, companies, clinicians and countries. Funding initiatives can also be consulted (e.g. Wellcome Trust). In addition, the need for data sharing is common between researchers, clinicians and policy makers. This should drive efforts to find the best way to achieve the establishment of a global database.

- **Data use:**

  - Risks/hazards: We do not know how to treat patients based only on sequencing results for most of the mutations: we have some (few) high confidence mutations; for many other mutations we are not confident with. How to manage clinical regulatory rules for the use of the database for predicting drug resistances? How to regulate the clinical use of these results in clinical practice?

    Mitigations: reproducibility on a large number of isolates can help in identifying high confidence markers of resistance; clinical trials could also be set up. Good quality data on (i.e.) 100.000 isolates (DST, WGS, etc.) would lead to discussion at the WHO and evidence-based analysis would promote endorsements/policies for the use of specific mutations.

Presentations

All the presentations from the workshop are available online on the NDWG website at:

http://www.stoptb.org/wg/new_diagnostics/datasharing.asp

Participants

Eric Adam
Otsuka SA
Switzerland

Enrique Avilés
CPTR - Critical Path Institute
United States

Jean-Luc Berland
Fondation Mérieux
France

Catharina Boehme
FIND
Switzerland

Sonia Borrell
Swiss TPH
Switzerland

Tobias Broger
FIND
Switzerland

Andrea Cabibbe
Fondazione Centro San Raffaele
Italy

Daniela Cirillo
Fondazione Centro San Raffaele
Italy

Davide Cittaro
Fondazione Centro San Raffaele
Italy

Inaki Comas
FISABIO
Spain

Derrick Crook
University of Oxford
United Kingdom

Ana Cruz
University of Oxford
United Kingdom

Siva Danaviah
Africa Centre Virology Laboratory, UKZN
South Africa

Dalene von Delft
TB Proof
South Africa

Anne Marie Demers
Stellenbosch University
South Africa

Claudia Denkinger
FIND
Switzerland

Keertan Dheda
University of Cape Town
South Africa

Gregory Dolganov
Stanford University
United States

David Dolinger
FIND
Switzerland

Katja Einer-Jensen
Qiagen
United Kingdom

Maha Farhat
Harvard School of Public Health
United States

Jim Gallarda
Bill & Melinda Gates Foundation
United States

Sebastien Gagneux
Swiss TPH
Switzerland

Debra Hanna
Critical Path Institute
United States

Rumina Hasan
Aga Khan University
Pakistan

Zahra Hasan
Aga Khan University
Pakistan

Sven Hoffner
SMI
Sweden

Nazir Ismail
NICD
South Africa

Vivian Jonas
Hologic
United States

Claudio Köser
University of Cambridge
United Kingdom

Davide Manissero
Qiagen
United Kingdom

Arne Materna
Qiagen
United Kingdom

Ruth McNerney
LSHTM
United Kingdom

Ruben van der Merwe
Stellenbosch University
South Africa

Paolo Miotto
Fondazione Centro San Raffaele
Italy

Igor Mokrousov
St. Petersburg Pasteur Institute
FEI

Stefan Niemann
Research Centre Borstel
Germany

Tim Peto
University of Oxford
United Kingdom

Alain Pluquet
bioMérieux
France

James Posey
CDC
United States

Leen Rigouts
ITM, Antwerp
Belgium

Timothy Rodwell
University of California, San Diego
United States

Marco Schito
NIAID / NIH
United States

Gary Schoolnik
Stanford TB molecular diagnostics group
United States

Alena Skrahina
National TB Programme
Belarus

Angela Starks
CDC
United States

Lynsey Stewart-Isherwood
University of Witwatersrand
South Africa

Philip Supply
Institut de Biologie de Lille
France

Alessandra Varga
FIND/NDWG
Switzerland

Maragretha de Vos
Stellenbosch University
South Africa

Timothy Walker
University of Oxford
United Kingdom

Marieke van der Werf
ECDC
Sweden

Viacheslav Zhuravlev
Research Institute of Phthisiopulmonology
Russia

Matteo Zignol
World Health Organization
Switzerland

Thank you

We wish to thank the Bill & Melinda Gates Foundation for their invaluable support.

BILL & MELINDA GATES foundation