SUPPLEMENT ARTICLE

Integration of Published Information Into a Resistance-Associated Mutation Database for *Mycobacterium tuberculosis*

Hugh Salamon,[1] Ken D. Yamaguchi,[1] Daniela M. Cirillo,[2] Paolo Miotto,[2] Marco Schito,[3] James Posey,[4] Angela M. Starks,[4] Stefan Niemann,[5] David Alland,[6] Debra Hanna,[7] Enrique Aviles,[7] Mark D. Perkins,[8] and David L. Dolinger[8]

[1]Knowledge Synthesis Inc., Berkeley, California; [2]IRCCS San Raffaele Scientific Institute, Milan, Italy; [3]HJF-DAIDS, a Division of The Henry M. Jackson Foundation for the Advancement of, Military Medicine, Inc., NIH, DHHS, Bethesda, Maryland; [4]Center for Disease Control and Prevention, Atlanta, Georgia; [5]Forschungszentrum Borstel, Germany; [6]Rutgers University, New Jersey; [7]Critical Path Institute, Tucson, Arizona; and [8]FIND, Geneva, Switzerland

Tuberculosis remains a major global public health challenge. Although incidence is decreasing, the proportion of drug-resistant cases is increasing. Technical and operational complexities prevent *Mycobacterium tuberculosis* drug susceptibility phenotyping in the vast majority of new and retreatment cases. The advent of molecular technologies provides an opportunity to obtain results rapidly as compared to phenotypic culture. However, correlations between genetic mutations and resistance to multiple drugs have not been systematically evaluated. Molecular testing of *M. tuberculosis* sampled from a typical patient continues to provide a partial picture of drug resistance. A database of phenotypic and genotypic testing results, especially where prospectively collected, could document statistically significant associations and may reveal new, predictive molecular patterns. We examine the feasibility of integrating existing molecular and phenotypic drug susceptibility data to identify associations observed across multiple studies and demonstrate potential for well-integrated *M. tuberculosis* mutation data to reveal actionable findings.

Keywords. tuberculosis; drug resistance; resistance-associated mutations; genomic sequencing; drug susceptibility testing; database.

Since 2002 there has been a gradual 1.3% annual decrease in the incidence of tuberculosis worldwide. Although this trend is encouraging, it is too weak to lead to elimination of tuberculosis as a public health problem by 2050, which is the goal of the World Health Organization (WHO) [http://www.who.int/tb/strategy/stop_tb_strategy/en/]. The challenge to stop tuberculosis is severely complicated by the increasing incidence of drug-resistant tuberculosis. Although rapid and accurate detection of tuberculosis will be a key factor in conquering tuberculosis, both treatment of the disease and better success preventing its transmission are significantly boosted by accurate information on drug susceptibility [1]. The WHO defines multidrug-resistant tuberculosis (MDR-TB) as resistance to isoniazid and rifampicin, with or without resistance to other first-line drugs [http://www.who.int/tb/challenges/mdr/tdrfaqs/en/]

With the introduction of new drug combinations and regimens, and patients with potentially more complex resistance profiles, it is imperative to be able to provide a comprehensive profile of drug susceptibility in order to select the correct therapies. Bacterial culture-based drug susceptibility testing (DST) is current "gold standard," but is technically difficult and time-consuming. DST methods are not standardized and results may vary depending on culture techniques employed [2–4], which is especially true for isolates with low-level resistance or when testing resistance to certain second-line drugs. Phenotypic DST methods can also expose laboratory workers to potential infection. Thus new approaches to determining drug resistance are needed.

Correspondence: Hugh Salamon, PhD, Knowledge Synthesis Inc., 725 Folger Ave, Berkeley, CA 94710 (hugh@knowledgesynthesis.com).

The Journal of Infectious Diseases® 2015;211(S2):S50–7

DOI: 10.1093/infdis/jiu816

Detection of resistance-conferring mutations with methods such as polymerase chain reaction and hybridization, targeted sequencing of specific genes, or whole genome sequencing are attractive and promising alternatives to phenotypic DST methods. The data from sequencing especially have been encouraging, with the identification of genes and intergenic noncoding regions associated with drug resistance. However, no systematic study has been performed to correlate genotypic output of either a targeted or whole genome sequencing approach with phenotypic drug response in culture and clinical outcome. In addition, there remains a need to inform logistical decisions on developing simple, rapid, affordable molecular tuberculosis drug-resistance diagnostics, particularly in light of challenges faced by clinical laboratories in low- to middle-resource settings. Currently available DNA sequencers all have reproducibility and performance issues, technical biases, and provide data that require informatics-intensive activities. While welcomed in a research setting, sequencing protocols provide data that need to be reduced to actionable knowledge and may contain false calls (errors) that require base-by-base review. Before sequencing technologies can impact tuberculosis on a global scale, we need to learn how to translate data into statements on drug resistance in a well-supported and reproducible fashion.

To affect guidance in the treatment of patients with tuberculosis, all current analysis and modelling point to the necessity for a solution based on sequence-level results generated as near to the patient as possible [1, 5]. This need for technology proximal to the point of care must be balanced with requirements for data quality. There are multiple sequencing instruments, multiple sample-processing approaches, multiple options for user interfaces, and as-yet incomplete global data associating specific mutations with degrees of resistance to different drugs. Both a protocol to measure mutations and algorithms to interpret the drug resistance implied by mutation data will need to be developed.

Although new sequencing methods are providing an avalanche of genomic information at continually lower cost [6–8], this wealth of information is currently underexploited for diagnostics. Although there are regional and research activities using raw sequencing results from various platforms, there is no generally accepted sequence data-handling approach that is vetted, quality controlled, and readily accessible to the scientific community, let alone usable globally in a clinically constructive way. Efforts to exploit sequencing technologies for tuberculosis treatment and control will require development of solutions tailored to translate this information into easy-to-use and intuitive diagnostic devices.

Data on mutations identified in drug-resistant isolates alone is not enough to determine the association of the mutation with resistance or to demonstrate causality. Therefore, we should focus on data incorporating both susceptible and resistant samples.

Currently, mutation and drug-resistance data for *M. tuberculosis* are scattered across multiple independent databases, journal articles, and their Supplementary Materials. We sought to identify data on isolates appropriate to integrate into a single database to enable querying data across study sources. The purpose of this work was to determine (i) the challenges to integrating mutation and DST data from multiple sources, and (ii) if data, once integrated, allow for systematic analysis that could inform diagnostics development.

METHODS

Published articles and online repositories were identified and reviewed for data on drug resistance-associated single nucleotide polymorphisms. The following criteria were used to prioritize the data sets for inclusion in this integration project:

(i) DST and mutation data on individual isolates preferred over summary statistics (desired, not required), (ii) number of isolates reported (desired more than 50 isolates in a publication), (iii) mutation data on susceptible isolates (required), (iv) easily understandable methods and results (required), (v) DST data on wild-type isolates (required), (vi) drug-level DST results rather than isolate classification solely by MDR or XDR criteria (required), (vii) publications post-2009 (required since data from pre-2010 publications were included via integration of the tuberculosis drug resistance mutation database (TBDReaMDB) [9]), (viii) understandable, well documented, and readily available data tables in articles, Supplementary Tables, or website portals (required), and (ix) publication not on hold or retracted (required).

Table 1 documents the article sources that were included [10–15]. Table 2 documents the online database resources.

Data Integration

The tuberculosis drug resistance database (TBDR) was established to integrate drug resistance mutation data from the different sources and to enable querying for those mutations that have supporting evidence from multiple studies. The intent was to capture as much information as possible from each study yet also allow analysis across diverse studies. During the development of the database specific issues were identified, and flexibility was built into the database. For instance, some studies reported mutations only at the amino acid level while others reported codon changes. Trivial conversions were implemented, such as translating codons to amino acids or nucleotides to the negative strand if appropriate. The database structure enabled on-the-fly summarization of data at the nucleotide level (eg, "S315T (AGC/ACC)"), the most granular level we store, or at the amino acid level (eg, "S315T"). Some source information, such as the genomic coordinates reported by the Broad/GTBDR database, was captured to permit future efforts to bring identical mutations together.

Table 1. Articles Used as Data Sources

PMID	Compact Citation	Comments
21 300 839	Campbell et al, 2011	314 clinical isolates with varied resistance patterns[a]
22 294 518	Casali et al, 2012	1000 sequenced isolates, multiple DST results
23 019 190	Nosova et al, 2013	68 strains were selected at random (38 strains resistant, 30 susceptible to OFX)
24 353 002	Rodwell et al, 2014	Tables 3, 4, 5, 6*A*, 6*B*, and 6*C* provide tabulated results for mutations and DST
24 478 476	Lin et al, 2014	Table 2 summarizes mutations and DST
25 336 456	Miotto et al, 2014	1066 isolates not in Casali et al, 2012, profiled for pncA mutations and PZA DST[b]

Abbreviations: DST, drug susceptibility testing; PMID, PubMed identifier.

[a] Isolate level data communicated by author Dr Posey.

[b] Communicated by authors Drs Cirillo and Miotto.

Some researchers reported mutation observations at the isolate level, while most authors reported just tabulated findings. Sometimes these tabulated findings are co-occurring mutations, which could give some insight into isolate-level observations. TBDR captured as much structure from each study as possible. For isolate-level reports TBDR stored the information on individual isolates then automatically summarized the tabulated results for each drug. For co-occurring mutations, the structure of the data was preserved to enable future analyses that rely on isolate-level information.

Researchers reported *rpoB* numbering using either *Escherichia coli* or *M. tuberculosis* numbering. TBDR uses *M. tuberculosis* numbering and reports the conversion from *E. coli* if and when it is performed. Researchers reported promoter mutations using a variety of genes in the same operon. TBDR normalizes these names only for the *fabG1-hemZ* operon promoter, mutations of which appear in TBDR as the *inhA* promoter. For gyrB, numbering systems have varied [16], and TBDR numbers 714 amino acids (NCBI protein accession number WP_003901763.1).

TBDR was built to provide reproducible results through automated processes. To minimize the manual manipulation of source material, computer programs were written to parse the 1523 data files in the case of the Broad/GTBDR database, an Excel file for TBDReaMDB, and various primary and Supplementary Tables in other publications. The main exception to the automation was PDF document table extraction, which typically required some manual cleanup after a copy and paste.

For published work TBDR stores PubMed identifiers (PMIDs). Since the TBDReaMDB source material did not provide PMIDs, these were identified manually. TBDR connects to PubMed to load reference details via the PMID.

RESULTS

Database Summary

The TBDR database currently contains 39 756 mutations across 29 genes and DST results for 23 drugs and one unspecified fluoroquinolone category. Because some studies performed DST for multiple drugs, the data from 80 studies, including the 73 found in the TBDReaMDB, comprised 148 investigations into drug resistance.

Across the 29 genes, mutations consisted of 1417 distinct amino acid substitutions, 89 distinct regulatory code changes, 105 insertions, and 106 deletions. Table 3 shows the mutations observed in the context of each drug. For example, there were 21 398 mutations observed in isolates subjected to DST for amikacin, and 5556 of these were observed in amikacin-resistant isolates. The numbers in Table 3 are purely descriptive, include mutations measured for genes not expected to be associated with resistance to the particular drug, and do not by themselves inform us about mutation-resistance associations. Table 4 summarizes the number of isolates subjected to DST for each of the drugs in the database. The database content described in Tables 3 and 4 serve as the basis for the calculations and queries described below.

Web Portal to Database

A web portal was established to enable simple access to the database. The portal both facilitated integration efforts and allowed sharing of results among coauthors. There are four main types of tables provided by the portal: (i) a list of drugs with data contained within the database, (ii) a summary of all

Table 2. Online Sources of Data

Description	Compact Citation	Web Site	Comments
TBDReaMDB compiled a comprehensive list of the genetic polymorphisms associated with first- and second-line drug resistance in clinical *M. tuberculosis* isolates throughout the world.	Sandgren et al, 2009 [9]	https://tbdreamdb.ki.se/Info/	High-confidence mutations were integrated into TBDR.org.
Broad's Gates Tuberculosis Drug Resistance Database (Broad/GTBDR). Downloadable isolate mutation and DST results.	None identified	http://www.broadinstitute.org/annotation/genome/mtb_drug_resistance.1/DirectedSequencingHome.html	Download includes 1398 isolates, which were integrated into TBDR.org.

Abbreviations: DST, drug susceptibility testing; TBDR, tuberculosis drug resistance database; TBDReaMDB, tuberculosis drug resistance mutation database.

Table 3. Numbers of Mutations Integrated into TBDR

Drug	Resistant	Susceptible	Total
Amikacin	5556	15 842	21 398
Amoxiclav	240	62	302
Capreomycin	11 407	9415	20 822
Ciprofloxacin	4372	12 170	16 542
Clarithromycin	1233	592	1825
Clofloxacin	17	2418	2435
Cycloserine	129	14 727	14 856
Ethambutol	19 544	8727	28 271
Ethionamide	11 012	6314	17 326
Fluoroquinolones	1091	129	1220
Gatifloxacin	556	518	1074
Isoniazid	28 647	2476	31 123
Kanamycin	5417	11 018	16 435
Levofloxacin	2182	7332	9514
Linezolid	21	1791	1812
Moxifloxacin	1433	3395	4828
Ofloxacin	2650	6525	9175
Para-aminosalicylic Acid	1979	15 247	17 226
Prothionamide	1558	3692	5250
Pyrazinamide	13 370	9094	22 464
Rifabutin	2686	925	3611
Rifampicin	27 150	3842	30 992
Streptomycin	21 250	7497	28 747
Thioacetazone	447	1698	2145

Shown are the numbers of observed mutations at any genetic locus investigated, summed across resistant and susceptible isolates for 23 drugs and the (unspecified) fluoroquinolones category.

Abbreviation: TBDR, tuberculosis drug resistance database.

Table 4. The Number of Isolates Found Resistant or Susceptible to 23 Drugs

Drug	Resistant	Susceptible	Total
Amikacin	593	1622	2215
Amoxiclav	15	4	19
Capreomycin	945	1253	2198
Ciprofloxacin	309	912	1221
Clarithromycin	68	34	102
Clofloxacin	1	137	138
Cycloserine	8	855	863
Ethambutol	2447	2993	5440
Ethionamide	613	372	985
Fluoroquinolones	1048	897	1945
Gatifloxacin	64	58	122
Isoniazid	5142	2299	7441
Kanamycin	655	973	1628
Levofloxacin	148	467	615
Linezolid	1	87	88
Moxifloxacin	170	401	571
Ofloxacin	307	741	1048
Para-aminosalicylic Acid	126	1061	1187
Prothionamide	194	377	571
Pyrazinamide	2350	2503	4853
Rifabutin	259	212	471
Rifampicin	4712	4825	9537
Streptomycin	2594	1486	4080
Thioacetazone	59	331	390

mutations for a given drug, (iii) a list of mutations provided by a reference source, and (iv) a summary for each drug of all canonical mutations for that drug and an array of resistance-mutation statistics. Sensitivity, specificity, positive predictive value, and negative predictive value were calculated on the pooled counts across studies for the following: the number of (i) clinical isolates with both phenotypic and genotypic results for the mutation, (ii) isolates resistant to the drug, (iii) isolates susceptible to the drug, (iv) isolates with the specific mutation, (v) mutant isolates resistant to the drug, and (vi) mutant isolates susceptible to the drug.

Two modes by which the database reports mutations across studies were deemed potentially useful to inform diagnostics research. The first mode merges all mutations with the same amino acid substitution or promoter position. The second mode merges mutations if the nucleotide-level information is also identical. Because some references did not report codon changes and others reported resistance mutations with the nucleotide change but without the codon, the nucleotide-level reports contain multiple rows for identical mutations. The interface allows selecting studies to exclude from the results report.

Results on Drug Resistance-associated SNPs

For each drug category in the TBDR database, mutations were queried to determine which were observed in at least 3 studies, exhibited a nominal specificity greater than 95% and were found at a higher rate in resistant isolates than in susceptible isolates. Many associations identified were noncanonical since resistance mutations for one drug carry information about resistance to another drug tested in the same study. This phenomenon is to be expected, as resistance to multiple drugs is, unfortunately, not uncommon. Table 5 lists the canonical gene-drug associations we used to limit our presentation on drug resistance-associated mutations. Eleven drugs yielded mutations in canonically associated genes that met the above criteria (Table 5). Supplementary Table 1 lists 106 amino acid substitutions and 11 regulatory resistance mutations that were defined by the query. The table is sorted sequentially on three columns: drug, specificity (highest first), and sensitivity (highest first). Table 6 lists the substitutions and regulatory mutations for 2 first-line drugs, isoniazid and rifampicin, sorted as in Supplementary Table 1. This relatively simple query represents a first attempt at using the integrated data. Further refinement by a panel of experts would surely improve the utility of the resistance mutations list for each drug.

There exist mutations in canonically drug-resistance-associated genes that do not reliably predict DST results. Data

Table 5. Drug-gene Associations

Drug	Canonically Associated Genes	Resistance-Associated Genes From TBDR Query (Number of Mutations Matching Criteria)
Aminoglycosides		
Amikacin	*rrs*	*rrs (2)*
Kanamycin	*rrs, eis*	*rrs (3)*
Capreomycin	*rrs, tlyA*	*rrs (3)*
Clarithromycin	*rrl*	*NA*
Ethambutol	*embB*	*embB (9)*
Ethionamide	*inhA, ethA*	*NA*
Fluoroquinolones[a]	*gyrA, gyrB*	*gyrA (7)*
Ciprofloxacin	*gyrA, gyrB*	*NA*
Clofloxacin	*gyrA, gyrB*	*NA*
Gatifloxacin	*gyrA, gyrB*	*NA*
Levofloxacin	*gyrA, gyrB*	*NA*
Moxifloxacin	*gyrA, gyrB*	*gyrA (5)*
Ofloxacin	*gyrA, gyrB*	*gyrA (7)*
Isoniazid	*katG, inhA, ahpC, kasA*	*katG (5), inhA (2)*
Linezolid	*rrl, rplC*	*NA*
Para-aminosalicylic Acid	*thyA*	*NA*
Prothionamide	*ethA, inhA*	*NA*
Pyrazinamide	*pncA*	*pncA (66)*
Rifabutin	*rpoB*	*NA*
Rifampicin	*rpoB*	*rpoB (16)*
Streptomycin	*rpsL, gid, rrs*	*rpsL (3), rrs (8)*

Abbreviation: TBDR, tuberculosis drug resistance database.

[a] For studies that reported results for unspecified fluoroquinilones.

integration can help us identify departures from wild type that do not confer drug resistance. This is a specific strength of TBDR, as it brings together the results of multiple studies in a manner that can be queried to address such concerns. For example, TBDR contains *pncA* mutations observed only in pyrazinamide-susceptible isolates and in more than one study, including C14G, S59F, F81S, H82Y, Y103C, and A143T. Similarly, 4 studies found *rrs* 1402 (C->N) mutations in a total of 5 isolates tested for amikacin susceptibility, and all were susceptible, indicating that this mutation may indeed be a poor marker of resistance to this particular drug [17].

Caveats

Table 6 and Supplementary Table 1 identify mutations that are supported by multiple studies and therefore have a higher confidence in their association with drug resistance. These results are not a comprehensive investigation into each drug and mutation, or prediction of resistance to the drug.

The sensitivities for mutations for a given drug in Table 6 and Supplementary Table 1 are not cumulative for predicting drug resistance. First of all, isolates may have multiple mutations and thus contribute to multiple rows in the table. Second, we do not have all data at the isolate level, as summary tables generally do not preserve this important information. If we had all the data at the isolate level (ie, all mutations reported for each isolate), we could indeed ask what sensitivity (and specificity) combinations of mutations would provide for drug resistance across this idiosyncratic collection of samples.

The predictive value of mutations for drug resistance and diagnostics depends strongly on the proportion of mutation-typed samples that are *a priori* phenotypically resistant. Because the data from many studies are highly biased toward analysis of resistant isolates, the statistics reported are likely quite unreliable as predictions in any particular population. Researchers often avoided typing phenotypically wild-type isolates at their true rates in the population.

DISCUSSION

Integrating tuberculosis drug-resistance data into TBDR required addressing a number of data-handling issues. A straightforward query enabled by the database revealed 96 mutations informative of drug resistance and observed in multiple independent studies.

For sources where complete DST and mutation results were available, TBDR cataloged the association of all resistance mutations with drug sensitivity, including noncanonical associations. Significant noncanonical associations likely arise in part from the evolution of drug resistance to multiple drugs and in part because of ascertainment bias since many studies target populations with MDR and XDR isolates.

The quantitative science required for diagnostics development cannot be addressed fully in an analysis of data such as prepared here. Nevertheless, analyses of existing data should help prioritize mutations. A large impact on diagnostics and patient treatment could be made by using properly integrated data to help the community reach a data-driven consensus regarding tuberculosis drug-resistance predictive mutations.

The Grading of Recommendations Assessment, Development and Evaluation (http://www.gradeworkinggroup.org/) criteria need to be kept in mind when proposing mutations for diagnosis of drug-resistant tuberculosis. A key consideration is how well a mutation actually predicts drug resistance. Because there are associations of particular mutations with phylogeny and of phylogeny with geographic region, these data are important to capture for future drug resistance mutation studies. By including complete or at least expanded sequence results, it should be possible to better understand which mutations are likely to be causal and which are simply markers of resistance in specific populations. It may be important to determine when phylogenetic information has no appreciable impact on mutation-based prediction of drug resistance. The quality of evidence for predicting resistance to drugs will necessarily be better for some mutations than others.

Table 6. Drug Resistance-Associated Mutations for Isoniazid and Rifampicin[a]

Drug	Gene	Mutation	Total Isolates Typed[b]	Resistant Isolates Typed	Susceptible Isolates Typed	Isolates With Mutation	Resistant Isolates With Mutation	Susceptible Isolates With Mutation	Sensitivity	Specificity	PPV	NPV	Number of Studies
Isoniazid	*katG*	S315R	1719	1458	261	46	46	0	3.2	100.0	100.0	15.6	3
Isoniazid	*katG*	del	2424	1172	1252	21	21	0	1.8	100.0	100.0	52.1	6
Isoniazid	*katG*	S315I	4099	2689	1410	18	18	0	0.7	100.0	100.0	34.6	8
Isoniazid	*inhA*	−8 (T/N)	5808	3884	1924	101	100	1	2.6	99.9	99.0	33.7	11
Isoniazid	*katG*	S315N	6227	4239	1988	77	75	2	1.8	99.9	97.4	32.3	15
Isoniazid	*inhA*	−15 (C/N)	6984	4754	2230	895	875	20	18.4	99.1	97.8	36.3	17
Isoniazid	*katG*	S315T	7441	5142	2299	3623	3586	37	69.7	98.4	99.0	59.2	19
Rifampicin	*rpoB*	S450W	8323	3927	4396	68	68	0	1.7	100.0	100.0	53.3	14
Rifampicin	*rpoB*	Q432K	1887	1554	333	11	11	0	0.7	100.0	100.0	17.8	3
Rifampicin	*rpoB*	Q432L	2713	1974	739	11	11	0	0.6	100.0	100.0	27.4	7
Rifampicin	*rpoB*	H445D	9537	4712	4825	155	154	1	3.3	100.0	99.4	51.4	20
Rifampicin	*rpoB*	S441L	6718	2963	3755	24	23	1	0.8	100.0	95.8	56.1	11
Rifampicin	*rpoB*	H445Y	9537	4712	4825	325	323	2	6.9	100.0	99.4	52.4	20
Rifampicin	*rpoB*	H445R	9537	4712	4825	134	132	2	2.8	100.0	98.5	51.3	20
Rifampicin	*rpoB*	D435G	2972	2113	859	33	32	1	1.5	99.9	97.0	29.2	4
Rifampicin	*rpoB*	D435V	9537	4712	4825	293	287	6	6.1	99.9	98.0	52.1	20
Rifampicin	*rpoB*	Q432P	2393	1671	722	8	7	1	0.4	99.9	87.5	30.2	3
Rifampicin	*rpoB*	H445L	9095	4304	4791	96	88	8	2.0	99.8	91.7	53.2	18
Rifampicin	*rpoB*	L452P	9126	4429	4697	141	127	14	2.9	99.7	90.1	52.1	17
Rifampicin	*rpoB*	H445N	7956	3512	4444	40	27	13	0.8	99.7	67.5	56.0	13
Rifampicin	*rpoB*	S450L	9537	4712	4825	2939	2923	16	62.0	99.7	99.5	72.9	20
Rifampicin	*rpoB*	L430P	8375	3881	4494	74	54	20	1.4	99.6	73.0	53.9	14
Rifampicin	*rpoB*	D435Y	5400	3879	1521	73	62	11	1.6	99.3	84.9	28.3	16

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.

[a] Detailed results for a total of 11 drugs are included in Supplementary Table 1.

[b] For each row, the quantity in this column is the total number of isolates typed for the mutation irrespective of other mutations.

In short, an effort to include as many isolate-level records (as opposed to summary table information), and to gather enough information to address phylogeny, should allow for analyses that will boost confidence in drug resistance prediction.

Factors other than phylogeny are confusing to evaluate when associating mutations with drug resistance. Importantly, for some mutations, even causal mutations, imperfect correlation with phenotypic DST likely results from varying phenotypic testing methods or drug concentration thresholds across studies. TBDR records information on DST methods reported by different investigators and further analysis could support expert interpretation. Additionally, although some published studies are clearly annotated by geographic region, others are less clearly annotated and also may include isolates from multiple regions.

The integrated data provide a platform for addressing the multivariate properties of the resistance mutations, although we have not demonstrated such an analysis in this article. A multivariate analysis, using statistical, machine-learning, or other mathematical methods, could determine which of the resistance mutations are most informative, and specifically which do not offer additional information when other resistance mutations have been measured. The results of such a study could inform the selection of a panel of resistance mutations that most efficiently uses resources by reducing redundancy.

The current database and web interface are proof-of-principle tools that enabled the generation of the main results as presented herein. The tools demonstrate that data integration is an important component in development of analysis algorithms to identify drug resistance-predictive mutations.

Much of the data we encountered was presented only in summary tables. For data to be most useful to other investigators and inform diagnostics development, information on isolates should always be reported as complete reports on the isolates, including all DST and mutation typing. While most useful would be entry of the data into an appropriate database, at the very least these complete reports should be released as Supplementary Data. For analysis to best demonstrate the diagnostic potential of molecular patterns, it would be most useful to have data from studies that record treatment regimens and outcome data together with mutation and drug susceptibility phenotypes. The completeness of data gathered and other aspects of data quality control should be carefully targeted in future efforts to collect and analyze tuberculosis drug resistance data.

This demonstration of data integration for *M. tuberculosis* drug resistance-associated mutations provides two important lessons. First, the knowledge in the community is currently larger than perhaps has been understood by many researchers and diagnostics developers, and could better inform diagnostic development decisions in the near future. Second, we can anticipate many important issues for gathering and analyzing data with more modern tools, such as bacterial whole genome sequencing, which could better inform microbial profiling efforts in the near future.

A database such as TBDR could be expanded or incorporated into another database to address the growing needs for knowledge sharing with respect to sequence data and markers for tuberculosis drug resistance. To develop a relevant data repository, the database will need to clearly address objectives from the community, namely development of tests for detection of drug resistance and clinical impact. At the same time, to develop a sustainable data repository, appropriate partnerships among researchers, clinical trial groups, reference labs, and commercial entities will need to drive the technology development, as different parties have distinct needs. For example, clinical trial groups are required to anonymize data, and commercial parties may need to compare in-house results with database contents in a confidential manner. As an example, the Critical Path to TB Drug Regimens (CPTR) initiative (http://cptrinitiative.org) has developed strong partnerships with clinical trial groups and commercial entities to tackle the challenges facing tuberculosis drug development. As part of its ongoing work, CPTR has established data management practices and technology that can be adapted to assist needs for knowledge sharing with respect to sequence data and markers for tuberculosis resistance. The power of TBDR and subsequent databases that incorporate existing and prospectively gathered genotypes, phenotypes, and metadata lies in the ability to compose and execute queries. These queries will need to be designed by a collaborative effort among data scientists, stakeholders in diagnostics development, expert committees on tuberculosis drug resistance, and computational biologists. In this way the complexities of different drugs and mutation interactions can be addressed, and a consensus for predicting resistance to each drug should be reached.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online (http://jid.oxfordjournals.org). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Disclaimer. The views and opinions expressed in this article are those of the authors and do not necessarily represent an official position of the US Centers for Disease Control and Prevention.

Financial support. This work was supported by the Bill and Melinda Gates Foundation. M. S. is funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract HHSN272200800014C. The funders had no role in the analysis of data and decision to publish.

Potential conflicts of interest. D. L. D. and M. D. P. are employed by FIND, a nonprofit organization that collaborates with industry partners,

including Cepheid and Hain diagnostics among others, for the development, evaluation and demonstration of new diagnostic tests for poverty-related diseases. H. S. and K. D. Y. are the beneficial owners of Knowledge Synthesis Inc. D. A. reports grants from Cepheid, and royalties from a molecular beacons patent pool. D. H. and E. A. report funding from the Bill and Melinda Gates Foundation outside of this work. There are no patents or products with respect to this article. All other authors report no potential conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Jain A, Mondal R. Extensively drug-resistant tuberculosis: current challenges and threats. FEMS Immunol Med Microbiol **2008**; 53: 145–50.
2. Van Deun A, Wright A, Zignol M, Weyer K, Rieder HL. Drug susceptibility testing proficiency in the network of supranational tuberculosis reference laboratories. Int J Tuberc Lung Dis **2011**; 15:116–24.
3. Angra PK, Taylor TH, Iademarco MF, Metchock B, Astles JR, Ridderhof JC. Performance of tuberculosis drug susceptibility testing in U.S. laboratories from 1994 to 2008. J Clin Microbiol **2012**; 50: 1233–9.
4. Jiang G-L, Chen X, Song Y, Zhao Y, Huang H, Kam KM. First proficiency testing of second-line anti-tuberculosis drug susceptibility testing in 12 provinces of China. Int J Tuberc Lung Dis **2013**; 17:1491–4.
5. Boehme CC, Nabeta P, Hillemann D, et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med **2010**; 363: 1005–15.
6. Catanho M, Mascarenhas D, Degrave W, de Miranda AB. GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. Genet Mol Res GMR **2006**; 5:115–26.
7. Stucki D, Gagneux S. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. Tuberc Edinb Scotl **2013**; 93:30–9.
8. Lu JT, Campeau PM, Lee BH. Genotype-phenotype correlation—promiscuity in the era of next-generation sequencing. N Engl J Med **2014**; 371:593–6.
9. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med **2009**; 6:e2.
10. Campbell PJ, Morlock GP, Sikes RD, et al. Molecular detection of mutations associated with first- and second-line drug resistance compared with conventional drug susceptibility testing of *Mycobacterium tuberculosis*. Antimicrob Agents Chemother **2011**; 55:2032–41.
11. Casali N, Nikolayevskyy V, Balabanova Y, et al. Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Res **2012**; 22:735–45.
12. Nosova EY, Bukatina AA, Isaeva YD, Makarova MV, Galkina KY, Moroz AM. Analysis of mutations in the gyrA and gyrB genes and their association with the resistance of *Mycobacterium tuberculosis* to levofloxacin, moxifloxacin and gatifloxacin. J Med Microbiol **2013**; 62 (Pt 1):108–13.
13. Rodwell TC, Valafar F, Douglas J, et al. Predicting extensively drug-resistant *Mycobacterium tuberculosis* phenotypes with genetic mutations. J Clin Microbiol **2014**; 52:781–9.
14. Lin S-YG, Rodwell TC, Victor TC, et al. Pyrosequencing for rapid detection of extensively drug-resistant *Mycobacterium tuberculosis* in clinical isolates and clinical specimens. J Clin Microbiol **2014**; 52:475–82.
15. Miotto P, Cabibbe AM, Feuerriegel S, et al. *Mycobacterium tuberculosis* Pyrazinamide Resistance Determinants: a Multicenter Study. MBio **2014**; 5:e01819–14.
16. Maruri F, Sterling TR, Kaiga AW, et al. A systematic review of gyrase mutations associated with fluoroquinolone-resistant *Mycobacterium tuberculosis* and a proposed gyrase numbering system. J Antimicrob Chemother **2012**; 67:819–31.
17. Georghiou SB, Magana M, Garfein RS, Catanzaro DG, Catanzaro A, Rodwell TC. Evaluation of genetic mutations associated with *Mycobacterium tuberculosis* resistance to amikacin, kanamycin and capreomycin: a systematic review. PloS One **2012**; 7:e33275.