ORIGINAL ARTICLE

Genetic Determinants of Drug Resistance in *Mycobacterium tuberculosis* and Their Diagnostic Value

Maha R. Farhat[1,2], Razvan Sultana[3], Oleg Iartchouk[4], Sam Bozeman[5], James Galagan[6,7,8], Peter Sisk[9], Christian Stolte[10], Hanna Nebenzahl-Guimaraes[11,12,13,14,15], Karen Jacobson[16,17], Alexander Sloutsky[2,18], Devinder Kaur[18], James Posey[19], Barry N. Kreiswirth[20], Natalia Kurepina[20], Leen Rigouts[21,22], Elizabeth M. Streicher[17], Tommie C. Victor[17], Robin M. Warren[17], Dick van Soolingen[12,13,14], and Megan Murray[2,23]

[1]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, Massachusetts; [2]Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts; [3]Genomics England, Queen Mary University, London, United Kingdom; [4]Novartis Institutes for Biomedical Research, Cambridge, Massachusetts; [5]Abt Associates, Boston, Massachusetts; [6]Department of Biomedical Engineering, [7]Department of Microbiology, and [8]Bioinformatics Program, Boston University, Boston, Massachusetts; [9]Gen9, Inc., Cambridge, Massachusetts; [10]CSIRO, North Ryde, New South Wales, Australia; [11]National Institute for Public Health and the Environment, Bilthoven, the Netherlands; [12]Department of Pulmonary Diseases and [13]Department of Medical Microbiology, Radboud University Nijmegen Medical Center, Nijmegen, the Netherlands; [14]Life and Health Sciences Research Institute, School of Health Sciences, University of Minho, Braga, Portugal; [15]Life and Health Sciences Research Institute/3Bs, PT Government Associate Laboratory, Braga/Guimaraes, Portugal; [16]Section of Infectious Diseases, Boston University School of Medicine, Boston, Massachusetts; [17]DST/NRF Center of Excellence for Biomedical TB Research/SAMRC Center for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa; [18]University of Massachusetts Medical School, Worcester, Massachusetts; [19]Division of Tuberculosis Elimination, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia; [20]Public Health Research Institute Tuberculosis Center, Rutgers University, Newark, New Jersey; [21]Mycobacteriology, Institute of Tropical Medicine, Antwerp, Belgium; [22]Biomedical Sciences, Antwerp University, Antwerp, Belgium; and [23]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts

ORCID ID: 0000-0002-3871-5760 (M.R.F.).

Abstract

Rationale: The development of molecular diagnostics that detect both the presence of *Mycobacterium tuberculosis* in clinical samples and drug resistance–conferring mutations promises to revolutionize patient care and interrupt transmission by ensuring early diagnosis. However, these tools require the identification of genetic determinants of resistance to the full range of antituberculosis drugs.

Objectives: To determine the optimal molecular approach needed, we sought to create a comprehensive catalog of resistance mutations and assess their sensitivity and specificity in diagnosing drug resistance.

Methods: We developed and validated molecular inversion probes for DNA capture and deep sequencing of 28 drug-resistance loci in *M. tuberculosis*. We used the probes for targeted sequencing of a geographically diverse set of 1,397 clinical *M. tuberculosis* isolates with known drug resistance phenotypes. We identified a minimal set of mutations to predict resistance to first- and second-line antituberculosis drugs and validated our predictions in an independent dataset. We constructed and piloted a web-based database that provides public access to the sequence data and prediction tool.

Measurements and Main Results: The predicted resistance to rifampicin and isoniazid exceeded 90% sensitivity and specificity but was lower for other drugs. The number of mutations needed to diagnose resistance is large, and for the 13 drugs studied it was 238 across 18 genetic loci.

Conclusions: These data suggest that a comprehensive *M. tuberculosis* drug resistance diagnostic will need to allow for a high dimension of mutation detection. They also support the hypothesis that currently unknown genetic determinants, potentially discoverable by whole-genome sequencing, encode resistance to second-line tuberculosis drugs.

Keywords: multidrug-resistant tuberculosis; molecular diagnostics; sensitivity and specificity

(Received in original form October 28, 2015; accepted in final form February 22, 2016)

Correspondence and requests for reprints should be addressed to Maha R. Farhat, M.D., 55 Fruit Street, Building 148, Boston, MA 02114. E-mail: mrfarhat@partners.org

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org

Am J Respir Crit Care Med Vol 194, Iss 5, pp 621–630, Sep 1, 2016

Originally Published in Press as DOI: 10.1164/rccm.201510-2091OC on February 24, 2016
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Drug resistance threatens to undermine tuberculosis control. To tackle the drug-resistant threat, better diagnostic tests are needed that can accurately determine the sensitivity of the bacterium to the full panel of drugs used for tuberculosis treatment. Present tests only detect resistance to a small portion of these drugs, and for several the test accuracy is moderate or poor.

What This Study Adds to the Field: Our study investigated bacterial mutations that can be used to diagnose drug resistance to 13 antituberculosis drugs. The findings significantly expand the list of mutations that can be used for resistance diagnostics and imply that only diagnostics technologies that can detect hundreds of mutations are likely to achieve the goal of a comprehensive diagnostic test for tuberculosis drug resistance.

Global surveillance for drug-resistant (DR) tuberculosis (TB) suggests that at least 3.5% of the 9 million incident TB cases are multidrug resistant (MDR) (i.e., resistant to isoniazid [INH] and rifampicin [RIF]), and that 9% of these MDR cases are also extensively DR (XDR) (i.e., also resistant to amikacin [AMI], kanamycin [KAN], or capreomycin [CAP] and at least one fluoroquinolone [FLQ]) (1). The World Health Organization (WHO) estimates that MDR-TB is detected in fewer than 45% of the 480,000 people affected and of these, at most 70% receive appropriate drug therapy (1). The remainder are not only likely to fail treatment but also to spread resistant organisms (2). WHO cites MDR-TB as a public health crisis and a priority area that needs to be addressed for TB control (1).

One of the main challenges faced in the control of DR-TB is the lack of laboratory capacity for the diagnosis of resistance (3). Several problems limit the utility of conventional drug susceptibility tests (DSTs). First, culture-based methods are expensive and require a specialized biosafety environment that is usually present only in centralized reference laboratories. Second, the slow growth of *Mycobacterium tuberculosis* (MTB) implies that results may take weeks to months to be reported. Finally, methods for DST for several of the second-line drugs have not yet been sufficiently standardized (4, 5).

Molecular diagnostics are now available that offer multiple advantages for the diagnosis of DR-TB (6–8). Some can be performed directly on sputum and therefore do not require the biosafety facilities needed for conventional culture and can be performed by relatively unskilled workers. In some cases, results can be available within 3 hours (2). However, recommended assays only test for resistance to RIF (6, 8) and INH (8) and consequently, the WHO recommends that conventional culture and DST should continue to be used to "confirm or exclude XDR-TB" and individualize MDR-TB treatment regimens (5). Although expanded diagnostics that test resistance to FLQs and second-line injectables are now commercially available their sensitivity is only moderate ranging from 69.1 to 99.2% in different reports, and their use has not been endorsed by the WHO (9–11). The limited performance of these tests, which rely on detecting mutations within the narrow resistance-determining regions of *gyrA*, *gyrB*, *rrs*, and the *eis* promoter, has raised questions about the optimal molecular technology needed, including the level of multiplexing of genes and mutations that is needed for a comprehensive and accurate diagnostic. Here in the largest collection of prospectively collected DR isolates to date (12, 13), we identify molecular determinants of resistance to 13 anti-TB drugs using molecular inversion probes (MIPs), and present a validated prediction model based on the detection of mutations within the full length of 28 putative DR loci. Some of the results of these studies have been previously reported in the form of an abstract (14).

Methods

Archive Assembly

We identified 1,748 MTB isolates archived at six reference laboratories: the U.S. CDC, the New Jersey Public Health Research Institute (PHRI), the Massachusetts Supranational TB Laboratory (MSLI), Stellenbosch University (SU) in South Africa, the National Institute for Public Health and the Environment of the Netherlands (RIVM), and the Institute of Tropical Medicine housing the WHO Tropical Disease Research (TDR) strain bank (15). These laboratories were selected because they belonged to the WHO network of supranational reference laboratories, which participate in a three-layer quality control: (*1*) routine testing of control strains with known minimum inhibitory concentrations, (*2*) a blinded exchange of samples with another national laboratory, and (*3*) the international WHO proficiency testing (RIVM, MSLI, CDC, and TDR). PHRI and SU were chosen because they had a track record of research associated with a well-characterized clinical strain collection.

Isolate Culture, DST, and Fingerprinting Methodology

All isolates underwent DST to at least INH, RIF, ethambutol (EMB), and one of the injectable agents (AMI, KAN, and CAP).

Supported by the Bill and Melinda Gates Foundation, the Parker B. Francis Fellowship (M.R.F.), and National Institutes of Health/National Institute of Allergy and Infectious Diseases (CETR U19AI109755-01 [M.R.F. and M.M.], BD2K K01-ES026835 [M.R.F.], and U19 AI-076217 [M.M.]). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions: This study was designed and conducted by M.M., S.B., O.I., R.S., and M.R.F. M.R.F. wrote the first draft of the paper, and all authors contributed to its final version. K.J. and H.N.-G. helped with curation of the isolate phenotypes. R.M.W., E.M.S., and T.C.V. performed the molecular characterization, drug susceptibility testing (DST), and selection of isolates from South Africa. A.S. and D.K. performed molecular and DST characterization and selected isolates from Peru. J.P. performed the molecular characterization and selection of isolates from the Centers for Disease Control and Prevention (CDC). The CDC Division of Tuberculosis Elimination Reference Laboratory performed the DST. B.N.K. and N.K. performed the molecular characterization, DST, and selection of isolates from the Public Health Research Institute. D.v.S. performed the molecular characterization, DST, and selection of isolates from the Netherlands National Institute for Public Health and the Environment. L.R. contributed to the molecular characterization, DST, and Sanger sequencing of selected isolates from the World Health Organization Tropical Disease Research archive. J.G., P.S., and C.S. constructed the public user interface to access these data.

DSTs were performed using the indirect 7H10 agar proportions method (PHRI, CDC, MSLI), 7H11 agar proportions method (SU, TDR), or BACTEC MGIT 960 (RIVM). A subset of isolates was tested for pyrazinamide (PZA) resistance by BACTEC MGIT 960 (RIVM, SU, and CDC), BACTEC 450 (MSLI, CDC), and indirect 7H10 agar proportions (CDC). Molecular fingerprinting by spoligotyping, IS*6110* restriction fragment length polymorphism, or mycobacterial interspersed repetitive unit-variable number tandem repeats was performed for a subset of the isolates using standard methodology (16, 17) and lineages were identified by comparison with those from publically available databases (*see* Table E1 in the online supplement) (18, 19).

Genetic Sequencing Using MIPs

MIPs (20) were designed to cover both DNA strands of the open reading frames, promoter regions, and 100 flanking bases on either side of the 28 selected loci (*see* Figures E1–E3, Tables E2 and E3). A total of 10 ng–100 pg of DNA was extracted from sputum cultures using standard methods. Barcodes and Illumina (San Diego, CA) adapters were attached to the captured sequences during the amplification phase followed by 75-bp read parallel sequencing on an Illumina GAIIx device (*see* Figure E4). We repeated this process on isolates for which fewer than 95% of the targeted nucleotide positions were covered by at least 20 reads and we retained in the analysis only those resequenced isolates that met these criteria.

Variant Identification and Heterogeneity

We used a custom bioinformatics pipeline to clean and filter the raw reads. We aligned filtered reads to the reference MTB isolate H37Rv and included in the analysis variants called by either Bowtie (21) 0.12.7/SAMtools (22) 0.1.18 or Stampy 1.0.23 (23)/Platypus 0.5.2 (24) (*see* Table E4). We classified a variant as "heterogeneous" (i.e., representing a population of mixed bacteria) if more than one base type was present in the reads aligning to that site. We included variants in our analysis if they were present in at least 40% of reads and conducted a sensitivity analysis lowering this threshold to 10% (*see* Table E5).

Validation of MIP Sequencing Results

We assessed the sequencing performance in three ways. First, we measured the concordance between variants identified by MIP-capture and Illumina sequencing with those identified by Sanger sequencing in eight loci among 249 isolates that had been sequenced using both methods. Second, we compared MIP-identified variants with variants identified in the same regions in Illumina whole genome sequences from 40 isolates. Third, we followed up possible false-negative MIP results by performing Sanger resequencing of relevant loci in a subset of 133 isolates in which our MIP-based sequencing failed to identify variants in DR isolates.

Phylogeny Construction and Isolate Diversity

After excluding variants predictive of resistance, we constructed and annotated a neighbor-joining tree using the Phylip (25) Neighbor program and Figtree v1.4.0. We classified isolates into three principal genetic groups on the basis of mutations in the genes *gyrA* and *katG* as described by Sreevatsan and coworkers (26). Strain diversity was measured using the Kimura two-parameter model as implemented by MEGA6 (27). Mutations in the sequenced DR genes that were previously determined to be lineage defining were also assessed (*see* Table E6).

Univariate Associations

We tested for an association between nonsynonymous and presumptive promoter variants and the DR phenotype to specific drugs using parallel Fisher exact tests with a Bonferroni correction.

Random Forest Modeling and Validation

For the full prediction model, we excluded mutations if they were silent, occurred only in sensitive isolates or a single resistant isolate, and if they were one of the following variants known not to code for resistance: *gyrA*: E21T, S95T, G668D, and *katG*: R463L (28–30). We performed a sensitivity analysis including singleton mutations and the accuracy of the resistance prediction was similar (*see* Table E6). For drugs other than ofloxacin and paraaminosalicylic acid (PAS), we randomly split the data into training and validation sets containing 67% and 33% of the isolates, respectively. Because of the low numbers of isolates resistant to either ofloxacin or PAS, we developed predictions for these drugs using the entire isolate set and measured the prediction error using a 10-fold cross-validation procedure (31).

Random forest predictive modeling was performed using R version 2.15.2 and randomForest R package version 4.6.7. The randomForest classwt variable was varied to maximize the sum of sensitivity and specificity (*see* online supplement). The

Table 1. Isolate Resistance and Genes Sequenced by Drug

Drug	Resistant	Sensitive	Genes Sequenced
INH	1,219	136	*katG, inhA* (+promoter), *fabG1*, *embB*, *kasA*, *ahpC* (+promoter), *oxyR'*, *iniA*, *iniB*, *iniC*, *ndh*
RIF	1,163	206	*rpoB*
EMB	914	416	*embB*, *embA*, *embC*, *iniA*, *iniB*, *iniC*
PZA	611	374	*pncA*
SM	941	414	*rpsL*, *rrs*, *gid*
ETH	612	374	*ethA*, *inhA* (+ promoter)
CIP	215	695	*gyrA*, *gyrB*
LEVO	110	437	
OFLX	69	201	
AMK	228	729	*rrs*, *rrl*
KAN	257	631	
CAP	577	363	*rrs*, *rrl*, *tlyA*
PAS	78	849	*thyA*
CYS	8	855	*alr*, *ddl*
Total	1,397		

Definition of abbreviations: AMK = amikacin; CAP = capreomycin; CIP = ciprofloxacin; CYS = cycloserine; EMB = ethambutol; ETH = ethionamide; INH = isoniazid; KAN = kanamycin; LEVO = levofloxacin; OFLX = ofloxacin; PAS = paraaminosalicylic acid; PZA = pyrazinamide; RIF = rifampicin; SM = streptomycin.

weighted model was then run with serially smaller subsets of mutations, eliminating one variable at a time in increasing order of importance. We used the unscaled permutation mean decrease in accuracy as our measure of variable importance (32, 33). We ran the serial models on 100 bootstrap samples of the training sets for each drug (34). For each bootstrap sample, a candidate minimum set of mutations was identified when any further removal of a mutation resulted in a decrease of more than one SD from the model's bootstrapped mean accuracy. The consensus minimum number of variables were those variables that we selected in most (>50%) of the bootstrap replicates for each drug. We finally constructed 1,000 tree random forest using this final set of variables for each drug and this constituted our final model. We calculated the SD of the sensitivity and specificity of full and minimal models by 100-fold boostrapping. We validated our classification of predictive mutations by comparing with mutation lists previously defined as lineage defining and likely benign (*see* Table E6).

Additional sequencing and method description is provided in the online supplement.

Public Database and Prediction Tool

We created a public data-sharing tool (http://www.broadinstitute.org/annotation/genome/mtb_drug_resistance.1/DirectedSequencingHome.html) that includes the genetic data and DR phenotypes. The resistance prediction model is provided with the online supplement.

Results

Phenotypic Drug Resistance Profiles

Isolates underwent culture-based DST to a median of 11 drugs (Table 1; *see* Table E1), 78 (6%) were fully drug sensitive, 141 (10%) were resistant to one or more first-line drugs but not to both INH and RIF, and 1,130 (81%) were resistant to both INH and RIF. Of the MDR isolates, 51% were also resistant to PZA, 62% to EMB, 23% to at least one FLQ, and 53% to at least one second-line injectable. Nineteen percent of the MDR isolates were XDR (i.e., also resistant to both an FLQ and an injectable) (*see* Table E8).

Table 2. Most Frequent Variants by Region

Drug	Gene	Base Position	Codon	Resistant Isolates with Mutation [*n* (%)]	Sensitive Isolates with Mutation [*n* (%)]
FLQ (CIP or OFLX)	***gyrA***	269	90	45 (16)	8 (1)
		280*	94	19 (7)	1 (0.1)
		281*	94	94 (33)	14 (2)
RIF	***rpoB***†	1303	435	25 (2)	4 (2)
		1304	435	146 (13)	1 (0.5)
		1333*	445	98 (8)	6 (3)
		1334	445	76 (7)	1 (0.5)
		1348	450	6 (0.5)	0
		1349	450	767 (66)	5 (2)
		2083	695	82 (7)	5 (2)
SM	*rpsL*	128	43	225 (24)	5 (1)
		262	88	1 (0.1)	0
		263*	88	67 (7)	2 (0.4)
	Gid	276	92	211 (22)	56 (13)
		275	92	1 (0.1)	0
		274*	92	0	1 (0.2)
	Rrs	513	—	19 (2)	1 (0.2)
		517	—	79 (8)	3 (0.7)
AG (AMK)	*Rrs*	1401	—	184 (82)	17 (2)
INH	**promoter-*inhA***	−15	—	265 (22)	2 (1)
	katG	943	315	2 (0.1)	0
		944*	315	909 (74)	3 (2)
		945	315	47 (4)	0
	kasA	805	269	209 (17)	7 (5)
	ahpC	146	49	162 (13)	4 (3)
	iniB	208	70	72 (6)	1 (0.7)
EMB	***embB***	916	306	285 (31)	14 (3)
		918*	306	258 (28)	28 (6)
		1216	406	40 (4)	6 (2)
		1217*	406	97 (10)	22 (5)
	embC	2320	774	97 (11)	27 (6)

Definition of abbreviations: AG = aminoglycosides; AMK = amikacin; CIP = ciprofloxacin; EMB = ethambutol; FLQ = fluoroquinolone; INH = isoniazid; OFLX = ofloxacin; RIF = rifampicin; SM = streptomycin.

*Two or more nonreference alleles were present at the same base position. Regions currently targeted by commercial molecular diagnostics (6, 49) are shown in bold. Here, we only include mutations that were more prevalent in resistant versus sensitive isolates and exclude variants with a frequency of <5% per codon or noncoding site in resistant isolates (except for rrs mutations in relation to SM resistance, for which we include the two most common mutations). †H37Rv rpoB codon numbering used here. Table E18 provides a conversion to *Escherichia coli* numbering. Table E19 details the variants by laboratory and DST method.

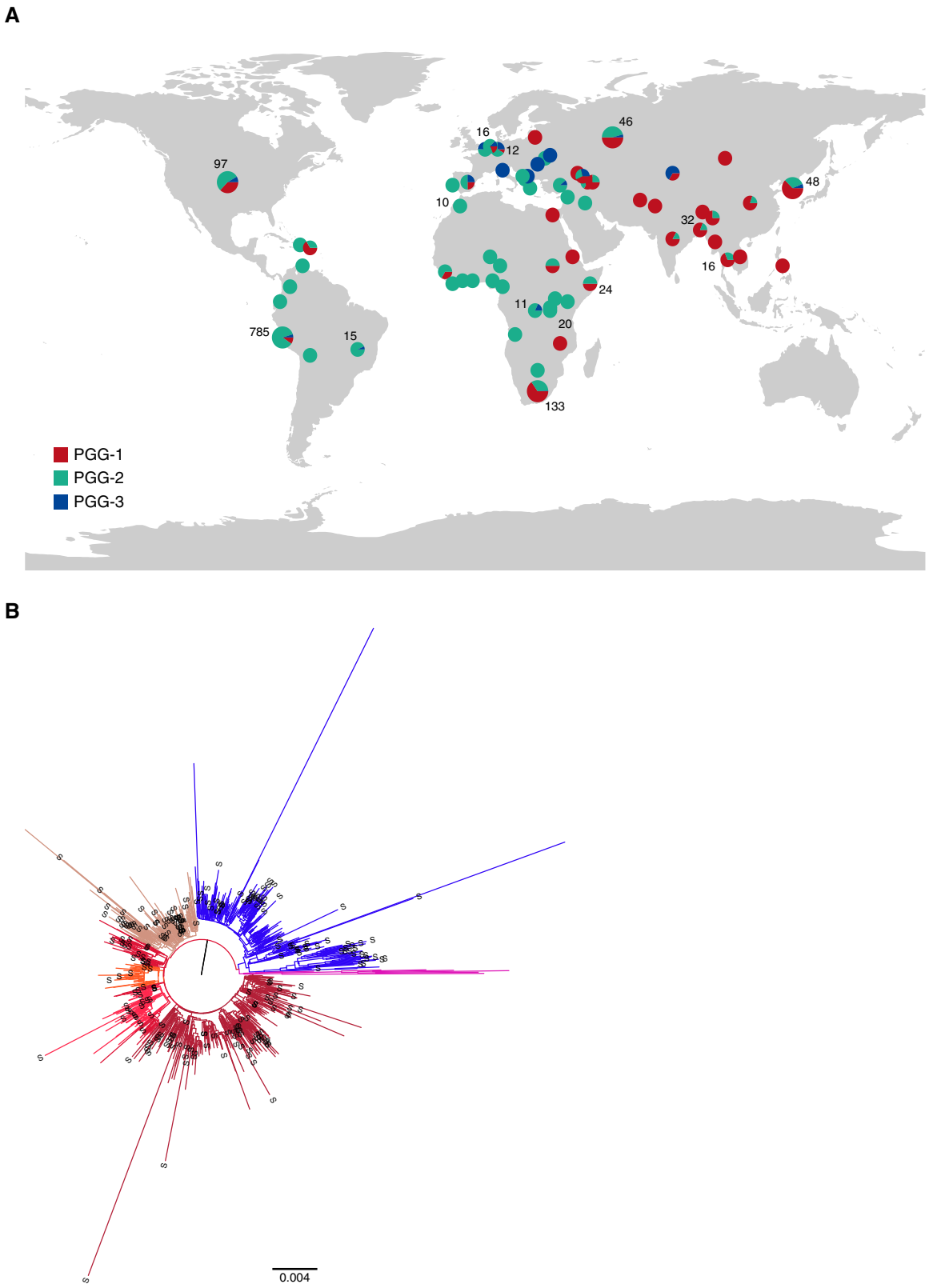

Figure 1. Geographic and genetic diversity of the isolates. (*A*) Isolate principal genetic group (PGG) (26) by country of origin. The digits represent the number of isolates collected from each country. Isolate numbers less than 10 are not displayed. (*B*) Neighbor-joining phylogenetic tree of *Mycobacterium tuberculosis* isolates. All *red tones* belong to the Europe–Africa–Americas lineage 4 (*taupe*, Haarlem; *magenta*, X/low-copy clade; *orange*, T; *maroon*, Latin American–Mediterranean). *Blue*, East Asia lineage 2 (e.g., Beijing); *purple*, East Africa and India lineage 3 (e.g., Central Asian Strain). S = sensitive.

MIP Sequencing

We selected 26 putative or known resistance genes and two promoter regions through a literature review (35) and consultation with experts (Table 1). We designed MIPs (20) to sequence these regions (*see* Figures E1–E3, Tables E2 and E3) because of the expected higher depth of MIP sequencing relative to whole genome sequencing (WGS) (20). Of 1,748 isolates sequenced with MIPs, 351 isolates were excluded because less than 95% of their bases were covered by 20 or more reads. In the remaining 1,397 isolates, the MIPs amplified uniformly with 85% producing between 100 and 1,000 reads (*see* Figure E5). Overall, MIPs captured an average of 99.9% of the targeted bases, and an average of 97.1% of the bases were covered with at least 20 reads (*see* Table E9).

In validation experiments, MIP-based sequencing captured all variants called by Sanger in 99% of the isolates (n = 249) and 100% of variants identified by WGS (40 isolates). MIPs also captured 84 additional variants not identified by WGS in these isolates. More than 95% of these variants were missed by WGS because of low coverage (*see* Table E10). Among the 133 isolates for which MIPs did not identify a relevant variant, 115 of 133 (87%) of the MIP results were confirmed by Sanger sequencing (*see* Table E11).

Gene Diversity

We targeted 42,367 bases for sequencing in the 1,397 isolates and identified 30,747 genetic variants starting at 2,673 distinct genomic sites (Table 2; *see* Figure E6). Of these variants, 5,987 (19%) were heterogeneous (i.e., detected at a read frequency of 40–95%; mean, 61%), and 24,760 were called with greater than 95% purity. Seventy percent of the variants (21,655) were protein-altering or occurred in promoter/intergenic regions (*see* Table E12).

Isolate Diversity

Among the isolates sequenced, 785 isolates (56%) originated from Peru, 133 (10%) were from South Africa, 97 (7%) were from the United States, 48% (3%) were from Korea, and the remaining 334 were from 63 other countries (Figure 1). Among the 509 isolates for which molecular fingerprints were available, 25% belonged to the Latin American–Mediterranean lineage, 22% to Beijing, 21% to T, and the remaining 32% to other lineages. Sensitive isolates were evenly distributed across MTB lineages (Figure 1; *see* Table E6). After we excluded DR-associated variation, the median pairwise genetic distance was 3.1 substitutions/10 kbp (interquartile range, 2.3–3.8) across the 42 kbp sequenced.

Univariate Associations

We found univariate associations between 47 genetic variants and a DR phenotype (*see* Table E13). These include many of the known resistance mutations and the following novel associations that reached statistical significance: the *iniB* A70T and *embA* N54D mutations and EMB resistance, and the *embB* M306I and M306V mutations and INH resistance even after stratification by the EMB resistance status (*see* Table E14). We also found strong associations between the *thyA* H207R and L8Q mutations and PAS resistance and between the *embA/B* promoter region and both INH and EMB resistance (*see* Table E15). We noted more than 800 novel variants (35) (*see* Table E16) that occurred more often in resistant than sensitive isolates, but these associations did not reach statistical significance.

Diagnostic Performance

Table 3 gives the sensitivity and specificity of the full and minimal genetic models for the prediction of the resistance phenotype. For PZA, the large number of very rare variants that contributed to resistance prediction meant that the minimal set of predictive mutations could not be chosen reliably. The final list of mutations encompassed 124 of the 127 nonsynonymous variants we observed in the *pncA* gene and promoter yet still underperformed in the validation set of isolates. For the other drugs, the minimal set of genetic variants predicted resistance

Table 3. Genetic Predictive Model Performance

	Mutations Included	All Variables on Learning		Selected Mutation	Selected Variables on Learning Isolate Set		Selected Variables on Validation Isolate Set	
		Sensitivity (%)	Specificity (%)	Number	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
INH	220	96 ± 1	98 ± 2	18	95 ± 1	98 ± 2	94 ± 1	94 ± 3
RIF	85	93 ± 1	98 ± 1	14	92 ± 1	98 ± 1	93 ± 1	95 ± 2
PZA	127	72 ± 2	97 ± 1	124	72 ± 3	96 ± 1	64 ± 3	92 ± 3
EMB	126	84 ± 2	91 ± 2	18	83 ± 2	89 ± 2	80 ± 2	82 ± 3
STR*	176	65 ± 2	97 ± 1	37	61 ± 2	97 ± 1	54 ± 3	94 ± 2
ETH	110	65 ± 2	92 ± 2	20	55 ± 3	90 ± 2	54 ± 3	89 ± 3
KAN	19	66 ± 4	98 ± 1	2	62 ± 4	99 ± 0.5	66 ± 5	98 ± 1
CAP	66	43 ± 3	96 ± 1	5	38 ± 2	96 ± 1	38 ± 3	95 ± 2
AMK	47	85 ± 3	98 ± 1	2	82 ± 3	98 ± 1	79 ± 5	97 ± 1
CIP	26	56 ± 4	98 ± 1	7	52 ± 5	99 ± 0.4	51 ± 5	100 ± 0.0
LEVO	18	77 ± 5	99 ± 0.3	8	74 ± 5	99 ± 0.4	63 ± 9	99 ± 1
OFLX†	19	83 ± 5	88 ± 3	6	77 ± 5	90 ± 2	74 ± 15	90 ± 6
PAS†	13	18 ± 5	99 ± 0.3	4	14 ± 5	99 ± 0.2	13 ± 9	99 ± 1

Definition of abbreviations: AMK = amikacin; CAP = capreomycin; CIP = ciprofloxacin; EMB = ethambutol; ETH = ethionamide; INH = isoniazid; KAN = kanamycin; LEVO = levofloxacin; OFLX = ofloxacin; PAS = paraaminosalicylic acid; PZA = pyrazinamide; RIF = rifampicin; STR = streptomycin. Bootstrap SEs are reported.

*For STR we also ran the prediction model after removal of gid_E92D. This resulted in a decrease in the sensitivity of prediction model by 2% but no change in the specificity.

†Tenfold cross-validation results shown for OFLX and PAS in seventh and eighth columns.

in the validation set with equivalent sensitivity and specificity as the full model (Table 3, Figure 2).

The model predicted INH resistance with 96% (±1%) sensitivity for MDR isolates but only 84% (±4%) sensitivity for monoresistant isolates. *katG* 315T mutations were less frequent and *inhA* -15T mutations more common in mono-INH resistant than in MDR isolates (42 vs. 73%, $P = 4 \times 10^{-8}$, and 30 vs. 21%, $P = 0.07$, respectively) (*see* Figure E7).

The minimal lists of predictive mutations included the following novel variants not previously recognized as diagnostically relevant: *embA/B* promoter, and the *ahpC*, *iniB*, and *gyrB* genes (*see* Table E17, Figure E8). Mutations excluded from the lists and their distribution are provided in Table E7. Twenty-four mutations were previously determined to be lineage defining (12, 36, 37) and were in a region sequenced in this study. Of these *gid E92D* was classified as

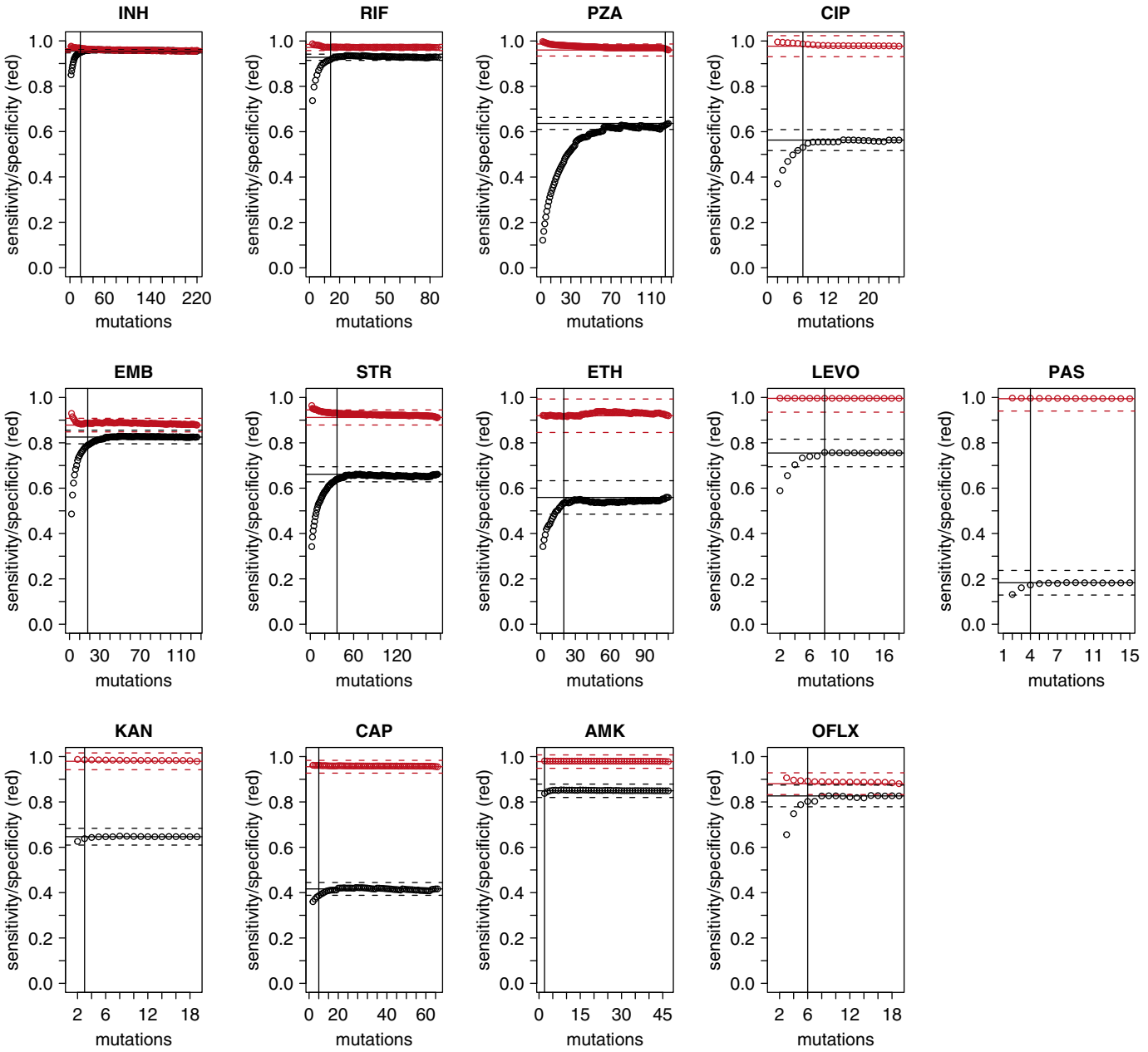

Figure 2. Diagnostic performance of serially pared predictive models by drug. Sensitivity is plotted in *black*, specificity in *red*. *Dashed lines* represent ±1 SD. *Vertical lines* represent the minimal set of predictive mutations chosen (in an automated step-down fashion) beyond which the sensitivity drops >1 SD from the mean sensitivity for the full model. AMK = amikacin; CAP = capreomycin; CIP = ciprofloxacin; EMB = ethambutol; ETH = ethionamide; INH = isoniazid; KAN = kanamycin; LEVO = levofloxacin; OFLX = ofloxacin; PAS = paraaminosalicylic acid; PZA = pyrazinamide; RIF = rifampicin; STR = streptomycin.

predictive and the 23 others were classified as nonpredictive of drug resistance (*see* Table E6).

Discussion

This analysis of almost 1,400 comprehensively sampled MTB clinical isolates, including more than 1,100 MDR isolates, has expanded the list of genetic determinants for drug resistance. The large number of genetic determinants found (238 mutations in 18 genetic loci) emphasizes that future MTB drug resistance diagnostics need to allow for a high dimension of mutation detection. This may render WGS technology the most attractive approach especially as it becomes more affordable and more readily available even in resource-limited settings.

Our analysis also shows that although the genetic determinants of resistance to RIF and INH are well defined, the full complement of mutations encoding resistance to other first- and second-line drugs is not yet established. These findings support previous work showing that rapid diagnostic tests for detecting mutations that confer resistance to INH and RIF are highly sensitive and specific but those targeting other drugs require further optimization if they are to replace conventional DSTs (6, 38, 39).

Several possible mechanisms may account for this sensitivity gap. First, there are likely as-yet-undetected DR loci and epistatic effects that code resistance to one or more drugs. Genome-wide analysis studies may identify these targets in the near future. Here we focused only on genes known or suspected to be associated with resistance, but we nevertheless identified multiple novel variants associated with clinical DR.

Second, some discrepancies may be caused by errors in "gold standard" DSTs. For example, the reproducibility of DST for some agents, such as PZA and EMB, is low, and results vary both by laboratory and technician (40). We tried to limit these discrepancies by choosing isolates well-characterized with respect to DR tested in national and supranational reference laboratories using WHO-recommended methods (5), but it is possible that some discrepancies remain and account for the low sensitivity and specificity of targeted sequencing for these more problematic drugs. It was not possible to retest all isolates that had discordant genotype and phenotype results because of the large number of isolates resistant to second-line drugs in these study. We did observe a DST false-negative rate of 0.1–6% as determined by the frequency of isolates that were phenotypically sensitive and found to have canonical resistance mutations (indicated by genes in bold in Table 2). Although factors that determine the false-positive rate are somewhat different, a false-positive rate of a similar magnitude to the observed false-negative rate is unlikely to explain most of the genotypic sensitivity gap.

Third, despite the high depth of our MIP sequencing, it is possible that minority resistant bacterial populations that resulted in a resistant DST were not adequately amplified and sequenced. Finally, it is possible that some genetic variants that lead to antibiotic resistance may involve rearrangements or recombination events that are not detected by the sequencing tools used here, which yield short DNA sequence reads optimized for detecting short nucleotide polymorphisms rather than these structural changes. It is well documented that rearrangements (41) can lead to resistance to chemotherapeutic drugs used to treat human malignancies and that resistance to antibiotics can result from large duplications that result in increased gene dosage (42).

Although our results are consistent with several previous reports on targeted sequencing of DR-TB, some of these have reported higher sensitivities for specific drugs. For example, two other groups obtained higher sensitivities for KAN because they included the *eis* gene among the loci sequenced (38, 39). *Eis* had not been identified as a resistance-associated gene at the time our study began, but mutations in this locus have since been found to explain up to 20% of KAN resistance (43). Other recently identified resistance genes in MTB include *panD* and *rpsA*, which was reported to confer PZA resistance in isolates that lack *pncA* mutations (12, 44, 45). Previous studies have also focused exclusively on either MDR (39) or XDR (38) isolates, which may have a narrower range of resistance mutations. This is supported by our observation of a lower genotypic sensitivity for INH resistance in monoresistant as compared with MDR isolates.

Although some of the variants associated with resistance phenotypes may cause resistance, others are likely to be mutations that interact with a causative mutation or compensate for its fitness cost. For example, one study showed that mutations in *rpoC* ameliorated fitness costs incurred by RIF resistance mutations in *rpoB* (46). Even if these mutations do not themselves confer resistance, it may be useful to include them in molecular diagnostic tools if they reliably predict resistance. In this study, we oversampled DR isolates to detect rarer genetic determinants and develop a more sensitive genotypic prediction model. This was at the expense of undersampling isolates sensitive to INH and RIF and may negatively impact the specificity of the resistance variants selected for these two drugs. Despite this oversampling, the variant-based model's specificity for these two drugs was validated at greater than or equal to 94% on an independent set of patient isolates. For all other drugs studied, at least 31% of the sample were phenotypically sensitive.

We do expect the sensitivity and specificity gaps to close as more clinical and research teams move to routine WGS of resistant isolates (47). The success of this endeavor depends on creation of public databases pooling data across laboratories and geographic regions and on the further refinement of predictive models similar to that proposed here that can update DR predictions as soon as new data become available (48). With WGS and user-friendly public databases, we expect that it will be possible to conduct routine diagnosis of resistance to the full spectrum of TB drugs, thereby allowing effective individualized treatment for DR-TB. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Acknowledgment: The authors thank Dr. Nancy Cook from the Brigham and Women's Hospital in Boston for providing biostatistical input.

References

1. World Health Organization. Global tuberculosis report 2014 [accessed 2015 Sept]. Available from: http://www.who.int/tb/publications/global_report/en/
2. Small PM, Pai M. Tuberculosis diagnosis: time for a game change. *N Engl J Med* 2010;363:1070–1071.
3. World Health Organization. Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 global report on surveillance and response [accessed 2015 Sept]. Available from: http://www.who.int/tb/features_archive/m_xdrtb_facts/en/index.html
4. Horne DJ, Pinto LM, Arentz M, Lin S-YG, Desmond E, Flores LL, Steingart KR, Minion J. Diagnostic accuracy and reproducibility of WHO-endorsed phenotypic drug susceptibility testing methods for first-line and second-line antituberculosis drugs. *J Clin Microbiol* 2013;51:393–401.
5. World Health Organization. Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis. 2014 [accessed 2015 Sept]. Available from: http://apps.who.int/iris/bitstream/10665/130918/1/9789241548809_eng.pdf?ua=1&ua=1
6. Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, Allen J, Tahirli R, Blakemore R, Rustomjee R, *et al*. Rapid molecular detection of tuberculosis and rifampin resistance. *N Engl J Med* 2010;363:1005–1015.
7. Brossier F, Veziris N, Aubry A, Jarlier V, Sougakoff W. Detection by GenoType MTBDRsl test of complex mechanisms of resistance to second-line drugs and ethambutol in multidrug-resistant Mycobacterium tuberculosis complex isolates. *J Clin Microbiol* 2010;48:1683–1689.
8. Hanrahan CF, Dorman SE, Erasmus L, Koornhof H, Coetzee G, Golub JE. The impact of expanded testing for multidrug resistant tuberculosis using genotype [correction of geontype] MTBDRplus in South Africa: an observational cohort study. *PLoS One* 2012;7:e49898.
9. Miotto P, Cirillo DM, Migliori GB. Drug resistance in *Mycobacterium tuberculosis*: molecular mechanisms challenging fluoroquinolones and pyrazinamide effectiveness. *Chest* 2015;147:1135–1143.
10. Jin J, Shen Y, Fan X, Diao N, Wang F, Wang S, Weng X, Zhang W. Underestimation of the resistance of *Mycobacterium tuberculosis* to second-line drugs by the new GenoType MTBDRsl test. *J Mol Diagn* 2013;15:44–50.
11. Tagliani E, Cabibbe AM, Miotto P, Borroni E, Toro JC, Mansjö M, Hoffner S, Hillemann D, Zalutskaya A, Skrahina A, *et al*. Diagnostic performance of the new version (v2.0) of GenoType MTBDRsl assay for detection of resistance to fluoroquinolones and second-line injectable drugs: a multicenter study. *J Clin Microbiol* 2015;53:2961–2969.
12. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, *et al*.; Modernizing Medical Microbiology (MMM) Informatics Group. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 2015;15:1193–1202.
13. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, *et al*. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* 2015;7:51.
14. Farhat M, Sultana R, Murray M. Large scale sequencing of genetic determinants of drug resistance in *Mycobacterium tuberculosis*: implications for diagnostic design [abstract]. *Am J Respir Crit Care Med* 2015;191:A2184.
15. Vincent V, Rigouts L, Nduwamahoro E, Holmes B, Cunningham J, Guillerm M, Nathanson C-M, Moussy F, De Jong B, Portaels F, *et al*. The TDR Tuberculosis Strain Bank: a resource for basic science, tool development and diagnostic services. *Int J Tuberc Lung Dis* 2012;16:24–31.
16. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, *et al*. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* 1997;35:907–914.
17. van Soolingen D, Hermans PW, de Haas PE, Soll DR, van Embden JD. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol* 1991;29:2578–2586.
18. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, *et al*. *Mycobacterium tuberculosis* complex genetic diversity: mining the Fourth International Spoligotyping Database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol* 2006;6:23.
19. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. *Nucleic Acids Res* 2010;38:W326–331.
20. Hardenbol P, Banér J, Jain M, Nilsson M, Namsaraev EA, Karlin-Neumann GA, Fakhrai-Rad H, Ronaghi M, Willis TD, Landegren U, *et al*. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* 2003;21:673–678.
21. Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* 2010;Chapter 11:Unit 11.7.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
23. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 2011;21:936–939.
24. Rimmer A, Mathieson I, Lunter G, McVean G. Wellcome Trust Centre for Human Genetics - Platypus. Platypus: an integrated variant caller. 2012 [accessed 2013 Oct]. Available from: http://www.well.ox.ac.uk/platypus
25. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989;5:164–166.
26. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA* 1997;94:9869–9874.
27. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;28:2731–2739.
28. Nebenzahl-Guimaraes H, Jacobson KR, Farhat MR, Murray MB. Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in *Mycobacterium tuberculosis*. *J Antimicrob Chemother* 2014;69:331–342.
29. Meacci F, Orrù G, Iona E, Giannoni F, Piersimoni C, Pozzi G, Fattorini L, Oggioni MR. Drug resistance evolution of a *Mycobacterium tuberculosis* strain from a noncompliant patient. *J Clin Microbiol* 2005;43:3114–3120.
30. Maruri F, Sterling TR, Kaiga AW, Blackman A, van der Heijden YF, Mayer C, Cambau E, Aubry A. A systematic review of gyrase mutations associated with fluoroquinolone-resistant *Mycobacterium tuberculosis* and a proposed gyrase numbering system. *J Antimicrob Chemother* 2012;67:819–831.
31. Ewout W. Steyerberg. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer-Verlag; 2009.
32. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics* 2008;9:307.
33. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006;7:3.
34. Chen SL, Hung C-S, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P, *et al*. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc Natl Acad Sci USA* 2006;103:5977–5982.
35. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. *PLoS Med* 2009;6:e2.
36. Feuerriegel S, Köser CU, Niemann S. Phylogenetic polymorphisms in antibiotic resistance genes of the *Mycobacterium tuberculosis* complex. *J Antimicrob Chemother* 2014;69:1205–1210.
37. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* 2014;5:4812.

38. Rodwell TC, Valafar F, Douglas J, Qian L, Garfein RS, Chawla A, Torres J, Zadorozhny V, Kim MS, Hoshide M, *et al*. Predicting extensively drug-resistant *Mycobacterium tuberculosis* phenotypes with genetic mutations. *J Clin Microbiol* 2014;52:781–789.
39. Campbell PJ, Morlock GP, Sikes RD, Dalton TL, Metchock B, Starks AM, Hooks DP, Cowan LS, Plikaytis BB, Posey JE. 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; 55:2032–2041.
40. World Health Organization (WHO). A roadmap for ensuring quality tuberculosis diagnostics services within national laboratory strategic plans. 2010 [accessed 2015 Sept]. Available from: http://www.who.int/tb/laboratory/gli_roadmap.pdf
41. Huff LM, Lee J-S, Robey RW, Fojo T. Characterization of gene rearrangements leading to activation of MDR-1. *J Biol Chem* 2006; 281:36501–36509.
42. Sandegren L, Andersson DI. Bacterial gene amplification: implications for the evolution of antibiotic resistance. *Nat Rev Microbiol* 2009;7:578–588.
43. Zaunbrecher MA, Sikes RD Jr, Metchock B, Shinnick TM, Posey JE. Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 2009;106:20004–20009.
44. Shi W, Zhang X, Jiang X, Yuan H, Lee JS, Barry CE III, Wang H, Zhang W, Zhang Y. Pyrazinamide inhibits trans-translation in *Mycobacterium tuberculosis*. *Science* 2011;333:1630–1632.
45. Shi W, Chen J, Feng J, Cui P, Zhang S, Weng X, Zhang W, Zhang Y. Aspartate decarboxylase (PanD) as a new target of pyrazinamide in *Mycobacterium tuberculosis*. *Emerg Microbes Infect* 2014;3: e58.
46. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* 2012; 44:106–110.
47. Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, Carmichael AJ, Parkhill J, Smith GP, Peacock SJ. Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *N Engl J Med* 2013;369:290–292.
48. Shafer RW. Rationale and uses of a public HIV drug-resistance database. *J Infect Dis* 2006;194:S51–S58.
49. Huang W-L, Chi T-L, Wu M-H, Jou R. Performance assessment of the GenoType MTBDRsl test and DNA sequencing for detection of second-line and ethambutol drug resistance among patients infected with multidrug-resistant *Mycobacterium tuberculosis*. *J Clin Microbiol* 2011;49:2502–2508.