Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance

Manson, Abigail L.; Cohen, Keira A; Abeel, Thomas; Desjardins, Christopher A.; Armstrong, Derek T.; Barry III, Clifton E.; Brand, Jeannette; Chapman, Sinéad B.; Cho, Sang-Nae; Gabrielian, Andrei

DOI
10.1038/ng.3767

Publication date
2017

Document Version
Final published version

Published in
Nature Genetics

nature genetics

Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance

Abigail L Manson[1,27], Keira A Cohen[1–3,27], Thomas Abeel[1,4], Christopher A Desjardins[1], Derek T Armstrong[5], Clifton E Barry III[6], Jeannette Brand[7], TBResist Global Genome Consortium[8], Sinéad B Chapman[1], Sang-Nae Cho[9], Andrei Gabrielian[10], James Gomez[1], Andreea M Jodals[11], Moses Joloba[12], Pontus Jureen[13], Jong Seok Lee[9], Lesibana Malinga[7], Mamoudou Maiga[14], Dale Nordenberg[15], Ecaterina Noroc[16], Elena Romancenco[16], Alex Salazar[1,4], Willy Ssengooba[12], A A Velayati[17], Kathryn Winglee[5], Aksana Zalutskaya[18], Laura E Via[6], Gail H Cassell[19], Susan E Dorman[5], Jerrold Ellner[20], Parissa Farnia[17], James E Galagan[1,21], Alex Rosenthal[10], Valeriu Crudu[16], Daniela Homorodean[11], Po-Ren Hsueh[22], Sujatha Narayanan[23], Alexander S Pym[2], Alena Skrahina[18], Soumya Swaminathan[23], Martie Van der Walt[7], David Alland[24], William R Bishai[2,5], Ted Cohen[25,26], Sven Hoffner[13], Bruce W Birren[1] & Ashlee M Earl[1]

Multidrug-resistant tuberculosis (MDR-TB), caused by drug-resistant strains of *Mycobacterium tuberculosis*, is an increasingly serious problem worldwide. Here we examined a data set of whole-genome sequences from 5,310 *M. tuberculosis* isolates from five continents. Despite the great diversity of these isolates with respect to geographical point of isolation, genetic background and drug resistance, the patterns for the emergence of drug resistance were conserved globally. We have identified harbinger mutations that often precede multidrug resistance. In particular, the *katG* mutation encoding p.Ser315Thr, which confers resistance to isoniazid, overwhelmingly arose before mutations that conferred rifampicin resistance across all of the lineages, geographical regions and time periods. Therefore, molecular diagnostics that include markers for rifampicin resistance alone will be insufficient to identify pre-MDR strains. Incorporating knowledge of polymorphisms that occur before the emergence of multidrug resistance, particularly *katG* p.Ser315Thr, into molecular diagnostics should enable targeted treatment of patients with pre-MDR-TB to prevent further development of MDR-TB.

Drug-resistant *M. tuberculosis* is a threat to global TB control efforts. Failure to identify and appropriately treat patients with drug-resistant TB can lead to increased mortality, nosocomial outbreaks and the expansion of drug resistance[1]. Five percent of *M. tuberculosis* cases worldwide are now MDR, which is defined as having resistance to both isoniazid and rifampicin[2]. Therapeutic regimens for MDR-TB can exceed 18 months and include agents that often entail significant adverse effects[3]. As of the present, 0.5% of global TB cases are considered extensively drug resistant (XDR), which is defined as MDR with additional resistance both to fluoroquinolones and at least one second-line

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]KwaZulu-Natal Research Institute for TB and HIV (K-RITH), Durban, South Africa. [3]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [4]Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands. [5]Center for Tuberculosis Research, Johns Hopkins University, Baltimore, Maryland, USA. [6]National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, Maryland, USA. [7]Medical Research Council, TB Platform, Pretoria, South Africa. [8]A full lists of members and affiliations appear at the end of the paper. [9]International Tuberculosis Research Center, Changwon and Department of Microbiology, Yonsei University College of Medicine, Seoul, South Korea. [10]Office of Cyber Infrastructure and Computational Biology, National Institutes of Health, Rockville, Maryland, USA. [11]Clinical Hospital of Pneumology Leon Daniello, Cluj Napoca, Romania. [12]Department of Medical Microbiology, Mycobacteriology Laboratory, Makerere University, Kampala, Uganda. [13]Public Health Agency of Sweden, Solna, Sweden. [14]University of Sciences, Techniques and Technologies of Bamako (USTTB), Bamako, Mali. [15]Novasano Health and Science, New York, New York, USA. [16]Microbiology and Morphology Laboratory, Phthisiopneumology Institute, Chisinau, Moldova. [17]Mycobacteriology Research Centre, National Research Institute of Tuberculosis and Lung Disease (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran. [18]Republican Research and Practical Centre for Pulmonology and Tuberculosis, Minsk, Belarus. [19]Department of Global Health and Social Medicine, Harvard Medical School, Division of Global Health Equity, Brigham and Women's Hospital, Boston, Massachusetts, USA. [20]Section of Infectious Diseases, Boston Medical Center, Boston, Massachusetts, USA. [21]Department of Biomedical Engineering and Microbiology, Boston University, Boston, Massachusetts, USA. [22]National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan. [23]National Institute for Research in Tuberculosis, Chennai, India. [24]Rutgers–New Jersey Medical School, Newark, New Jersey, USA. [25]Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [26]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA. [27]These authors contributed equally to this work. Correspondence should be addressed to A.M.E. (aearl@broadinstitute.org).

Received 30 June 2016; accepted 14 December 2016; published online 16 January 2017; doi:10.1038/ng.3767

injectable drug[2]. XDR-TB has incredibly poor treatment outcomes: in one long-term cohort of patients with XDR-TB in South Africa, only 19% of patients had a favorable outcome[4].

The global frontline molecular diagnostic for drug-resistant *M. tuberculosis*, Xpert MTB/RIF, simultaneously detects the presence of *M. tuberculosis* and identifies rifampicin resistance[5]. Although this assay identifies patients harboring rifampicin-resistant strains for the initiation of MDR-TB treatment, it may not identify resistance at the earliest available opportunity. In a recent analysis of genomes from a large collection of *M. tuberculosis* clinical isolates from South Africa[6], Cohen *et al.* showed that the overwhelming majority of MDR-TB and XDR-TB strains evolved resistance to isoniazid before resistance to rifampicin. This result was consistent with another recent genomic analysis of strains from Russia[7] and from an MDR-TB outbreak in Argentina[7]. In addition, analysis of phenotypic drug-susceptibility tests from a large, global collection of strains collected during TB drug resistance surveys indicated that isoniazid resistance is acquired before rifampicin resistance[8]. Furthermore, a recent meta-analysis revealed that patients harboring isoniazid-resistant strains have higher rates of treatment failure, relapse and acquisition of multidrug resistance relative to patients with drug-susceptible strains[9].

Collectively, these results suggest that, to detect resistance as soon as possible and to prevent MDR-TB and XDR-TB strains from evolving, molecular diagnostic tests for *M. tuberculosis* should include the earliest resistance-conferring mutations to emerge; however, the identities of these MDR 'harbinger' mutations remain undefined. To close this gap in understanding, we undertook a large-scale analysis of a global data set of whole-genome sequences from 5,310 *M. tuberculosis* strains, including 868 newly sequenced strains and 4,442 previously published strains, to determine the order of acquisition of drug-resistance mutations and to identify which mutations occur early along the pathway toward MDR and which might, therefore, serve as early sentinels in the development of MDR.

RESULTS

Drug resistance arises by similar mechanisms across the globe

To examine global phylogeographical patterns, including the order of evolution of drug-resistance mutations in *M. tuberculosis*, we compiled a set of 8,316 whole-genome sequences from clinical *M. tuberculosis* strains that were either newly sequenced as part of this study or were sequenced as part of 14 published studies that used Illumina technology (**Supplementary Table 1**)[6,10–23]. After quality filtering (Online Methods), our data set included 5,310 genome sequences that represented *M. tuberculosis* strains from 48 countries and 17 United Nations (UN)-defined geographical regions (**Supplementary Figs. 1–4**, **Supplementary Tables 1–3** and **Supplementary Note**). Although our data set represented a broad diversity of TB strains from many global regions, the phylogeographical distribution of the strains did not perfectly match the actual distribution of TB burden worldwide (**Supplementary Fig. 1b**); however, all seven known global lineages of *M. tuberculosis*[24] were represented (lineage 1, EAI or Indo-Oceanic lineage; lineage 2, Beijing lineage; lineage 3, CAS or Central Asian lineage; lineage 4, Euro-American lineage; lineage 5, *Mycobacterium africanum* West African type I; lineage 6, *M. africanum* West African type II; and the deep-branching lineage 7), as well as *Mycobacterium bovis*. As expected, lineages 1–4 were predominant (99.2%), consistent with the previously described limited geographical and host distributions of lineages 5–7 (**Supplementary Table 3**)[22,25,26].

To examine the distribution of drug resistance in our sample, for each of the isolate genomes, we computationally predicted resistance to eight drugs[27] using a curated list of polymorphisms associated with resistance (**Supplementary Table 4**). Because phenotypic drug-resistance information was unavailable for most of the data sets, we did not incorporate phenotypic information into our analysis. We identified a total of 392 unique drug-resistance-associated polymorphisms in at least one strain (**Supplementary Tables 2** and **5**). Relative to the expected global rates for resistance, we observed higher

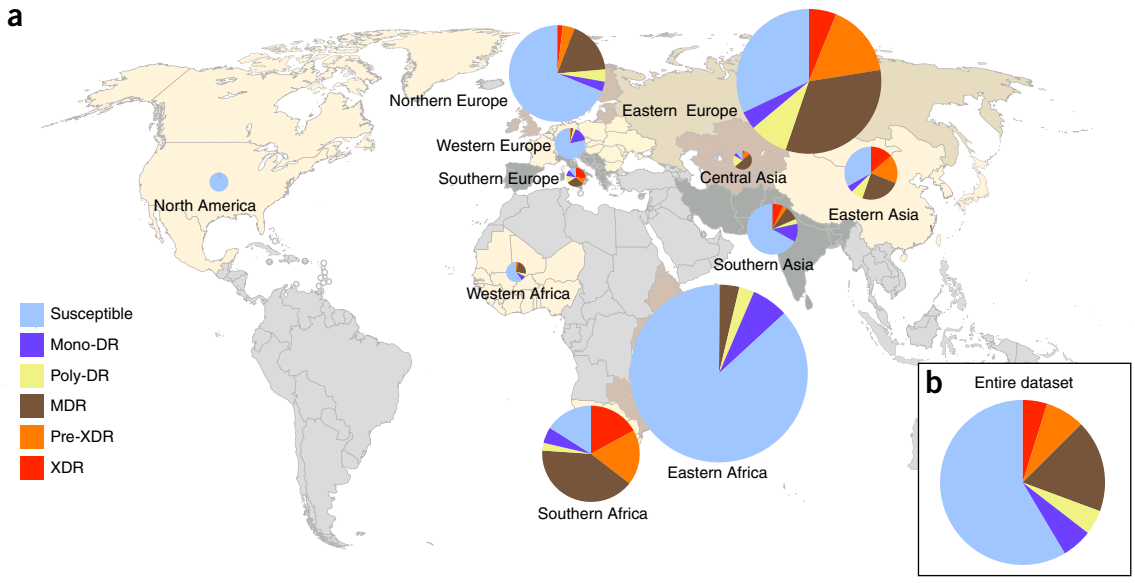

Figure 1 Geographical distribution of *M. tuberculosis* isolates by drug-resistance (DR) pattern. (**a**) Distribution of the 5,310 *M. tuberculosis* isolates included in our data set by DR genotype (pie charts) and by 11 UN geographical subregions (coloring); the plot is not meant to indicate the overall global incidence of TB or drug resistance. There were no strains in our data set from geographical regions that are shaded in gray. UN geographical subregions with fewer than 30 strains were excluded from this figure. The map was modified from a blank map of UN geographical subregions (https://commons.wikimedia.org/wiki/File:Geografiaj_subregionoj_la%C5%AD_Unui%C4%9Dintaj_Nacioj_malplene.svg; licensed under CC BY-SA 3.0 via Wikimedia Commons; http://commons.wikimedia.org/wiki/). (**b**) The overall proportion of drug-resistant strains identified among all 5,310 *M. tuberculosis* isolates in our data set.

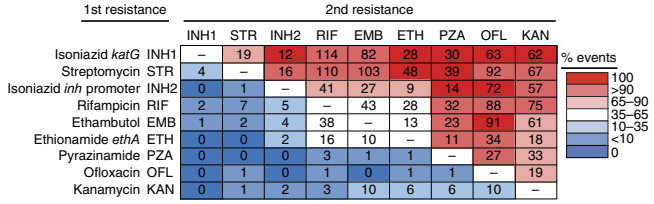

Figure 2 Across the globe, isoniazid resistance was overwhelmingly the first step toward drug resistance. Acquisition of a *katG* mutation affecting Ser315 preceded that of all other drug-resistance-conferring mutations for the majority of instances in which the order of acquisition could be disambiguated. We quantified the pairwise number of evolutions in which resistance to one drug preceded resistance to a second drug. Reported numbers represent the number of independent evolution events (not the number of strains) in which resistance to the drug indicated in the row labeled "first resistance" was acquired before resistance to the drug indicated in the column labeled "second resistance." Shading color indicates the percentage of evolutionary events in which resistance to the first drug clearly predates resistance to the second drug in that drug pair. Although mutations in *inhA* can confer resistance to both isoniazid and ethionamide[64], we defined genotypic ethionamide resistance as being conferred by mutations in only *ethA* to simplify the analysis and to avoid double counting.

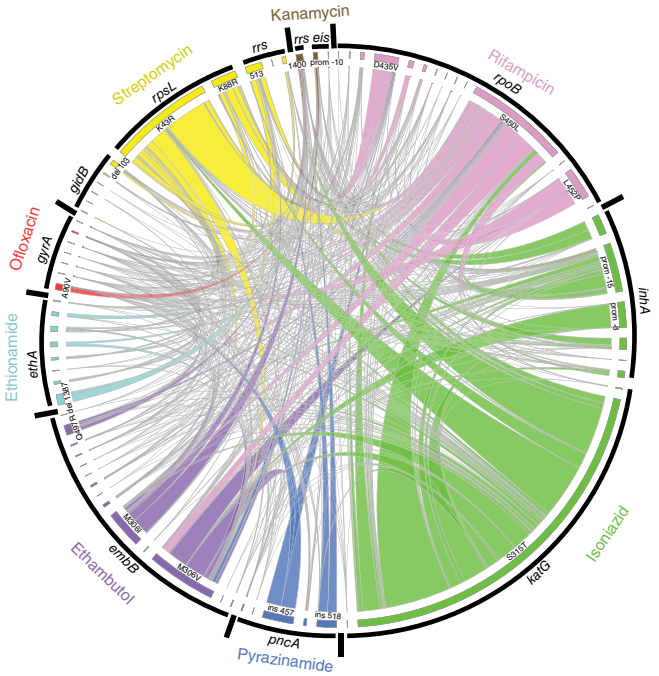

Figure 3 Sequential acquisition of drug-resistance-conferring mutations shows that isoniazid-resistance-conferring mutations, specifically *katG* mutation encoding p.Ser315Thr, most often come first in sequential pairs of mutations. This figure includes data from 71 drug-resistance-conferring mutations with at least ten occurrences in our data set, which represent 93% of all drug-resistance-conferring mutations in our data set. We used PAUP analysis to assign gains of specific mutations to individual nodes on the phylogeny and tabulated all routes of drug-resistance acquisition across the full strain phylogeny, examining only nodes on the tree where drug-resistance mutations arose (i.e., node 1 (mutation A) → node 2 (mutations B and C) → node 3 (mutation D)). We tabulated the number of times each pair of mutations arose sequentially at adjacent nodes (i.e., mutations A → B, A → C, B → D, and C → D). We removed node pairs that did not meet specific bootstrap and branch-length criteria (Online Methods). The ribbons in this figure depict the number of times that each pair of mutations arose sequentially at adjacent nodes across the entire data set. The width of the ribbon at each end is proportional to the number of times mutation A arose before mutation B, or vice versa (i.e., a ribbon with a thick end at *katG* p.Ser315Thr and a thin end at *rpoB* p.Ser450Leu indicates that *katG* p.Ser315Thr arose before *rpoB* p.Ser450Leu much more frequently than the reverse). Each ribbon is colored according to the mutation that more often occurred first in each sequential pairing.

rates of resistance; 962 strains (18%) had mutations for both rifampicin and isoniazid resistance and lacked mutations for ofloxacin and kanamycin resistance (MDR *sensu stricto*), and 257 (5%) of strains had mutations for resistance to all four drugs that define XDR (rifampicin, isoniazid, ofloxacin and kanamycin) (**Supplementary Table 6**). Another 409 (8%) strains carried mutations causing pre-XDR levels of resistance (MDR genotype plus mutations conferring resistance to either ofloxacin or kanamycin). Over half of the sequenced strains did not have any resistance-conferring mutation and were thus predicted to be drug susceptible (**Fig. 1**, **Supplementary Figs. 4–6** and **Supplementary Table 2**).

Drug resistance was identified in nearly all (15 of 17) of the UN regions (http://unstats.un.org/unsd/methods/m49/m49regin.htm) for which we had data, although its regional distribution varied considerably (**Supplementary Figs. 7–10** and **Supplementary Note**). In certain regions of the globe, we observed large numbers of closely related strains with nearly identical sets of resistance-conferring mutations, which could be attributed to clonal transmission. Because our data set contained isolates from several known outbreaks[6,13], rather than focusing on the total number of strains with each mutation, we instead examined the number of times each mutation evolved in different global regions by counting independent arisals, or the number of separate evolutions of a specific mutation occurring at defined positions in the phylogeny for a specific geography. Using parsimony-based analysis to reconstruct mutation gains and losses at all nodes across the phylogeny (Online Methods), we observed that the distribution of arisals of specific mutations was fairly constant across the globe, in contrast to the uneven distribution of strains with these mutations (**Supplementary Tables 7** and **8**, and **Supplementary Note**), suggesting that drug resistance has arisen via similar mechanisms irrespective of geography. This was also true for the evolution of MDR and XDR strains, which we calculated within our data set to have evolved independently 573 and 138 times, respectively. Along with frequent, repeated, *de novo* arisals, person-to-person transmission—as predicted when all strains descending from a common ancestor in the phylogeny shared the same MDR genotype—was also an important contributor to the observed MDR cases. Of the 573 arisals, 360 (63%) led to a single MDR strain in our data set (*de novo* evolution), whereas 213 arisals (37%) resulted in two or more descendant MDR strains, probably indicating person-to-person transmission of MDR-TB.

Isoniazid resistance overwhelmingly arises before rifampicin resistance across all lineages, geographical regions and time

In an earlier analysis involving a smaller data set of strains from South Africa[6], we showed that isoniazid resistance evolved before rifampicin resistance in almost all cases. To determine whether this ordering of mutation acquisition was also observed in a globally diverse set of strains, we used a parsimony-based analysis to examine the order of pairwise arisals of drug-resistance mutations. We filtered out portions of the phylogeny with ambiguous topologies (Online Methods) and only included nodes at which explicit ordering could be established[6]. In agreement with our previous results[6], we found that resistance to first-line drugs generally evolved before resistance to second-line drugs (**Supplementary Note**), as would be expected from the order

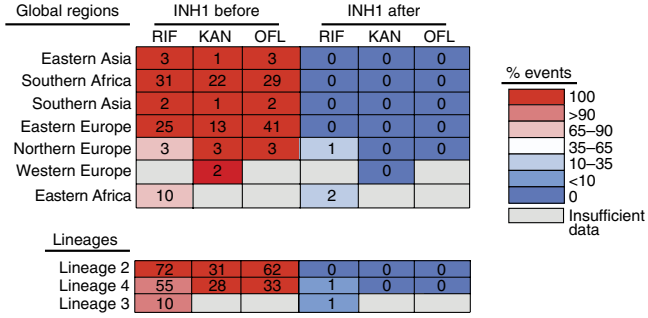

Figure 4 In all lineages and global regions, the *katG* mutation encoding p.Ser315Thr occurs first, and few examples of converse ordering are observed. We separately recalculated phylogenies for isolates from patients in each of the 11 UN subregions and five lineages with greater than 30 representatives. This figure depicts the pairwise ordering of the *katG* p.Ser315Thr mutation (INH1) in relation to mutations conferring resistance to the three other XDR-defining drugs (rifampicin (RIF), kanamycin (KAN) and ofloxacin (OFL)), within each individual *M. tuberculosis* lineage and geographical region. The numbers here do not necessarily add up to the same total number as that in **Figure 2**, as the analyses of regions and lineages were performed individually, which can affect the number of arisals. Gray shading indicates that there were not sufficient pairings for analysis. Data are not shown for the following regions and lineages, as there were insufficient pairings: West Africa, Southern Europe, Central Asia, Northern America, lineage 1 and *M. bovis*.

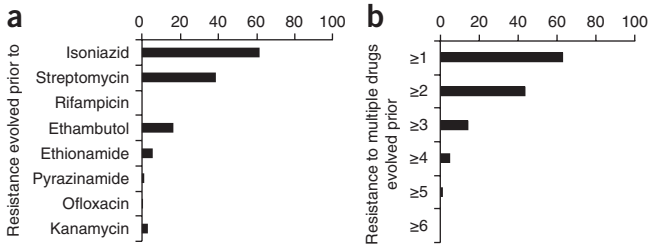

Figure 5 Non-rifampicin drug resistance often precedes the arisal of mutations that are detectable by the Xpert MTB/RIF assay. Data are shown here for nodes at which an Xpert MTB/RIF–detectable mutation arose. (**a**) Percentage of nodes at which an Xpert MTB/RIF–detectable mutation arose and for which resistance to each of eight drugs unambiguously preceded its arisal. Drug-resistance events that appeared to arise coincidentally with the Xpert MTB/RIF node were excluded from this representation. More than one additional drug-resistance event could precede a single Xpert MTB/RIF node. No strains contained additional rifampicin-resistance-conferring mutations that arose before those detectable by the Xpert MTB/RIF assay. The percentage of Xpert-assay-detectable mutations that are preceded by the presence of additional mutations that cause drug resistance is likely much higher, as we were unable to disambiguate ordering for a substantial number of nodes at which additional mutations arose at the same node (**Supplementary Fig. 11**). (**b**) Percentage of nodes at which resistance to one or more other drugs unambiguously preceded the arisal of mutations that were detectable by the Xpert MTB/RIF assay. 13% of Xpert MTB/RIF arisal nodes unambiguously had no additional drug-resistance-conferring mutations arising prior to the arisal of the Xpert MTB/RIF–detectable mutation.

in which antituberculosis drugs are used in clinical practice. We also observed that mutations conferring isoniazid resistance overwhelmingly arose before any other mutation implicated in resistance (**Fig. 2**, **Supplementary Table 9** and **Supplementary Note**), despite substantial complexity in the types and ordering of the mutations within our data set (**Fig. 3**). Notably, isoniazid resistance predated rifampicin resistance in 96% of pairwise comparisons (155 of 162), a pattern that remained true regardless of lineage or geographical source (**Fig. 4**). Although the majority of this effect was due to mutations in the *katG* gene encoding catalase–peroxidase (98%; 114 of 116 pairings), non-*katG*-associated mutations for isoniazid resistance followed this same pattern (89%; 41 of 46 pairings). Thus, the provenance of global MDR was overwhelmingly isoniazid-resistant strains. In particular, strains carrying a *katG* mutation encoding a p.Ser315Thr substitution frequently gained rifampicin resistance, whereas only a very small minority (4%) of global MDR arisals were due to a gain of isoniazid resistance on a rifampicin-resistant background, despite the presence of 48 rifampicin-monoresistant strains and of 152 (3% of total) isolates that were rifampicin resistant but not resistant to isoniazid.

One possible explanation for this notable result is that isoniazid entered into clinical use approximately 20 years before rifampicin (rifampicin was introduced between 1971 and 1993, depending on geography)[28], resulting in ancestral *M. tuberculosis* populations that had different amounts of exposure to these drugs, which could have affected the order of acquisition of drug resistance to favor resistance to isoniazid before rifampicin resistance. To test this hypothesis, we predicted the date for the arisal of each isoniazid- or rifampicin-resistance mutation using the BEAST software[29] and then tallied the number of co-arisals of resistance to both drugs that occurred during various time periods, starting from a fixed date in the past and extending to the present, (Online Methods), starting at 1971 (the date of introduction of rifampicin) and ending at 2000 (a later date included to account for the lag in timing for the widespread use of rifampicin) (**Supplementary Tables 9–11**). Our results showed that, regardless of the time period or evolutionary rate chosen (Online Methods and **Supplementary Note**), resistance mediated by *katG* mutation affecting Ser315 arose before rifampicin resistance 92–98% of the time (**Supplementary Table 10**), indicating that, even during the era when isoniazid and rifampicin were given in combination, the emergence of isoniazid resistance predated that of rifampicin resistance.

Diagnostics for early detection of pre-MDR *M. tuberculosis*

Contributors to the current global burden of MDR-TB include not only historical emergences of MDR strains, which led to person-to-person transmission of MDR-TB, but also ongoing *de novo* evolution. Of the 573 MDR arisals in our data set, we estimated that 67% occurred since 2004. Thus, new strategies for curbing the emergence of MDR strains, such as identifying strains that are precursors to MDR strains, will be critical to the control of MDR strains worldwide. Xpert MTB/RIF, currently one of the frontline diagnostic tests used to exclusively identify rifampicin-resistance-conferring mutations in the rifampicin-resistance-determining region (RRDR) of *rpoB*[5], is commonly used globally as a proxy for detecting MDR-TB. The most common ordering observed, of isoniazid resistance before rifampicin resistance, indicates that Xpert MTB/RIF serves as an appropriate proxy for MDR-TB and is well-suited to detect MDR strains in all geographical regions and all lineages of *M. tuberculosis* (**Fig. 4**). However, because mutations that result in rifampicin resistance (detectable by Xpert MTB/RIF) are rarely the first drug-resistance-conferring mutations to emerge, oftentimes by the time a mutation that is detectable by the Xpert MTB/RIF assay develops, there is pre-existing resistance to multiple additional drugs, including second-line drugs (**Fig. 5**). Because we excluded nodes at which we were unable to disambiguate the relative ordering of the acquisition of pairs of drug-resistance-associated mutations, our estimates represent a lower bound on the number

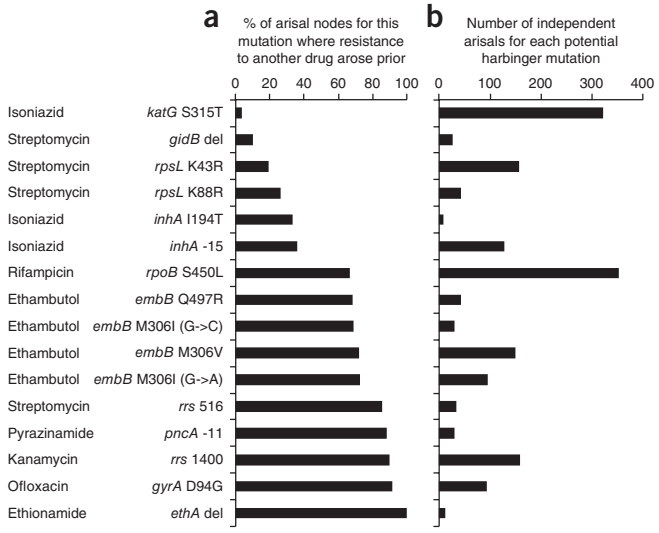

Figure 6 *katG* p.Ser315Thr is a commonly occurring mutation with very little resistance to other drugs arising prior to its occurrence. (**a**) The percentage of nodes at which resistance to another drug unambiguously preceded the arisal of the indicated mutation, for each of the 16 pre-MDR mutations. The percentage of nodes at which another drug-resistance-conferring mutation had already arisen prior to the pre-MDR mutation is likely much higher, as we were unable to disambiguate ordering for a substantial number of nodes for which additional mutations arose at the same node. (**b**) The number of independent arisals for each of the 16 pre-MDR (or harbinger) mutations. Because there are two mutations at the Met306 codon in *embB*, the nucleotide change at position 4,247,609 is also indicated for these two variants.

of nodes at which resistance to other drugs was gained before a rifampicin-resistance-conferring mutation could be detected by the Xpert MTB/RIF (**Supplementary Fig. 11**).

Diagnostics that identify mutations present before the emergence of multidrug resistance would provide an opportunity to identify drug resistance during a period in which there are both greater therapeutic options and improved treatment outcomes[30]. Although the results from our pairwise ordering of resistance acquisition clearly demonstrated that *katG* p.Ser315Thr–mediated isoniazid resistance is among the earliest to evolve, our pairwise approach necessarily oversimplified the complex process of MDR-TB evolution. Thus, in a complementary approach to identify other possible sentinels of complex resistance, we cataloged all of the resistance-conferring mutations that commonly evolved before the development of MDR-TB (which we refer to as 'pre-MDR-TB mutations'). For this set of mutations, we quantified the fraction of MDR-defining nodes at which one of these pre-MDR-TB mutations had evolved before the development of MDR-TB to determine how much resistance to other drugs had unambiguously arisen before the emergence of multidrug resistance (**Fig. 6**). Our analysis identified a set of 16 resistance-conferring mutations (of 340 total found among MDR and XDR strains) that arose before MDR-TB a minimum of two independent times (**Supplementary Table 12** and **Supplementary Note**).

Unexpectedly, we observed resistance-conferring mutations for all eight drugs among this set of pre-MDR-TB mutations. However, many of these mutations evolved infrequently (**Fig. 6b**), and thus would likely have low negative predictive value if their sequences were included on a diagnostic panel aimed at identifying pre-MDR-TB. In contrast, the *katG* mutation encoding p.Ser315Thr, which confers isoniazid resistance, stood out as a frequently occurring mutation (**Fig. 6b**) with very few instances of resistance to other drugs arising before its gain (**Fig. 6a**). Despite the high level of complexity in the stepwise acquisition of drug-resistance-conferring mutations in *M. tuberculosis* (**Fig. 3**), the *katG* mutation encoding p.Ser315Thr was by far the most common mutation to evolve before the emergence of multidrug resistance (**Fig. 2**, **Supplementary Tables 12** and **13**, and **Supplementary Note**). Of the 321 independent arisals of *katG* p.Ser315Thr in our data set, 302 (94%) occurred at the earliest node in which drug resistance was present.

DISCUSSION

We constructed the largest global data set of *M. tuberculosis* whole genomes analyzed to date, consisting of genomes from 5,310 diverse strains. Although the global distribution of strains in this data set does not reflect the global incidence of TB for some regions (**Supplementary Fig. 1b**), our unique data set had a broad geographical distribution and deep sampling of drug-resistant strains, including MDR and XDR strains from multiple lineages and regions. We were, therefore, able to dissect the step-by-step evolution of drug resistance and to identify harbinger resistance-conferring mutations that emerged before development of MDR-TB.

We observed that MDR-TB and XDR-TB evolved many independent times, in different lineages and regions of the world, suggesting that there are many 'permissive' environments that have allowed MDR-TB and XDR-TB to emerge repeatedly. Molecular diagnostic tests for drug-resistant TB could be improved by incorporating knowledge of the global patterns of resistance emergence. We observed that the distribution of arisals of specific resistance-conferring mutations was fairly constant across the globe, indicating that drug resistance has arisen via common mechanisms worldwide. Thus, a universal diagnostic for detecting resistance to the eight drugs examined here may be achievable without the need for regional specialization. Without phenotypic drug-susceptibility data for all included strains, we were not able to identify previously unknown drug-resistance-conferring mutations or to quantify the amount of drug resistance that remains unexplained by our curated list of polymorphisms; however, we expect this amount to be small[27].

By dissecting the step-by-step evolution of drug-resistant *M. tuberculosis* across the phylogeny, we observed that patterns in the order of emergence of drug resistance also appeared to be conserved globally. In particular, across all lineages and geographical regions, isoniazid resistance overwhelmingly arose before rifampicin resistance. Some regions of the world, such as Iran[31], are reported to have a high incidence of rifampicin resistance; however, our results suggest that rifampicin monoresistance rarely leads to MDR-TB. Although the effects of convergent evolution among frequently evolving mutations could cause isoniazid resistance evolutions to be dated further back in time than when they actually occurred, we took care to minimize such effects (**Supplementary Note**). In support of our results, this relative ordering of isoniazid and rifampicin resistance is consistent with prior findings based on genomic data showing that isoniazid resistance arises before rifampicin resistance in Russia[7], South Africa[6] and South America[7], as well as with analysis of a large global collection of phenotypic data[8].

Why would isoniazid resistance arise first? We showed that the earlier clinical introduction of isoniazid was not a major contributor to the earlier arisal of isoniazid resistance; our dating analysis indicates that isoniazid resistance arose before rifampicin resistance across all time periods, including recently (**Supplementary Note**). However, there are many alternative, although not definitive, explanations for this preferential ordering. Isoniazid is a prodrug, which must first be

activated by KatG (encoded by *katG*), the catalase–peroxidase[32], to form an adduct with nicotinamide adenine dinucleotide (NAD)[33], which then inhibits InhA (encoded by *inhA*), an NADH-dependent enoyl-acyl carrier protein reductase[34], and ultimately inhibits mycolic acid biosynthesis[35]. The major mechanisms of isoniazid resistance include mutations in *katG*, a non-essential gene, which result in failure to activate isoniazid, and either upregulation or target modification of InhA. Rifampicin inhibits the β subunit of the mycobacterial RNA polymerase, encoded by a single, essential gene, *rpoB*[36]. *M. tuberculosis* cells grown *in vitro* have higher spontaneous mutation rates toward isoniazid resistance than rifampicin resistance[37,38], which could be due to the greater number of mutations that can lead to isoniazid resistance as compared to rifampicin resistance, i.e., any inactivating mutation within *katG* can result in isoniazid resistance, whereas only specific non-inactivating mutations in *rpoB* can result in rifampicin resistance. However, we observed that a single mutation in *katG*, which results in a substitution of serine to threonine at position 315, accounted for the majority of isoniazid-resistance arisals and that, overall, there were ~20% more independent arisals of resistance to rifampicin than there were to isoniazid, indicating that the relative rates of resistance *in vivo* may differ from those calculated *in vitro*.

Another possible explanation for the ordering is that isoniazid-resistant strains, including those carrying the prevalent *katG* mutation encoding p.Ser315Thr, are more likely to develop resistance to other drugs. Although previous *in vitro* studies have shown a difference in the types of rifampicin-resistance-conferring mutations that arise on isoniazid-resistant backgrounds[39], there is no evidence that isoniazid-resistant strains are transformed into 'hypermutators' (ref. 40). Furthermore, the sequence surrounding the Ser315 codon in *katG* does not seem to be susceptible to mutation nor does it seem to be a mutational hotspot *in vitro*[37]. However, as we and others have shown, this specific mutation is common among isoniazid-resistant clinical isolates, indicating that it is well tolerated *in vivo*. This is probably due to the fact that this mutation preserves mycobacterial catalase activity while still preventing activation of isoniazid[41]. This preserved fitness may affect the evolutionary adaptive landscape[42,43] through which *M. tuberculosis* may acquire future resistance. Such a fitness landscape, which takes into account the relative fitness of different combinations of resistance-conferring and compensatory mutations, may produce a restricted set of evolutionary paths leading to MDR-TB.

A third possibility is that there is differential drug availability within the body, either due to pharmacokinetic effects[44,45] or to differential rates of clinical penetration of the drugs into lesions[46], that may influence the order of emergence of mutations. Current treatment regimens that result in suboptimal dosing of rifampicin[47–49] may result in effective mono-exposure to isoniazid, increasing the likelihood of developing isoniazid resistance first. Isoniazid preventative therapy (IPT)[50], the treatment of suspected cases of latent TB with only isoniazid, could provide an opportunity for isoniazid resistance to develop before exposure to other drugs. However, IPT is not commonly used in most of the countries for which we have assembled data and, therefore, is unlikely to have a major role in the early arisal of isoniazid resistance in our data set.

Early identification and appropriate treatment of individuals with isoniazid-monoresistant strains, including treatment with non-isoniazid-based regimens[51], may prevent the selection and eventual transmission of strains with resistance to additional drugs. The worldwide case rate of isoniazid monoresistance is estimated to be as high as 2–15% (refs. 52–56), or 200,000–1,400,000 cases per year. Several studies have shown that patients harboring isoniazid-monoresistant strains have worse clinical outcomes than those harboring susceptible strains[57–59], and enhanced treatment regimens for patients with such strains resulted in lower rates of treatment failure and acquired drug resistance[59]. A recent large meta-analysis revealed substantially worse treatment outcomes—including higher rates of MDR acquisition—when patients with isoniazid-resistant strains were treated with treatment regimens recommended by the World Health Organization that contained only first-line TB drugs[9]. One large retrospective cohort study also points to early detection of isoniazid monoresistance for improved outcomes[52]. In particular, the *katG* mutation encoding p.Ser315Thr has been associated with unfavorable treatment outcome and increased relapse in one population[60]. These and other results challenge the predictions of an earlier mathematical modeling study, which prognosticated that incorporation of isoniazid resistance onto a molecular test in India would provide only a negligible benefit to the control of MDR-TB[61].

Our large data set confirms that Xpert MTB/RIF, currently the most widely used rapid molecular diagnostic for the diagnosis of *M. tuberculosis* and MDR-TB, performs excellently as a surrogate marker for multidrug resistance, irrespective of strain lineage and/or regional source. However, as it detects only rifampicin-resistance-conferring mutations, Xpert MTB/RIF does not identify drug resistance at the earliest available opportunity. Thus, diagnostic algorithms that rely upon application of Xpert MTB/RIF alone will allow rifampicin-susceptible but otherwise drug-resistant strains to propagate unchecked. In fact, our evolutionary analysis indicates that, by the time a Xpert MTB/RIF–identifiable rifampicin-resistance-conferring mutation is acquired, oftentimes multiple additional resistance-conferring mutations are already present. Additional commercially available diagnostics, such as the Hain MTBDRplus and Hain MTBDRsl ver. 2.0 line-probe assays, are available and detect a broader set of drug-resistance-conferring mutations. Despite logistical considerations that make practical application of this technology more difficult—such as moderate turnaround times and the need for specialized laboratory facilities—line-probe assays are able to detect isoniazid resistance (Hain MTBDRplus) with excellent specificity[62]. However, current diagnostic algorithms in certain TB-endemic countries[63] call for the application of these tests only after rifampicin resistance has been identified by Xpert MTB/RIF. Therefore, more comprehensive diagnostic tests are not being appropriately used to their full potential to identify rifampicin-susceptible but otherwise drug-resistant strains.

Through an evolutionary analysis of a diverse, global data set encompassing the genomes of 5,310 strains of *M. tuberculosis*, we observed that recent *de novo* emergence of MDR-TB in the past 10 years is a substantial contributor to global MDR today. Thus, to stem the development of additional MDR strains, one should seek to identify drug-resistant strains in the pre-MDR stage, during which there are additional therapeutic options and improved treatment outcomes. The identification of harbinger mutations, such as the *katG* mutation encoding p.Ser315Thr, may serve as an early warning signal that multidrug resistance may soon develop. Focusing on common, early-occurring mutations could improve the design of diagnostic tests that are aimed at targeting the earliest-occurring signatures of drug-resistant bacteria. Future prospective research will be needed to determine whether these harbinger mutations increase the risk of MDR *M. tuberculosis* emergence in a given population. If substantiated, then surveillance efforts for harbinger mutations may assist organizations to better allocate intensified TB control resources to at-risk areas and to target drug resistance in the pre-MDR stage.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank the Broad Institute's Genome Sequencing Platform and Assembly and Annotation teams, including S.K. Young, M.E. Priest, T.P. Shea, B.J. Walker, L. Alvarado, M.G. Fitzgerald, S. Gujja, S. Hamilton, C. Howarth, J.D. Larimer, M.D. Pearson, Q. Zeng and J. Wortman. We would like to thank J. Romano and A. Keo for help with lineage detection, and M. Zambrano, B. Ferro and J.C. Rozo for isolation and phenotypic characterization of strains. We are also grateful to members of the TBResist Consortium for contribution of their strains, phenotypic data and expertise, and their help in forging collaborations, and to V. Dartois, D. Thomas, D. Hung and D. Plachetzki for helpful conversations. We also thank three anonymous reviewers of our manuscript for their insights and helpful suggestions. This project has been funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services (grant U19AI110818 to the Broad Institute (B.W.B. and A.M.E.)), contract HHSN272200900018C to the Broad Institute (B.W.B.) and contract HHSN2722000900050C to the TB Clinical Diagnostics Research Consortium, the Intramural Research Program of NIAID (C.E.B. and L.E.V.) and the Korean CDC, Korean Ministry of Health and Welfare. This work was also funded (in part) by the intramural research program of the NIAID, NIH (C.E.B.). Funding was also provided by NIH grant 5U01AI069924-07 for IeDEA (A.S.P.), the Howard Hughes Medical Institute (W.R.B.) and NIH grant R01 AI110386 for 'Host–pathogen interactions in a failing global lineage of MTBC: *M. africanum*' (W.R.B.). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

A.L.M., K.A.C., T.A., C.A.D., B.W.B. and A.M.E. conceived the project; A.L.M., K.A.C., T.A., C.A.D. and A. Salazar analyzed the data; A.L.M., K.A.C. and A.M.E. interpreted results; A.L.M. and K.A.C. wrote the manuscript; and D.T.A., C.E.B., J.B., S.B.C., S.-N.C., A.G., J.G., A.M.J., M.J., P.J., J.S.L., L.M., M.M., D.N., E.N., E.R., A. Skrahina, W.S., A.A.V., K.W., A.Z., L.E.V., G.H.C., S.E.D., J.E., P.F., J.E.G., A.R., V.C., D.H., P.-R.H., S.N., A.S.P., S.S., M.V.d.W., D.A., W.R.B., T.C. and S.H. were involved in sample acquisition and handling, including oversight of these activities. All authors critically read and revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Farmer, P. *et al.* The dilemma of MDR-TB in the global era. *Int. J. Tuberc. Lung Dis.* **2**, 869–876 (1998).
2. World Health Organization. Global Tuberculosis Report 2015 (World Health Organization, Geneva, Switzerland, 2015).
3. Awofeso, N. Antituberculosis medication side effects constitute major factor for poor adherence to tuberculosis treatment. *Bull. World Health Organ.* **86**, B–D (2008).
4. Pietersen, E. *et al.* Long-term outcomes of patients with extensively drug-resistant tuberculosis in South Africa: a cohort study. *Lancet* **383**, 1230–1239 (2014).
5. Boehme, C.C. *et al.* Rapid molecular detection of tuberculosis and rifampin resistance. *N. Engl. J. Med.* **363**, 1005–1015 (2010).
6. Cohen, K. *et al.* Evolution of extensively drug-resistant tuberculosis over four decades revealed by whole-genome sequencing of *Mycobacterium tuberculosis* from KwaZulu-Natal, South Africa. *PLoS Med.* **12**, e1001880 (2015).
7. Eldholm, V. *et al.* Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat. Commun.* **6**, 7119 (2015).
8. Izu, A., Cohen, T. & Degruttola, V. Bayesian estimation of mixture models with prespecified elements to compare drug resistance in treatment-naive and experienced tuberculosis cases. *PLoS Comput. Biol.* **9**, e1002973 (2013).
9. Gegia, M., Winters, N., Benedetti, A., van Soolingen, D. & Menzies, D. Treatment of isoniazid-resistant tuberculosis with first-line drugs: a systematic review and meta-analysis. *Lancet Infect. Dis.* http://dx.doi.org/10.1016/S1473-3099(16)30407-8 (2016).
10. Biek, R. *et al.* Whole-genome sequencing reveals local transmission patterns of *Mycobacterium bovis* in sympatric cattle and badger populations. *PLoS Pathog.* **8**, e1003008 (2012).
11. Blouin, Y. *et al.* Significance of the identification in the Horn of Africa of an exceptionally deep branching *Mycobacterium tuberculosis* clade. *PLoS One* **7**, e52841 (2012).
12. Bryant, J.M. *et al.* Inferring patient-to-patient transmission of *Mycobacterium tuberculosis* from whole-genome sequencing data. *BMC Infect. Dis.* **13**, 110 (2013).
13. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.* **46**, 279–286 (2014).
14. Clark, T.G. *et al.* Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment-experienced patients by whole-genome sequencing. *PLoS One* **8**, e83012 (2013).
15. Comas, I. *et al.* Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* **45**, 1176–1182 (2013).
16. Gardy, J.L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
17. Guerra-Assunção, J.A. *et al.* Recurrence due to relapse or re-infection with *Mycobacterium tuberculosis*: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *J. Infect. Dis.* **211**, 1154–1163 (2015).
18. Merker, M. *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat. Genet.* **47**, 242–249 (2015).
19. Perdigão, J. *et al.* Unraveling *Mycobacterium tuberculosis* genomic diversity and evolution in Lisbon, Portugal, a highly drug-resistant setting. *BMC Genomics* **15**, 991 (2014).
20. Walker, T.M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
21. Zhang, H. *et al.* Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.* **45**, 1255–1260 (2013).
22. Winglee, K. *et al.* Whole-genome sequencing of *Mycobacterium africanum* strains from Mali provides insights into the mechanisms of geographic restriction. *PLoS Negl. Trop. Dis.* **10**, e0004332 (2016).
23. Wollenberg, K.R. *et al.* Whole-genome sequencing of *Mycobacterium tuberculosis* provides insight into the evolution and genetic composition of drug-resistant tuberculosis in Belarus. *J. Clin. Microbiol.* http://dx.doi.org/10.1128/JCM.02116-16 (2016).
24. Gagneux, S. *et al.* Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **103**, 2869–2873 (2006).
25. Yimer, S.A. *et al. Mycobacterium tuberculosis* lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara region, Ethiopia. *J. Clin. Microbiol.* **53**, 1301–1309 (2015).
26. O'Reilly, L.M. & Daborn, C.J. The epidemiology of *Mycobacterium bovis* infections in animals and man: a review. *Tuber. Lung Dis.* **76** (Suppl. 1), 1–46 (1995).
27. Desjardins, C.A. *et al.* Novel d-cycloserine resistance mechanism in *Mycobacterium tuberculosis* revealed by whole-genome analysis. *Nat. Genet.* **48**, 544–551 (2016).
28. Ma, Z., Lienhardt, C., McIlleron, H., Nunn, A.J. & Wang, X. Global tuberculosis drug development pipeline: the need and the reality. *Lancet* **375**, 2100–2109 (2010).
29. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
30. Cattamanchi, A. *et al.* Clinical characteristics and treatment outcomes of patients with isoniazid-monoresistant tuberculosis. *Clin. Infect. Dis.* **48**, 179–185 (2009).
31. Velayati, A.A. *et al.* High prevalance of rifampin-monoresistant tuberculosis: a retrospective analysis among Iranian pulmonary tuberculosis patients. *Am. J. Trop. Med. Hyg.* **90**, 99–105 (2014).
32. Johnsson, K. & Schultz, P.G. Mechanistic studies of the oxidation of isoniazid by the catalase peroxidase from *Mycobacterium tuberculosis*. *J. Am. Chem. Soc.* **116**, 7425–7426 (1994).
33. Rozwarski, D.A., Grant, G.A., Barton, D.H., Jacobs, W.R. Jr. & Sacchettini, J.C. Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*. *Science* **279**, 98–102 (1998).
34. Quémard, A. *et al.* Enzymatic characterization of the target for isoniazid in *Mycobacterium tuberculosis*. *Biochemistry* **34**, 8235–8241 (1995).
35. Marrakchi, H., Lanéelle, G. & Quémard, A. InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. *Microbiology* **146**, 289–296 (2000).
36. Sassetti, C.M., Boyd, D.H. & Rubin, E.J. Genes required for mycobacterial growth defined by high-density mutagenesis. *Mol. Microbiol.* **48**, 77–84 (2003).
37. Bergval, I.L., Schuitema, A.R., Klatser, P.R. & Anthony, R.M. Resistant mutants of *Mycobacterium tuberculosis* selected *in vitro* do not reflect the *in vivo* mechanism of isoniazid resistance. *J. Antimicrob. Chemother.* **64**, 515–523 (2009).
38. Ford, C.B. *et al. Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* **45**, 784–790 (2013).
39. Bergval, I. *et al.* Pre-existing isoniazid resistance, but not the genotype of *Mycobacterium tuberculosis*, drives rifampicin resistance codon preference *in vitro*. *PLoS One* **7**, e29108 (2012).
40. O'Sullivan, D.M., McHugh, T.D. & Gillespie, S.H. The effect of oxidative stress on the mutation rate of *Mycobacterium tuberculosis* with impaired catalase–peroxidase function. *J. Antimicrob. Chemother.* **62**, 709–712 (2008).

41. Pym, A.S., Saint-Joanis, B. & Cole, S.T. Effect of *katG* mutations on the virulence of *Mycobacterium tuberculosis* and the implication for transmission in humans. *Infect. Immun.* **70**, 4955–4960 (2002).
42. Salverda, M.L. *et al.* Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* **7**, e1001321 (2011).
43. Ogbunugafor, C.B., Wylie, C.S., Diakite, I., Weinreich, D.M. & Hartl, D.L. Adaptive landscape by environment interactions dictate evolutionary dynamics in models of drug resistance. *PLoS Comput. Biol.* **12**, e1004710 (2016).
44. Gumbo, T. *et al.* Concentration-dependent *Mycobacterium tuberculosis* killing and prevention of resistance by rifampin. *Antimicrob. Agents Chemother.* **51**, 3781–3788 (2007).
45. Pasipanodya, J.G. & Gumbo, T. A new evolutionary and pharmacokinetic–pharmacodynamic scenario for rapid emergence of resistance to single and multiple antituberculosis drugs. *Curr. Opin. Pharmacol.* **11**, 457–463 (2011).
46. Jutte, P.C., Rutgers, S.R., Van Altena, R., Uges, D.R. & Van Horn, J.R. Penetration of isoniazid, rifampicin and pyrazinamide in tuberculous pleural effusion and psoas abscess. *Int. J. Tuberc. Lung Dis.* **8**, 1368–1372 (2004).
47. Diacon, A.H. *et al.* Early bactericidal activity of high-dose rifampin in patients with pulmonary tuberculosis evidenced by positive sputum smears. *Antimicrob. Agents Chemother.* **51**, 2994–2996 (2007).
48. Long, M.W., Snider, D.E. Jr. & Farer, L.S. U.S. Public Health Service Cooperative trial of three rifampin–isoniazid regimens in treatment of pulmonary tuberculosis. *Am. Rev. Respir. Dis.* **119**, 879–894 (1979).
49. Mitchison, D.A. Role of individual drugs in the chemotherapy of tuberculosis. *Int. J. Tuberc. Lung Dis.* **4**, 796–806 (2000).
50. Mills, H.L., Cohen, T. & Colijn, C. Community-wide isoniazid preventive therapy drives drug-resistant tuberculosis: a model-based analysis. *Sci. Transl. Med.* **5**, 180ra49 (2013).
51. Jindani, A. *et al.* High-dose rifapentine with moxifloxacin for pulmonary tuberculosis. *N. Engl. J. Med.* **371**, 1599–1608 (2014).
52. Chien, J.Y. *et al.* Treatment outcome of patients with isoniazid-monoresistant tuberculosis. *Clin. Microbiol. Infect.* **21**, 59–68 (2015).
53. Fasih, N., Rafiq, Y., Jabeen, K. & Hasan, R. High isoniazid resistance rates in rifampicin-susceptible *Mycobacterium tuberculosis* pulmonary isolates from Pakistan. *PLoS One* **7**, e50551 (2012).
54. Jenkins, H.E., Zignol, M. & Cohen, T. Quantifying the burden and trends of isoniazid-resistant tuberculosis, 1994–2009. *PLoS One* **6**, e22927 (2011).
55. Lumb, R. *et al.* Tuberculosis in Australia: bacteriologically confirmed cases and drug resistance, 2010. A report of the Australian Mycobacterium Reference Laboratory Network. *Commun. Dis. Intell. Q. Rep.* **37**, E40–E46 (2013).
56. Nagu, T.J. *et al.* Multidrug- and other forms of drug-resistant tuberculosis are uncommon among treatment-naive tuberculosis patients in Tanzania. *PLoS One* **10**, e0118601 (2015).
57. Villegas, L. *et al.* Prevalence, risk factors and treatment outcomes of isoniazid- and rifampicin-monoresistant pulmonary tuberculosis in Lima, Peru. *PLoS One* **11**, e0152933 (2016).
58. Espinal, M.A. *et al.* Standard short-course chemotherapy for drug-resistant tuberculosis: treatment outcomes in six countries. *J. Am. Med. Assoc.* **283**, 2537–2545 (2000).
59. Menzies, D. *et al.* Standardized treatment of active tuberculosis in patients with previous treatment and/or with monoresistance to isoniazid: a systematic review and meta-analysis. *PLoS Med.* **6**, e1000150 (2009).
60. Huyen, M.N. *et al.* Epidemiology of isoniazid-resistance mutations and their effect on tuberculosis treatment outcomes. *Antimicrob. Agents Chemother.* **57**, 3620–3627 (2013).
61. Denkinger, C.M., Pai, M. & Dowdy, D.W. Do we need to detect isoniazid resistance in addition to rifampicin resistance in diagnostic tests for tuberculosis? *PLoS One* **9**, e84197 (2014).
62. Bai, Y., Wang, Y., Shao, C., Hao, Y. & Jin, Y. GenoType MTBDRplus assay for rapid detection of multidrug resistance in *Mycobacterium tuberculosis*: a meta-analysis. *PLoS One* **11**, e0150321 (2016).
63. Department of Health, South Africa. *Management of Drug-Resistant Tuberculosis: Policy Guidelines* http://www.health-e.org.za/wp-content/uploads/2014/06/MDR-TB-Clinical-Guidelines-Updated-Jan-2013.pdf (2013).
64. Banerjee, A. *et al. inhA*, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*. *Science* **263**, 227–230 (1994).

TBResist Global Genome Consortium: Jeannette Brand[7], Pontus Jureen[13], Lesibana Malinga[7], Dale Nordenberg[15], A A Velayati[17], Gail H Cassell[19], Parissa Farnia[17], Daniela Homorodean[11], Alena Skrahina[18], Martie Van der Walt[7] & Sven Hoffner[13]

ONLINE METHODS

Whole-genome sequencing and data sets. References detailing the sequencing methods for the published data sets can be found in **Supplementary Table 1**. Sequencing data for the TB Antibiotic Resistance Catalog (TB-ARC) projects (**Supplementary Table 1**) were generated at the Broad Institute, as in Cohen *et al.*[6]. Additional information about samples for each of these unpublished projects can be found at the Broad Institute website (https://olive.broadinstitute.org/projects/tb_arc). The goal of the TB-ARC project was to create a catalog of mutations for antibiotic resistance in *M. tuberculosis* to inform diagnostics. As such, strains from each of the countries represented both drug-sensitive and drug-resistant isolates that would enable curation of such a catalog. For the data set from India, 223 drug-resistant and drug-sensitive strains, representative of lineages found in India (particularly lineages 1 and 3) were selected from studies in Tiruvallur and Madurai in Southern India and sequenced. For the MRC data set, 189 primarily drug-resistant strains from South Africa and Swaziland were selected for sequencing. For the CDRC data set, 179 genomes from South Korea and Uganda, with a wide variety of drug-resistance patterns, as well as extensive characterization of drug-susceptibility profiles, were sequenced. For the data set from Sweden, 150 genomes were collected, primarily from Sweden's immigrant population. This set included a complete collection of all 141 MDR and XDR strains identified nationwide in Sweden in the period from 2003 to 2013. For the data set from Moldova, 95 genomes were selected from the countrywide specimen and data repositories. For the data set from Romania, 34 genomes were sequenced, with the goal of describing drug-resistant strains that were circulating in Romania and their diversity. For the data set from Iran, 33 primarily highly drug-resistant samples, including totally drug-resistant (TDR) samples, were sequenced.

The study protocol for these TB-ARC projects was approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects. Informed consent was obtained from all subjects, or else an appropriate waiver of consent was obtained.

For all of these TB-ARC projects, genomic DNA was extracted using published methods[65]. Library preparation and whole-genome sequencing (WGS) were performed as previously described on the Illumina HiSeq 2000 at the Broad Institute[66]. Sequencing data were submitted to the Sequence Read Archive (SRA) at NCBI under the umbrella BioProject identifiers listed in **Supplementary Table 1**.

Alignments. Raw read data for 8,136 strains were downloaded from the SRA (see SRA accession codes in **Supplementary Table 1**). Reads were mapped onto the genomic sequence of the *M. tuberculosis* reference strain H37Rv (GenBank accession number CP003248.2) using BWA version 0.7.10 (ref. 67). Variants were identified using Pilon version 1.11 as described[66]. The global *M. tuberculosis* lineage designations used in our analysis, as well as each strain's spoligotype, were predicted using digital spoligotyping, as described in Cohen *et al.*[6].

We eliminated 824 strains that did not pass our quality control filters: average sequencing depth of coverage >20×; fraction of long insertions <0.2; ambiguity rate <0.5 (to remove samples that were suspected to represent mixes of different genotypes); number of low-coverage bases <250,000; and having a single match to one lineage in our lineage-prediction algorithm. We also eliminated strains for which Pilon analysis failed three times. Of the remaining 7,312 samples, we removed 1,970 strains with no 'country' metadata or description in a publication, 19 strains with substantial non-tuberculous mycobacterial contamination, as well as 13 additional duplicate patient samples. Use of these filters resulted in a final set of 5,310 strains for analysis.

The Emu software[68] was run to canonicalize variants. We conducted phylogenetic analyses for the entire set of 5,310 strains, as well as for a subset corresponding to each lineage and each United Nations geographical subregion with >30 strains (**Supplementary Table 3**). For each set, all sites with unambiguous single-nucleotide polymorphisms (SNPs) in at least one strain were combined into a concatenated alignment. Ambiguous positions were treated as missing data. The concatenated alignment was then were used to generate a midpoint-rooted phylogenetic tree using FastTree[69] version 2.1.8.

Drug-resistance mutations. A curated list of genomic polymorphisms that confer drug resistance was defined for eight drugs: rifampicin, ethambutol, isoniazid, ethionamide, ofloxacin, pyrazinamide, streptomycin and kanamycin. This was based on a literature review and consideration of current molecular drug-resistance diagnostics. All mutations incorporated in current molecular diagnostics in standard practice were included. This included the Xpert MTB/RIF[5] (http://www.cepheid.com/us/cepheid-solutions/clinical-ivd-tests/critical-infectious-diseases/xpert-mtb-rif), the Hain Genotype MTBDRplus (http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/genotype-mtbdrplus.html), the Hain Genotype MTBDRsl Line Probe Assay (http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/genotype-mtbdrsl.html), as well as the Hain MTBDRsl Line Probe Assay v. 2.0 (http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/genotype-mtbdrsl.html). Additional resistance-conferring mutations were selected for inclusion based on laboratory evidence and recent compelling genomic evidence that these mutations encode for resistance (**Supplementary Tables 4** and **5**). Because of greater uncertainty in calling longer variants in our data, we excluded insertions and deletions that were longer than 10 bp. Using this curated list, we identified 392 drug-resistance-conferring mutations among the 231,898 total variants observed in our full data set across 5,310 strains.

Recent reports suggest that currently tabulated mutation sets account for the majority of phenotypic resistance[27,70,71]. Based on tabulated mutation sets, pyrazinamide phenotypic resistance was predicted with lower sensitivity, based on genotype, than other drugs in one recent analysis[70]. However, including all loss-of-function mutations in *pncA* in the mutation set, as we have done in this study, would likely greatly improve sensitivity in predicting phenotypic pyrazinamide resistance based on genotype.

Evolution of drug-resistance-conferring mutations. We used PAUP[72] version 4.0b10 to reconstruct gains and losses of drug-resistance-conferring mutations across the phylogenetic tree. We performed this analysis both for the full phylogeny of all 5,310 strains, as well as for individual phylogenies for each of the 11 geographic subregions and five lineages with >30 strains. PAUP was run using a cost matrix that assigned a 20× greater cost for a loss event relative to that of a gain event.

When examining the relative ordering of resistance-conferring mutations at two different nodes, we removed portions of the tree with potentially ambiguous topology. We removed node pairs from our analysis when the ancestral node had a bootstrap value <90%, as well as node pairs for which the longest of the individual branch lengths separating them was $>10^{-4}$. 25% of the branches in our phylogeny had branch lengths $>10^{-4}$. Our combined branch length and bootstrap filtering removed a total of 48% of the node pairs.

Assigning dates to phylogeny nodes. BEAST[29] version 1.8.2 was used to estimate dates of acquisition of drug-resistance-conferring mutations in the phylogenies of strains belonging to lineages 1–4. Lineages 1 and 3 contained a small enough number of strains to run BEAST (494 and 431 strains, respectively). However, because the numbers of strains in lineages 2 and 4 were greater than those allowed by the current capabilities of the BEAST algorithm, we subdivided these lineages into subsets of 400–700 strains and ran BEAST separately on each of these subsets. First, we removed very closely related strains from lineages 2 and 4. To do this, we clustered strains using simple agglomerative hierarchical clustering. For each cluster that contained multiple sequences with <10 SNP differences in the core region aligning to H37RV, we kept only one strain. This reduced the number of unique strains in lineage 2 to 978, and the number of unique strains in lineage 4 to 1,556. We then manually examined the phylogenies to split the remaining lineage 2 strains into two clades (lineage 2a with 462 strains, and lineage 2b with 516 strains) and the remaining lineage 4 strains into three clades (lineage 4a with 423 strains, lineage 4b with 413 strains and lineage 4c with 720 strains). We constructed a phylogenetic tree for each of these seven subsets using FastTree[69] version 2.1.8 (one tree from lineage 1, two trees from lineage 2, one tree from lineage 3 and three trees from lineage 4). We then ran BEAST to estimate dates of acquisition of drug resistance mutations in these seven clades representing lineages 1–4.

Because of the large size of our data set and the small spread in sample isolation dates, we used simplified parameters and a fixed evolutionary rate. We ran BEAST twice, using two fixed values for the evolutionary rate (representing an upper and lower bound of possible evolutionary rates). Because of the wide

range of evolutionary rates previously observed in *M. tuberculosis*, including varying rates for different strains[6,12,20,38,73–75], we used a lower bound of 0.3 mutations per genome per year (ref. 73) and an upper bound of 0.6 mutations per genome per year (ref. 6) to cover the entire range of published rates across all lineages. Isolation dates for each sample were used as input to BEAST. If our metadata included only a range of isolation dates, then we selected the midpoint of this date range. We enforced the topology of our input tree that was generated using FastTree[69]. We used the following parameters when running BEAST: GTR, empirical base frequencies and no site heterogeneity model. BEAST was run for a minimum of 10 million iterations, sampling every 1,000 iterations. The program Tracer was used to examine mixing and effective sample size to assess chain length and model convergence. If the effective sample size (ESS) with 10 million iterations for all variables was not >100, then BEAST was run for additional iterations, until ESS values were all >100. The first 1 million iterations were excluded as 'burn-in'. Estimated dates are given with 95% highest-posterior-density (HPD) intervals.

We also used the BEAST data to calculate a date for the arisal of each MDR node. To calculate the number of strains descending from each MDR node, we included the closely related strains that were removed after performing the hierarchical clustering.

Data availability. Newly sequenced TB-ARC data have been deposited in the Sequence Read Archive (SRA) under accession codes PRJNA235852 (India), PRJNA217391 (MRC), PRJNA219826 (CDRC), PRJNA200335 (Belarus), PRJNA229360 (Sweden), PRJNA220218 (Moldova), PRJNA233386 (Romania), and PRJNA237443 (Iran). Accessions for all newly sequenced data, as well as previously published data, are listed in **Supplementary Table 1**. Data are also available on the Broad Institute's website (https://olive.broadinstitute.org/projects/tb_arc).

65. Larsen, M.H., Biermann, K., Tandberg, S., Hsu, T. & Jacobs, W.R. Jr. Genetic manipulation of *Mycobacterium tuberculosis*. *Curr. Protoc. Microbiol.* **6**, 10A.2 (2007).
66. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome-assembly improvement. *PLoS One* **9**, e112963 (2014).
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
68. Salazar, A., Earl, A., Desjardins, C. & Abeel, T. Normalizing alternate representations of large sequence variants across multiple bacterial genomes. *BMC Bioinformatics* **16**, 116 (2015).
69. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
70. Walker, T.M. *et al.* Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* **15**, 1193–1202 (2015).
71. Farhat, M.R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **45**, 1183–1189 (2013).
72. Swofford, D. *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0* (Illinois Natural History Survey, 1989).
73. Guerra-Assunção, J.A. *et al.* Large-scale whole-genome sequencing of *M. tuberculosis* provides insights into transmission in a high-prevalence area. *eLife* **4**, e05166 (2015).
74. Roetzer, A. *et al.* Whole-genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med.* **10**, e1001387 (2013).
75. Ford, C.B. *et al.* Use of whole-genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat. Genet.* **43**, 482–486 (2011).

doi:10.1038/ng.3767