HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2017 July 16.

Published in final edited form as:

Nat Genet. 2017 March ; 49(3): 395–402. doi:10.1038/ng.3767.

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

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

[1]Broad Institute of M.I.T. and Harvard, 415 Main Street, Cambridge, MA, 02142, 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, MA, USA [4]Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands [5]Johns Hopkins University, Baltimore, MD, USA [6]National Institute of Allergy and Infectious Disease, National Institute of Health, USA [7]Medical Research Council, TB Platform, Pretoria, South Africa

Corresponding author: Dr. Ashlee M. Earl, The Broad Institute of MIT & Harvard, 415 Main Street, Cambridge, MA 02142, USA, aearl@broadinstitute.org, Phone: +1 (617) 714-7927.

*These authors contributed equally to this work

URLs

GeneXpert, http://www.cepheid.com/us/cepheid-solutions/clinical-ivd-tests/critical-infectious-diseases/xpert-mtb-rif

Hain Genotype MTBDR*plus*, http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/genotype-mtbdrplus.html

Hain Genotype MTBDR*sl* Line Probe Assay, http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/genotype-mtbdrsl.html

Hain Genotype MTBDRsl Line Probe Assay version 2.0, http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/tuberculosis/genotype-mtbdrsl.html

Accession Codes

SRA accession codes for newly sequenced TB-ARC data: PRJNA235852 (India), PRJNA217391 (MRC), PRJNA219826 (CDRC), PRJNA200335 (Belarus), PRJNA229360 (Sweden), PRJNA220218 (Moldova), PRJNA233386 (Romania), PRJNA237443 (Iran).

Author Contributions

ALM, KAC, TA, CAD, BWB and AME conceived the project. ALM, KAC, TA, CAD, AS analyzed the data. ALM, KAC, and AME interpreted results. ALM and KAC wrote the manuscript. DTA, CB, JB, SBC, SNC, AG, JG, AMJ, MJ, PJ, JSL, LM, MM, DN, EN, ER, AS, WS, AAV, KW, AZ, LEV, GHC, SED, JE, PF, JEG, AR, VC, DH, PRH, SN, ASP, AS, SS, MVW, DA, WRB, TC, and SH were involved in sample acquisition and handling including oversight of these activities. All authors critically read and revised the manuscript.

Competing financial Interests Statement

The authors have no competing financial interests.

[8]International Tuberculosis Research Center, Changwon and Department of Microbiology, Yonsei University College of Medicine, Seoul, South Korea [9]Office of Cyber Infrastructure and Computational Biology, National Institute of Health, USA [10]Clinical Hospital of Pneumology Leon Daniello Cluj Napoca, Romania [11]Makerere University, Department of Medical Microbiology, Mycobacteriology Laboratory, Kampala, Uganda [12]The Public Health Agency of Sweden, Sweden [13]University of Sciences, Techniques and Technologies of Bamako (USTTB), Bamako, Mali [14]Novasano Health and Science [15]Microbiology & Morphology Laboratory Phthisiopneumology Institute, Chisinau, Moldova [16]Mycobacteriology Research Centre, National Research Institute of Tuberculosis and Lung Disease (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran [17]Republican Research and Practical Centre for Pulmonology and Tuberculosis, Belarus [18]Harvard T. H. Chan School of Public Health, Boston, MA, USA [19]Boston Medical Center, Boston, MA, USA [20]Boston University, Boston, MA, USA [21]National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan [22]National Institute for Research in Tuberculosis, India [23]Rutgers-New Jersey Medical School, New Jersey, USA [24]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA [25]Department of Global Health and Social Medicine, Harvard Medical School, Division of Global Health Equity, Brigham and Women's Hospital, Boston, MA [26]TBResist Global Genome Consortium

Abstract

Multidrug-resistant tuberculosis (MDR-TB), caused by drug resistant strains of *Mycobacterium tuberculosis*, is an increasingly serious problem worldwide. In this study, we examined a dataset of 5,310 *M. tuberculosis* whole genome sequences from five continents. Despite great diversity with respect to geographic point of isolation, genetic background and drug resistance, patterns of drug resistance emergence were conserved globally. We have identified harbinger mutations that often precede MDR. In particular, the *katG* S315T mutation, conferring resistance to isoniazid, overwhelmingly arose before rifampicin resistance across all lineages, geographic regions, and time periods. Molecular diagnostics that include markers for rifampicin resistance alone will be insufficient to identify pre-MDR strains. Incorporating knowledge of pre-MDR polymorphisms, particularly *katG* S315, into molecular diagnostics will enable targeted treatment of patients with pre-MDR-TB to prevent further development of MDR-TB.

Introduction

Drug-resistant *Mycobacterium tuberculosis* is a threat to global tuberculosis (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 multidrug-resistant (MDR), defined as having resistance to both isoniazid and rifampicin[2]. Therapeutic regimens for MDR-TB can exceed 18 months, and include agents that often confer significant adverse effects[3]. 0.5% of global TB cases are now considered extensively drug-resistant (XDR), defined as MDR with additional resistance both to fluoroquinolones and at least one second-line injectable drug[2].

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

The global front-line molecular diagnostic for drug-resistant *M. tuberculosis*, GeneXpert MTB/RIF (GeneXpert, Sunnyvale, California, USA, *see urls*), simultaneously detects the presence of *M. tuberculosis* and identifies rifampicin resistance[5]. While GeneXpert identifies patients harboring rifampicin-resistant strains for initiation of MDR-TB treatment, this test may not identify resistance at the earliest available opportunity. In a recent analysis of a large collection of *M. tuberculosis* clinical isolate genomes from South Africa[6], Cohen *et al.* showed that the overwhelming majority of MDR- and XDR-TB evolved resistance to isoniazid prior to resistance to rifampicin. This result was consistent with another recent genomic analysis of strains from Russia[7] and 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].

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

Results

Drug resistance arises by universal mechanisms across the globe

In order to examine global phylogeographic patterns, including the order of drug resistance mutation evolution 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 part of fourteen published studies that used Illumina technology (Supplementary Table 1)[6,9–22]. After quality filtering (see Methods), our dataset included 5,310 genome sequences representing *M. tuberculosis* strains from 48 countries and 17 United Nations (UN) defined geographic regions[23] (Supplementary Note A; Supplementary Figures 1–4; Supplementary Tables 1–3). Though our dataset represented a broad diversity of TB strains from many global regions, their phylogeographic distribution did not perfectly match the actual distribution of TB burden worldwide (Supplementary Figure 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, *M. africanum* West African Type I; lineage 6, *M. africanum* West African Type II; and the deep-branching lineage 7), as well as *M. bovis*. Unsurprisingly, lineages 1–4 predominated (99.2%), consistent with the previously described limited geographic and host distributions of lineages 5–7 (Supplementary Table 3)[21,25,26].

To examine the distribution of drug resistance in this sample, for each isolate genome we computationally predicted resistance to eight drugs[27] using a curated list of polymorphisms associated with resistance (see Methods; Supplementary Table 4). As phenotypic drug resistance information was unavailable for most datasets, we did not incorporate phenotypic information into our analysis. We identified a total of 392 unique drug-resistance polymorphisms in at least one strain (Supplementary Tables 2,5). In comparison to expected global resistance rates, we observed higher rates of resistance, with 962 strains (18%) carrying mutations for both rifampicin and isoniazid resistance and lacking mutations for ofloxacin and kanamycin resistance (MDR sensu stricto), and 257 (5%) of strains carrying 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 level resistance (MDR genotype plus mutations to either ofloxacin or kanamycin). Over half of the sequenced strains did not carry any resistance-conferring mutation so were predicted to be drug susceptible (Figure 1; Supplementary Figures 4–6; Supplementary Table 2).

Drug-resistance was identified in nearly all (15 of 17) UN regions for which we had data, though its regional distribution varied considerably (Supplementary Note A; Supplementary Figures 7–10). In certain regions of the globe, we observed high numbers of closely related strains with nearly identical sets of resistance mutations, which can be attributed to clonal transmission. Because our dataset contained isolates from several known outbreaks[6,12], 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 (see 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 Note B; Supplementary Tables 7–8), suggesting that drug resistance has arisen via similar mechanisms irrespective of geography. This was also true for the evolution of MDR and XDR, which we calculated evolved independently 573 and 138 times within our dataset, 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. 360 of the 573 arisals (63%) led to a single MDR strain in our dataset (*de novo* evolution), whereas 213 arisals (37%) resulted in two or more descendant MDR strains, likely indicating person-to-person transmission of MDR-TB.

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

In an earlier analysis involving a smaller, South African dataset[6], we showed that isoniazid resistance evolved prior to rifampicin resistance in almost all cases. To determine whether this ordering of mutation acquisition held 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 portions of the phylogeny with ambiguous topologies (see 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 prior to resistance to second-line drugs (Supplementary Note C), as would be expected from the order in which antituberculous drugs are utilized in clinical practice. We also observed that mutations conferring isoniazid resistance overwhelmingly arose prior to any other mutation implicated in resistance (Figure 2; Supplementary Table 9; and Supplementary Note C), despite substantial complexity in the types and ordering of evolved mutations (Figure 3). Strikingly, isoniazid resistance predated rifampicin resistance in 96% of pairwise comparisons (155 out of 162), a pattern that remained true regardless of lineage or geographic source (Figure 4). While the majority of this effect was due to *katG* mutations (98% in 114 out of 116 pairings), non-*katG* isoniazid mutations followed this same pattern (89%, 41 out of 46 pairings). Thus, the provenance of global MDR was overwhelmingly isoniazid-resistant strains. In particular, strains carrying the *katG* S315 mutation frequently gained rifampicin resistance, whereas only a very small minority (4%) of global MDR arisals were due to gain of isoniazid resistance within a rifampicin-resistant background, despite the presence of 48 rifampicin mono-resistant strains, and 152 (3% of total) isolates containing rifampicin resistance, but no isoniazid resistance.

One possible explanation for this striking result is that isoniazid entered into clinical use approximately 20 years before rifampicin (1971–1993, depending on geography)[28] resulting in ancestral *M. tuberculosis* populations having differing exposures to these drugs, and potentially impacting the ordering of acquisition of resistance in favor of isoniazid resistance before rifampicin resistance. To test this hypothesis, we predicted the date for the arisal of each isoniazid and rifampicin resistance mutation using BEAST[29], and then tallied the number of co-arisals of resistance to both drugs occurring at various posterior time cutoffs to the present (see Methods) starting at 1971 (the date of introduction of rifampicin) and ending at 2000 (a later date to account for lag in the timing of rifampicin's widespread use) (Supplementary Tables 9–11). Our results revealed that, regardless of the date increment or evolutionary rate chosen (Methods; Supplementary Note D), *katG* S315-mediated resistance arose prior to rifampicin resistance 92%–98% of the time (Supplementary Table 10), indicating that, even during the era when isoniazid and rifampicin were given in combination, isoniazid resistance predated rifampicin resistance.

Diagnostics for early detection of pre-MDR M. tuberculosis

Contributors to the current global burden of MDR-TB include not only historic emergences of MDR, which led to person-to-person transmission of MDR-TB, but also ongoing *de novo* evolution. Of the 573 MDR arisals in our dataset, we estimated that 67% occurred since 2004 (see Methods). Thus, new strategies for curbing the emergence of MDR, such as identifying MDR precursors, will be critical to the control of MDR worldwide. GeneXpert, currently one of the front-line diagnostic tests used to exclusively identify rifampicin resistance-conferring mutations in the RRDR of *rpoB*[5], is commonly used globally as a proxy for detecting MDR-TB. The habitual ordering of isoniazid resistance prior to rifampicin resistance indicates that GeneXpert MTB/RIF serves as an appropriate proxy for MDR-TB, and is well suited to detect MDR in all geographic regions and all lineages of *M. tuberculosis* (Figure 4). However, as rifampicin mutations (detectable by GeneXpert) are

rarely the first resistance mutations to emerge, by the time a mutation that is detectable by GeneXpert develops, oftentimes there is pre-existing resistance to multiple additional drugs, including second-line drugs (Figure 5). As we excluded nodes at which we were unable to disambiguate the ordering of resistance, our estimates represent a lower bound on the number of nodes where resistance to other drugs was gained prior to a GeneXpert-detectable node (Supplementary Figure 11).

Diagnostics that identify mutations present before MDR emerges would provide an opportunity to identify drug resistance during a period where there are both greater therapeutic options and improved treatment outcomes[30]. While the results from our pairwise ordering of resistance acquisition clearly demonstrated *katG* S315T-mediated isoniazid resistance as 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 mutations that commonly evolved prior to development of MDR-TB ("pre-MDR-TB mutations"). For this set of mutations, we quantified the fraction of MDR-defining nodes where one of these pre-MDR mutations had evolved prior to the development of MDR-TB in order to determine how much resistance to other drugs had unambiguously arose prior to MDR (Figure 6). Our analysis (see Methods) revealed a set of 16 resistance mutations (of 340 total found among MDR and XDR strains) that arose prior to MDR-TB a minimum of two independent times (Supplementary Note E; Supplementary Table 12).

Surprisingly, we observed resistance mutations for all eight drugs among this set of pre-MDR mutations. However, many of these mutations evolved infrequently (Figure 6a), and thus would likely have low negative predictive value if included on a diagnostic panel aimed at identifying pre-MDR-TB. Of the more frequently occurring mutations, we observed that many arose repeatedly after resistance to multiple other drugs (Figure 6b), indicating that their typical emergence pattern is not limited to early resistance. Conversely, *katG* S315T stood out as a frequently occurring mutation (Figure 6a) with very few instances of other resistance arising prior to its gain (Figure 6b). Despite the high level of complexity in the stepwise acquisition of drug resistance mutations in *M. tuberculosis* (Figure 3), *katG* S315T was by far the most common mutation to evolve prior to MDR (Figure 2, Supplementary Notes E–G; Supplementary Tables 12–13). Of the 321 independent arisals of *katG* S315T in our dataset, 302 (94%) occurred at the earliest node where drug resistance was present.

Discussion

We constructed the largest global dataset of *M. tuberculosis* whole genomes analyzed to date, consisting of 5,310 diverse strains. Though the global distribution of strains in this dataset does not reflect the global incidence of TB for some regions (Supplementary Figure 1b), our unique dataset had a broad geographic 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 identify harbinger resistance mutations that emerged prior to development of MDR-TB.

We observed that MDR- 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- 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 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 need for regional specialization. Without phenotypic drug susceptibility data for all included strains, we were not able to identify novel drug resistance mutations or quantify the amount of drug resistance that remains unexplained by our curated list of polymorphisms; however, we expect it 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 geographic 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 mono-resistance rarely leads to MDR-TB. Though 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 arose, we have taken care to minimize such effects (Supplementary Note H). Supporting our results, this relative ordering of isoniazid and rifampicin resistance was consistent with prior findings based on genomic data in Russia[7], South Africa[6] and South America[7], as well as analysis of a large global collection of phenotypic data[8].

Why would isoniazid resistance arise first? We show that the earlier clinical introduction of isoniazid is not a major contributor to the earlier arisal of isoniazid resistance; our dating analysis indicates that isoniazid resistance arises before rifampicin resistance across all time periods, including recently (Supplementary Note D). However, there are many alternative, though 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 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 1) mutations in *katG*, a non-essential gene, which result in failure to activate isoniazid, and 2) 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* 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*, S315T, accounted for the majority of isoniazid resistance arisals, and that, overall, there were ~20% more independent arisals of resistance to rifampicin than 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* S315T mutation, are more likely to develop resistance to other drugs. While previous *in vitro* studies have shown a difference in the types of rifampicin resistance mutations arising in isoniazid-resistant backgrounds[39], there is no evidence that isoniazid resistant strains are transformed into "hypermutators"[40]. Further, the sequence surrounding the *katG* S315 site does not appear susceptible to mutation nor does it appear to be a mutational hotspot *in vitro*[37]. However, as others and we have shown, this specific mutation is common among isoniazid resistant clinical isolates, indicating that it is well tolerated *in vivo*. This is likely due to the fact that this mutation preserves mycobacterial catalase activity while still preventing activation of isoniazid[41]. This preserved fitness may impact 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 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 differential clinical penetrations of the drugs into lesions[46], that may influence the ordering 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], treatment of suspected cases of latent TB with only isoniazid, could provide an opportunity for isoniazid resistance to develop prior to 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 dataset.

Early identification and appropriate treatment of individuals with isoniazid mono-resistant strains, such as with non-isoniazid based regimens[51], may prevent the selection and eventual transmission of additional MDR strains. The worldwide case rate of isoniazid mono-resistance is estimated to be as high as 2%–15%[52–56], or 200,000–1.4 million cases per year. Several studies have shown that patients harboring isoniazid mono-resistant strains have worse clinical outcomes than those harboring susceptible strains[57–59], and enhanced treatment regimens for such strains resulted in lower rates of treatment failure and acquired drug resistance[59]. One large retrospective cohort study also points to early detection of isoniazid mono-resistance for improved outcomes[52]. In particular, the *katG* S315T mutation 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 dataset confirms that GeneXpert, currently the most widely used rapid molecular diagnostic for the diagnosis of *M. tuberculosis* and MDR-TB, performs excellently as a surrogate marker for MDR, irrespective of lineage and/or region of the globe. However, as it

includes only rifampicin resistance mutations, GeneXpert does not identify drug resistance at the earliest available opportunity. Thus, diagnostic algorithms that rely upon application of GeneXpert alone will allow rifampicin-susceptible but otherwise drug-resistant strains to propagate unchecked. In fact, our evolutionary analysis revealed that, by the time a GeneXpert identifiable rifampicin resistance mutation is acquired, oftentimes multiple additional resistance-conferring mutations are already present. Additional commercially available diagnostics, such as the Hain MTBDR*plus* and Hain MTBDR*sl* ver 2.0 line probe assays, are available and detect a broader set of resistance mutations. Despite logistical considerations that make practical application of this technology more difficult—such as moderate turn-around times and need for specialized laboratory facilities—line probe assays are able to detect isoniazid resistance (HainMTBDR*plus*) 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 GeneXpert. 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 dataset encompassing 5,310 strains of *M. tuberculosis*, we observed that recent *de novo* emergence of MDR-TB in the last 10 years is a significant contributor to global MDR today. Thus, to stem the development of additional MDR strains, one should seek to identify resistant strains in the “pre-MDR” stage, during which there are additional therapeutic options and improved treatment outcomes. Identification of harbinger mutations, such as *katG* S315T, may serve as an early warning signal that MDR may soon develop. Focusing on common, early-occurring mutations could improve the design of diagnostic tests, aimed to target 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 emergence in a given population. If substantiated, 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.

Online Methods

Whole Genome Sequencing and datasets

References detailing the sequencing methods for the published datasets can be found in Supplementary Table 1. Sequencing data for the TB-ARC projects (Supplementary Table 1) was 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 (http://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 Indian dataset, 223 drug resistant and drug sensitive strains, representative of lineages found in India (particularly lineages 1 and 3) were selected and sequenced from studies in Tiruvallur and Madurai in Southern India. For the MRC dataset, 189 primarily drug-resistant strains from South African and Swaziland were selected for sequencing. For the CDRC dataset, 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 Swedish dataset, 150 genomes were collected, primarily from Sweden's immigrant population. This set includes a complete collection of all 141 MDR and XDR strains identified nationwide in Sweden in the period 2003–2013. For the dataset from Moldova, 95 genomes were selected from the countrywide specimen and data repositories. For the Romanian dataset, 34 genomes were sequenced, with the goal of describing drug resistant strains circulating in Romania and their diversity. For the Iranian dataset, 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[66]. Library preparation and whole genome sequencing (WGS) were performed as previously described on the Illumina HiSeq 2000 at the Broad Institute [67]. Sequencing data were submitted to the Sequence Read Archive 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 accessions in Supplementary Table 1). Reads were mapped onto a reference strain of H37Rv (GenBank accession number CP003248.2) using BWA version 0.7.10[68]. Variants were identified using Pilon version 1.11 as described [67]. The global *M. tuberculosis* lineage designations used in our analysis, as well as each strain's spoligotype, were predicted using digital spoligotyping as in Cohen et al., 2015[6].

We eliminated 824 strains that did not pass our quality control filters: average sequencing depth of coverage >20X; 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 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 mycobacteria contamination; as well as 13 additional duplicate patient samples. These filters resulted in a final set of 5,310 strains for analysis.

Emu[69] 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[23] 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[70] 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], the Hain Genotype MTBDR*plus*, the Hain Genotype MTBDR*sl* Line Probe Assay, as well as the Hain MTBDR*sl* ver 2.0 (*see urls*). Additional resistance mutations were selected for inclusion based on laboratory evidence and recent compelling genomic evidence that these mutations encode for resistance (Supplementary Tables 4–5). Because of greater uncertainty calling longer variants in our data, we excluded insertions and deletions longer than 10bp. Using this curated list, we identified 392 drug-resistance mutations among the 231,898 total variants observed in our full dataset across 5,310 strains.

Recent reports suggest that currently tabulated mutation sets account for the majority of phenotypic resistance[27,71,72]. While pyrazinamide was predicted at a lower sensitivity than other drugs in one recent analysis[71], including all loss-of-function mutations in *pncA*, as we have done in this study, would likely greatly improve sensitivity to detect pyrazinamide resistance.

Evolution of drug resistance mutations

We used PAUP[73] version 4.0b10 to reconstruct gains and losses of drug resistance 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 20x greater cost for a loss event relative to a gain event.

When examining the relative ordering of resistance 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 where the longest of the individual branch lengths separating them was >1e-4. 25% of the branches in our phylogeny had branch lengths >1e-4. Our combined branch length and bootstrap filtering removed a total of 48% of the node pairs.

Dating the phylogeny

BEAST[29] version 1.8.2 was used to estimate dates of acquisition of drug resistance mutations in the phylogenies of 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 size of lineages 2 and 4 were beyond the current capabilities of the BEAST algorithm, we subdivided these lineages into subsets of size 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 containing 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[70] 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 dataset and small spread in sample isolation dates, we used simplified parameters and a fixed evolutionary rate. We ran BEAST twice, using a fixed upper and lower bound for the evolutionary rate. Because of the wide range of rates previously observed in *M. tuberculosis*, including varying rates for different strains[6,11,19,38,74–76], we used a lower bound of 0.3 mutations/ genome/ year [74] and an upper bound of 0.6 mutations/ genome/ year[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, we selected the midpoint of this date range. We enforced the topology of our input tree generated using FastTree[70]. 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 in order 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. In order 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 Statement

Newly sequenced TB-ARC data has been deposited in the SRA database. Accessions for all newly sequence data, as well as previously published data, are listed in Supplementary Table 1. Data are also available on the Broad Institute’s website (http://olive.broadinstitute.org/projects/tb_arc)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank the Broad Institute’s Genome Sequencing Platform and Assembly and Annotation teams, including Sarah K. Young, Margaret E. Priest, Terrence P. Shea, Bruce J. Walker, Lucia Alvarado, Dr. Michael G. Fitzgerald, Sharvari Gujja, Susanna Hamilton, Clinton Howarth, Jeffrey D. Larimer, Matthew D. Pearson, Dr. Qiandong Zeng, and Jennifer Wortman. We would like to thank Joe Romano and Arlin Keo for help with lineage detection, and Maria Dr. Mercedes Zambrano, Beatriz Ferro and Juan Carlos Rozo for isolation and phenotypic characterization of strains. We are also grateful to members of TBResist for contribution of their strains, phenotypic

data, and expertise, and their help in forging collaborations, and to Dr. Veronique Dartois, Dr. Dick Thomas, Dr. Deb Hung and Dr. David 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, under Grant No. U19AI110818, and Contract No. HHSN272200900018C to the Broad Institute, and Contract No. HHSN27220009000050C to the TB Clinical Diagnostics Research Consortium, and the Intramural Research Program of NIAID (CB and LEV) together with 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. Barry). Funding was also provided by NIH Grant No. 5U01AI069924-07 for IeDEA (A. Pym); Howard Hughes Medical Institute and NIH R01 AI110386 for "Host-pathogen Interactions in a failing global lineage of MTBC: *M. africanum*" (W. Bishai). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. Farmer P, et al. The dilemma of MDR-TB in the global era. Int J Tuberc Lung Dis. 1998; 2:869–76. [PubMed: 9848606]
2. WHO. Global Tuberculosis Report 2015. 2015; 20
3. Awofeso N. Anti-tuberculosis medication side-effects constitute major factor for poor adherence to tuberculosis treatment. Bulletin of the World Health Organization. 2008; 86:161–240. [PubMed: 18368195]
4. Pietersen E, et al. Long-term outcomes of patients with extensively drug-resistant tuberculosis in South Africa: a cohort study. Lancet. 2014; 383:1230–9. [PubMed: 24439237]
5. Boehme CC, et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med. 2010; 363:1005–15. [PubMed: 20825313]
6. Cohen K, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, Shea TP, Walker BJ, Bantubani N, Almeida D, Alvarado L, Chapman S, Mvelase NR, Duffy EY, FitzGerald MG, Govender P, Gujja S, Hamilton S, Howarth C, Larimer JD, Maharaj K, Pearson MD, Priest ME, Zeng Q, Padayatchi N, Grosset J, Young SJ, Wortman J, Mlisana K, O'Donnell MR, Birren BW, Bishai WR, Pym AS. Evolution of extensively drug-resistant tuberculosis over four decades revealed by whole genome sequencing of Mycobacterium tuberculosis from KwaZulu-Natal, South Africa. Plos Medicine. 2015
7. Eldholm V, et al. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nat Commun. 2015; 6:7119. [PubMed: 25960343]
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. 2013; 9:e1002973. [PubMed: 23555210]
9. Biek R, et al. Whole genome sequencing reveals local transmission patterns of Mycobacterium bovis in sympatric cattle and badger populations. PLoS Pathog. 2012; 8:e1003008. [PubMed: 23209404]
10. Blouin Y, et al. Significance of the identification in the Horn of Africa of an exceptionally deep branching Mycobacterium tuberculosis clade. PLoS One. 2012; 7:e52841. [PubMed: 23300794]
11. Bryant JM, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis. 2013; 13:110. [PubMed: 23446317]
12. Casali N, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet. 2014; 46:279–86. [PubMed: 24464101]
13. Clark TG, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One. 2013; 8:e83012. [PubMed: 24349420]
14. Comas I, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 2013; 45:1176–82. [PubMed: 23995134]
15. Gardy JL, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011; 364:730–9. [PubMed: 21345102]
16. Guerra-Assuncao JA, et al. Recurrence due to relapse or reinfection 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. 2015; 211:1154–63. [PubMed: 25336729]

17. Merker M, et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat Genet. 2015; 47:242–9. [PubMed: 25599400]
18. Perdigao J, et al. Unraveling Mycobacterium tuberculosis genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. BMC Genomics. 2014; 15:991. [PubMed: 25407810]
19. Walker TM, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013; 13:137–46. [PubMed: 23158499]
20. Zhang H, et al. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013; 45:1255–60. [PubMed: 23995137]
21. Winglee K, Manson McGuire A, Maiga M, Abeel T, Shea T, Desjardins CA, Diarra B, Baya B, Sanogo M, Diallo S, Earl AM, Bishai WR. Whole genome sequencing of Mycobacterium africanum strains from Mali provides insights into the mechanisms of geographic restriction. PLoS Negl Trop Dis. 2015; 10:e0004332.
22. Wollenberg KR, Desjardins CA, Zalutskaya A, Slodovnikova V, Oler AJ, Quinones M, Abeel T, Chapman SB, Tartakovsky M, Gabrielan A, Hoffner S, Skrahin A, Birren BW, Rosenthal A, Skrahina A, Earl AM. Whole genome sequencing of Mycobacterium tuberculosis provides insight into the evolution and genetic composition of drug-resistant tuberculosis in Belarus. J Clin Microbiol. 2016 in press.
23. UN. United Nations Geoscheme. http://unstats.un.org/unsd/methods/m49/m49regin.htm
24. Gagneux S, et al. Variable host-pathogen compatibility in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2006; 103:2869–73. [PubMed: 16477032]
25. Yimer SA, 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. 2015; 53:1301–9. [PubMed: 25673798]
26. O'Reilly LM, Daborn CJ. The epidemiology of Mycobacterium bovis infections in animals and man: a review. Tuber Lung Dis. 1995; 76(Suppl 1):1–46.
27. Desjardins C, CK, Govender VM, Abeel T, Maharaj K, Walker B, Shea T, Almeida T, Manson A, Salazar A, Padayatchi N, O'Donnell M, Mlisana K, Wortman J, Birren B, Grosset J, Pym A, Earl A. Novel D-cycloserine Resistance Mechanism in Mycobacterium tuberculosis Revealed by Whole Genome Analysis. Nature Genetics. 2016; 48:544–551. [PubMed: 27064254]
28. Ma Z, Lienhardt C, McIlleron H, Nunn AJ, Wang X. Global tuberculosis drug development pipeline: the need and the reality. Lancet. 2010; 375:2100–9. [PubMed: 20488518]
29. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29:1969–73. [PubMed: 22367748]
30. Cattamanchi A, et al. Clinical characteristics and treatment outcomes of patients with isoniazid-monoresistant tuberculosis. Clin Infect Dis. 2009; 48:179–85. [PubMed: 19086909]
31. Velayati AA, et al. High prevelance of rifampin-monoresistant tuberculosis: a retrospective analysis among Iranian pulmonary tuberculosis patients. Am J Trop Med Hyg. 2014; 90:99–105. [PubMed: 24189362]
32. Johnsson K, Schultz PG. Mechanistic studies of the oxiation of isoniazid by the catalase peroxidase from *Mycobacterium tuberculosis*. J Am Chem Soc. 1994; 116:7425–7426.
33. Rozwarski DA, Grant GA, Barton DH, Jacobs WR Jr, Sacchettini JC. Modification of the NADH of the isoniazid target (InhA) from Mycobacterium tuberculosis. Science. 1998; 279:98–102. [PubMed: 9417034]
34. Quemard A, et al. Enzymatic characterization of the target for isoniazid in Mycobacterium tuberculosis. Biochemistry. 1995; 34:8235–41. [PubMed: 7599116]
35. Marrakchi H, Laneelle G, Quemard A. InhA, a target of the antituberculous drug isoniazid, is involved in a mycobacterial fatty acid elongation system, FAS-II. Microbiology. 2000; 146(Pt 2): 289–96. [PubMed: 10708367]
36. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003; 48:77–84. [PubMed: 12657046]

37. Bergval IL, Schuitema AR, Klatser PR, Anthony RM. Resistant mutants of Mycobacterium tuberculosis selected in vitro do not reflect the in vivo mechanism of isoniazid resistance. J Antimicrob Chemother. 2009; 64:515–23. [PubMed: 19578178]
38. Ford CB, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet. 2013; 45:784–90. [PubMed: 23749189]
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. 2012; 7:e29108. [PubMed: 22235262]
40. O'Sullivan DM, McHugh TD, Gillespie SH. The effect of oxidative stress on the mutation rate of Mycobacterium tuberculosis with impaired catalase/peroxidase function. J Antimicrob Chemother. 2008; 62:709–12. [PubMed: 18577539]
41. Pym AS, Saint-Joanis B, Cole ST. Effect of katG mutations on the virulence of Mycobacterium tuberculosis and the implication for transmission in humans. Infect Immun. 2002; 70:4955–60. [PubMed: 12183541]
42. Salverda ML, et al. Initial mutations direct alternative pathways of protein evolution. PLoS Genet. 2011; 7:e1001321. [PubMed: 21408208]
43. Ogbunugafor CB, Wylie CS, Diakite I, Weinreich DM, Hartl DL. Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. PLoS Comput Biol. 2016; 12:e1004710. [PubMed: 26808374]
44. Gumbo T, et al. Concentration-dependent Mycobacterium tuberculosis killing and prevention of resistance by rifampin. Antimicrob Agents Chemother. 2007; 51:3781–8. [PubMed: 17724157]
45. Pasipanodya JG, Gumbo T. A new evolutionary and pharmacokinetic-pharmacodynamic scenario for rapid emergence of resistance to single and multiple anti-tuberculosis drugs. Curr Opin Pharmacol. 2011; 11:457–63. [PubMed: 21807559]
46. Jutte PC, Rutgers SR, Van Altena R, Uges DR, Van Horn JR. Penetration of isoniazid, rifampicin and pyrazinamide in tuberculous pleural effusion and psoas abscess. Int J Tuberc Lung Dis. 2004; 8:1368–72. [PubMed: 15581207]
47. Diacon AH, et al. Early bactericidal activity of high-dose rifampin in patients with pulmonary tuberculosis evidenced by positive sputum smears. Antimicrob Agents Chemother. 2007; 51:2994–6. [PubMed: 17517849]
48. Long MW, Snider DE Jr, Farer LSUS. Public Health Service Cooperative trial of three rifampin-isoniazid regimens in treatment of pulmonary tuberculosis. Am Rev Respir Dis. 1979; 119:879–94. [PubMed: 110184]
49. Mitchison DA. Role of individual drugs in the chemotherapy of tuberculosis. Int J Tuberc Lung Dis. 2000; 4:796–806. [PubMed: 10985648]
50. Mills HL, Cohen T, Colijn C. Community-wide isoniazid preventive therapy drives drug-resistant tuberculosis: a model-based analysis. Sci Transl Med. 2013; 5:180ra49.
51. Jindani A, et al. High-dose rifapentine with moxifloxacin for pulmonary tuberculosis. N Engl J Med. 2014; 371:1599–608. [PubMed: 25337749]
52. Chien JY, et al. Treatment outcome of patients with isoniazid mono-resistant tuberculosis. Clin Microbiol Infect. 2015; 21:59–68. [PubMed: 25636929]
53. Fasih N, Rafiq Y, Jabeen K, Hasan R. High isoniazid resistance rates in rifampicin susceptible Mycobacterium tuberculosis pulmonary isolates from Pakistan. PLoS One. 2012; 7:e50551. [PubMed: 23226311]
54. Jenkins HE, Zignol M, Cohen T. Quantifying the burden and trends of isoniazid resistant tuberculosis, 1994–2009. PLoS One. 2011; 6:e22927. [PubMed: 21829557]
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. 2013; 37:E40–6. [PubMed: 23692157]
56. Nagu TJ, et al. Multi drug and other forms of drug resistant tuberculosis are uncommon among treatment naive tuberculosis patients in Tanzania. PLoS One. 2015; 10:e0118601. [PubMed: 25849784]

57. Villegas L, et al. Prevalence, Risk Factors, and Treatment Outcomes of Isoniazid- and Rifampicin-Mono-Resistant Pulmonary Tuberculosis in Lima, Peru. PLoS One. 2016; 11:e0152933. [PubMed: 27045684]
58. Espinal MA, et al. Standard short-course chemotherapy for drug-resistant tuberculosis: treatment outcomes in 6 countries. JAMA. 2000; 283:2537–45. [PubMed: 10815117]
59. Menzies D, et al. Standardized treatment of active tuberculosis in patients with previous treatment and/or with mono-resistance to isoniazid: a systematic review and meta-analysis. PLoS Med. 2009; 6:e1000150. [PubMed: 20101802]
60. Huyen MN, et al. Epidemiology of isoniazid resistance mutations and their effect on tuberculosis treatment outcomes. Antimicrob Agents Chemother. 2013; 57:3620–7. [PubMed: 23689727]
61. Denkinger CM, Pai M, Dowdy DW. Do we need to detect isoniazid resistance in addition to rifampicin resistance in diagnostic tests for tuberculosis? PLoS One. 2014; 9:e84197. [PubMed: 24404155]
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. 2016; 11:e0150321. [PubMed: 26934724]
63. Department of Health, R.o.S.A. Management of drug-resistant tuberculosis: policy guidelines. South Africa: 2013.
64. Pusch, t. 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/)
65. Banerjee A, et al. inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. Science. 1994; 263:227–30. [PubMed: 8284673]
66. Larsen MH, Biermann K, Tandberg S, Hsu T, Jacobs WR Jr. Genetic Manipulation of Mycobacterium tuberculosis. Curr Protoc Microbiol. 2007; Chapter 10(Unit 10A):2.
67. Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9:e112963. [PubMed: 25409509]
68. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. [PubMed: 19451168]
69. Salazar A, Earl A, Desjardins C, Abeel T. Normalizing alternate representations of large sequence variants across multiple bacterial genomes. BMC Bioinformatics. 2015; 16(Suppl 2):A8.
70. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009; 26:1641–50. [PubMed: 19377059]
71. Walker TM, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015; 15:1193–202. [PubMed: 26116186]
72. Farhat MR, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013; 45:1183–9. [PubMed: 23995135]
73. Swofford, D. Computer program distributed by the Illinois Natural History Survey. Champaign, Illinois: 1991. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1.
74. Guerra-Assuncao JA, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015; 4
75. Roetzer A, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013; 10:e1001387. [PubMed: 23424287]
76. Ford CB, et al. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet. 2011; 43:482–6. [PubMed: 21516081]

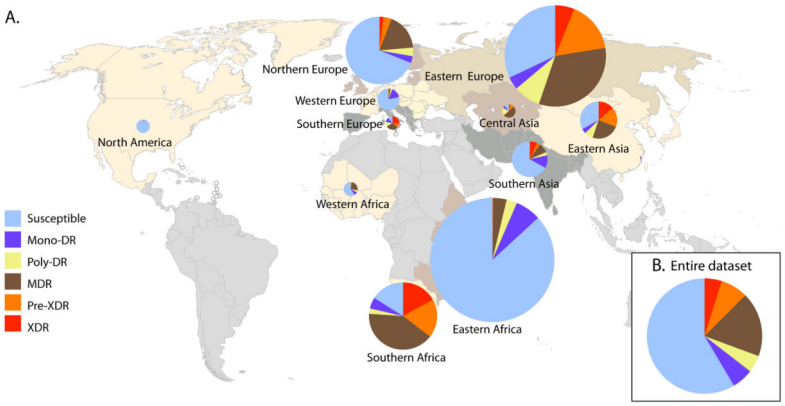

Figure 1.

A) Geographic distribution of *M. tuberculosis* isolates in our dataset by drug resistance pattern. This plot shows the distribution of the 5,310 *M. tuberculosis* isolates included in our dataset by drug resistance genotype (pie charts) and by 11 UN geographic subregions[23] (coloring), and is not meant to indicate the overall global incidence of TB or drug resistance. There were no strains in our dataset from geographic regions colored grey. UN geographic subregions with fewer than 30 strains were excluded from this figure. Map modified from a blank map of UN geographical subregions.[64] B) The overall proportion of drug-resistant strains identified among all 5,310 *M. tuberculosis* isolates in our dataset.

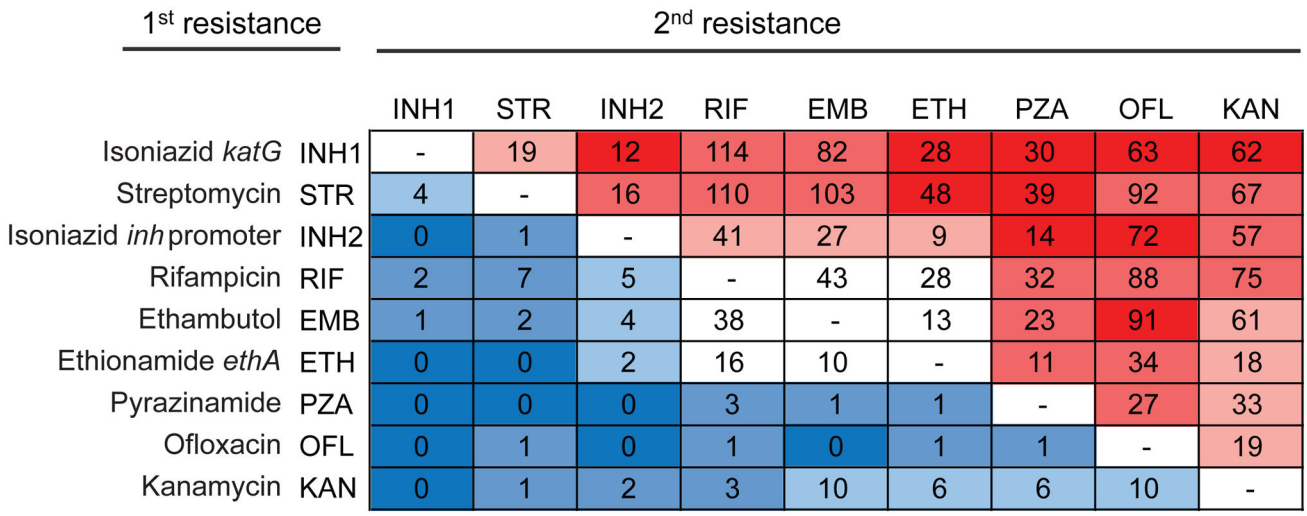

Figure 2.

Across the globe, isoniazid resistance was overwhelmingly the first step towards drug resistance. Acquisition of a *katG* S315 mutation precedes all other resistance mutations for the majority of instances in which the order of acquisition can 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 the drug resistance indicated by the row labeled “first resistance” was acquired before the drug resistance indicated by the column labeled “second resistance”. The shading color indicates the percentage of evolutionary events in which the “first resistance” clearly predates the “second resistance” for that drug pair. While *inhA* mutations can confer resistance to both isoniazid and ethionamide[65], we defined genotypic ethionamide resistance as mutations in only *ethA* to simplify the analysis and avoid double counting.

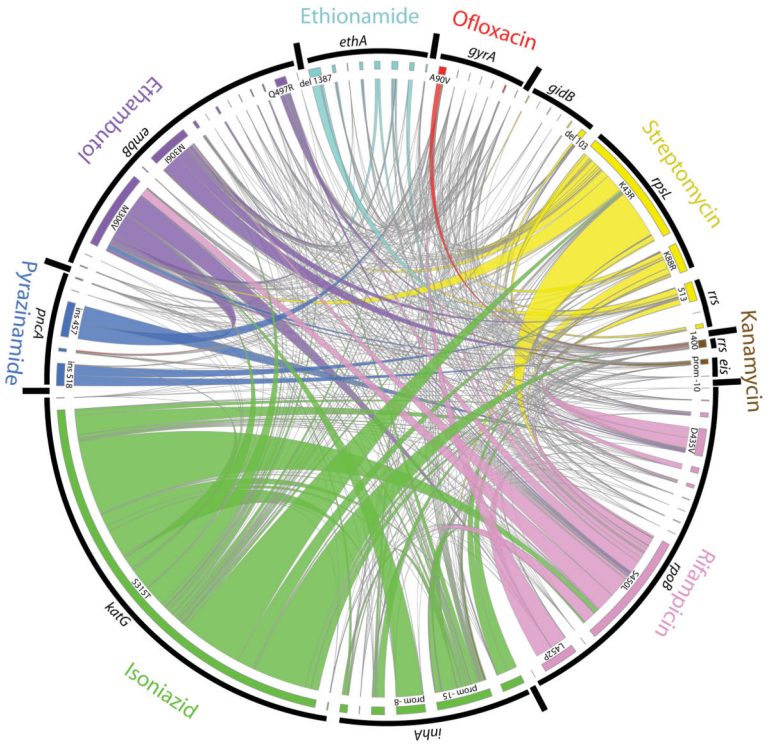

Figure 3.

Sequential acquisition of drug resistance mutations reveals that isoniazid resistance-conferring mutations, specifically *katG* S315T, most often come first in sequential pairs. This figure includes data from 71 mutations conferring drug resistance with at least 10 occurrences in our dataset, which represent 93% of all drug resistance mutations in our dataset. Using PAUP analysis to assign specific mutation gains to individual nodes on the phylogeny, we tabulated all routes of drug resistance acquisition across the full strain phylogeny, examining only those 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 (see Methods). The ribbons in this figure depict the number of times that each pair of mutations arose sequentially at adjacent nodes across the entire dataset. 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* S315T and a thin end at *rpoB* S450L indicates that *katG* S315T mutation arose prior to *rpoB* S450L much more frequently than the opposite). Each ribbon is colored according to the mutation more often occurring first in each sequential pairing.

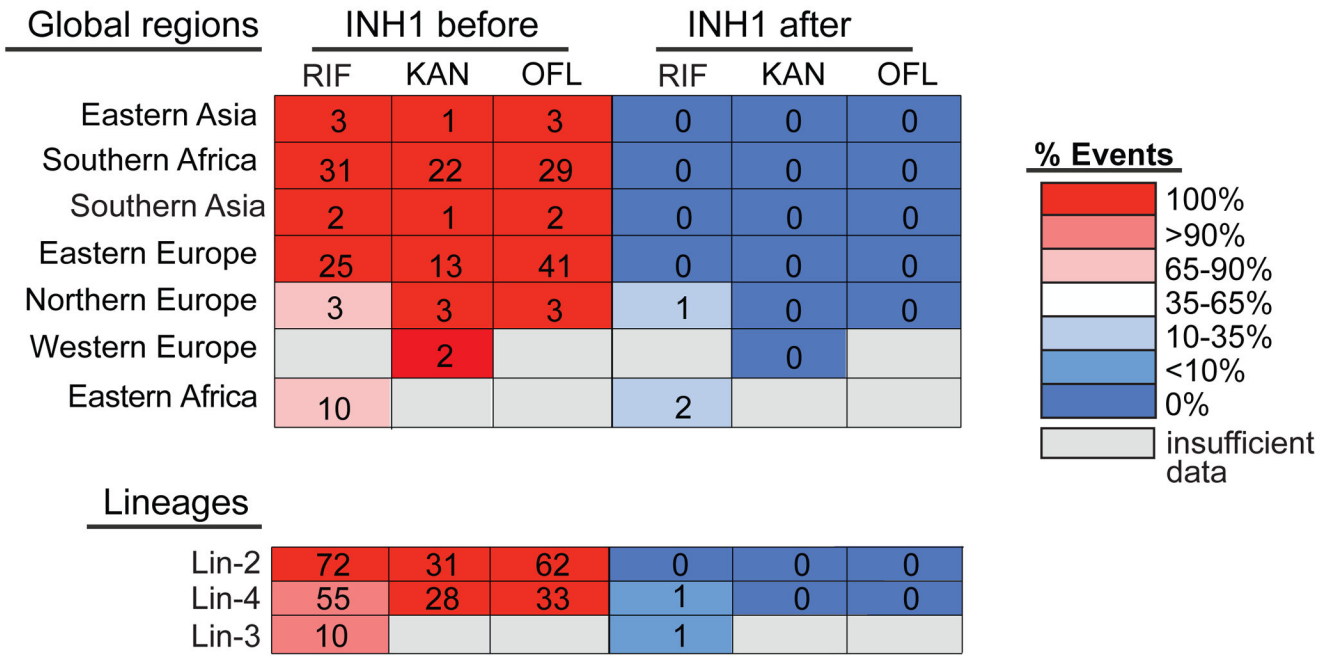

Figure 4.

In all lineages and global regions, the *katG* mutation S315T occurs first, and few examples of the reverse ordering are observed. We separately recalculated phylogenies for isolates from patients in each of the 11 UN subregions and 5 lineages with greater than 30 representatives (see Methods). This figure depicts the pairwise ordering of the *katG* S315T mutation in relation to mutations conferring resistance to the other three XDR-definings drugs (rifampicin (R), kanamycin (K), and ofloxacin (O)), within each individual *M. tuberculosis* lineage and geographic region. The numbers in panel B do not necessarily add up to the numbers in panel A, as the analyses of regions and lineages were performed individually, which can affect the number of arisals. Grey shading indicates that there were not sufficient pairings. 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*.

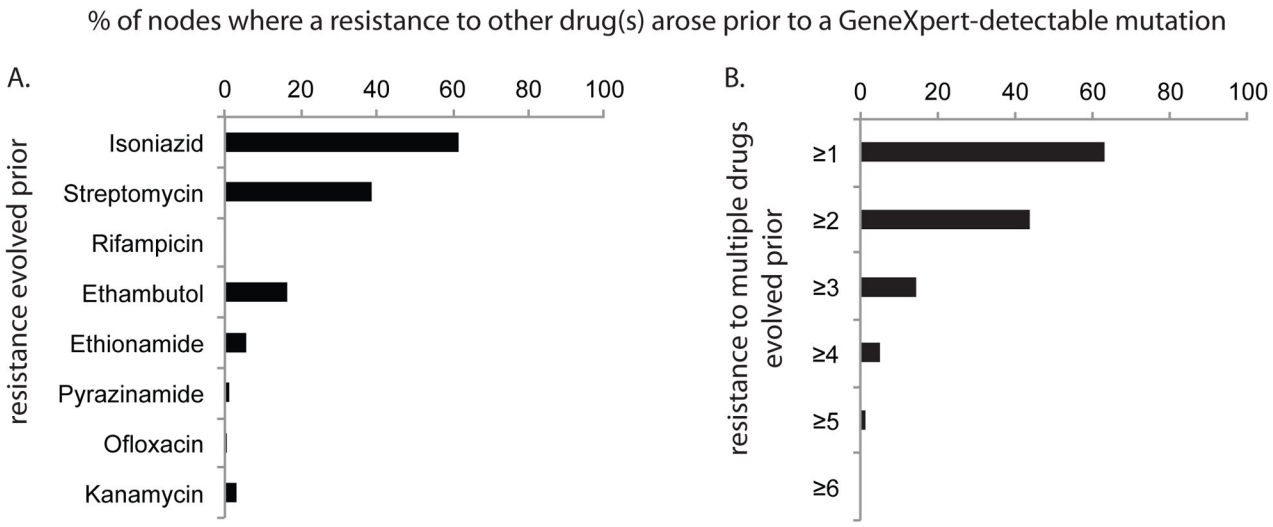

Figure 5.

Non-rifampicin drug resistance often precedes the arisal of GeneXpert mutations. Data are shown here for the nodes at which a GeneXpert mutation arose. A) This plot shows the percentage of GeneXpert nodes where resistance to each of eight drugs unambiguously preceded the arisal of GeneXpert mutations. Drug resistances that appeared to arise coincident to the GeneXpert node were excluded from this representation. More than one additional resistance could precede a single GeneXpert node. No strains contained additional rifampicin mutations arising before GeneXpert mutations. The bars represent a lower bound on the percentage of GeneXpert mutations preceded by additional resistance mutations, as we were unable to disambiguate ordering for a substantial number of nodes where additional mutations arose at the same node (see Supplementary Figure 11). B) Percentage of nodes where resistance to one or more other drugs unambiguously preceded the arisal of GeneXpert mutations. 13% of GeneXpert arisal nodes unambiguously had no additional drug resistance mutations arising prior.

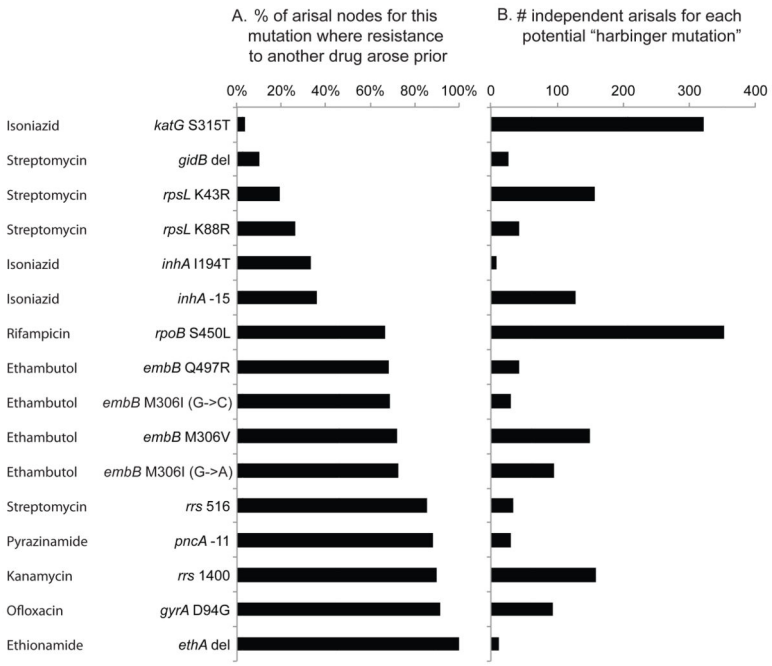

Figure 6.

katG S315T is a commonly occurring mutation with very little resistance to other drugs arising prior to its occurrence. A) For each of the 16 "pre-MDR" mutations, the percentage of nodes where resistance to another drug unambiguously preceded the arisal of that mutation. These bars represent a lower bound on the percentage of nodes preceded by another resistance mutation, as we were unable to disambiguate ordering for a substantial number of nodes where additional mutations arose at the same node. B) The number of independent arisals for each of 16 "pre-MDR" (or harbinger) mutations. Since there are two *embB* M306I mutations, the nucleotide change at position 4,247,609 is also indicated for these two variants.