View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Harvard University - DASH

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

Genomic Analysis Identifies Targets of Convergent Positive Selection in Drug Resistant Mycobacterium tuberculosis

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation	Farhat, M. R., B. J. Shapiro, K. J. Kieser, R. Sultana, K. R. Jacobson, T. C. Victor, R. M. Warren, et al. 2013. "Genomic Analysis Identifies Targets of Convergent Positive Selection in Drug Resistant Mycobacterium tuberculosis." Nature genetics 45 (10): 10.1038/ng.2747. doi:10.1038/ng.2747. http://dx.doi.org/10.1038/ng.2747.
Published Version	doi:10.1038/ng.2747
Accessed	February 19, 2015 3:54:02 PM EST
Citable Link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:12152881
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

(Article begins on next page)

NIH Public Access
Author Manuscript
Nat Genet. Author manuscript; available in PMC 2014 April 01.

Published in final edited form as:
Nat Genet. 2013 October ; 45(10): . doi:10.1038/ng.2747.

Genomic Analysis Identifies Targets of Convergent Positive Selection in Drug Resistant Mycobacterium *tuberculosis*

Maha R Farhat[*,1,†], **B Jesse Shapiro**[2,3,4,5,†], **Karen J Kieser**[6], **Razvan Sultana**[7], **Karen R Jacobson**[8,9], **Thomas C Victor**[9], **Robin M Warren**[9], **Elizabeth M Streicher**[9], **Alistair Calver**[10], **Alex Sloutsky**[11], **Devinder Kaur**[11], **Jamie E Posey**[12], **Bonnie Plikaytis**[12], **Marco R Oggioni**[13], **Jennifer L Gardy**[14], **James C Johnston**[15], **Mabel Rodrigues**[16], **Patrick K C Tang**[16], **Midori Kato-Maeda**[17], **Mark L Borowsky**[18,19], **Bhavana Muddukrishna**[18,19], **Barry N Kreiswirth**[20], **Natalia Kurepina**[20], **James Galagan**[21,22,23,2], **Sebastien Gagneux**[24,25], **Bruce Birren**[2], **Eric J Rubin**[6], **Eric S Lander**[2], **Pardis C Sabeti**[2,3,4,6], and **Megan Murray**[*,26,27]

[1]Pulmonary and Critical Care Division, Massachusetts General Hospital, Harvard Medical School, Boston, MA, 02114 [2]The Eli and Edythe L. Broad Institute, Cambridge, MA, 02142 [3]Department of Organismic and Evolutionary Biology, Faculty of Arts and Sciences, Harvard University, Cambridge, MA 02138 [4]Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA 02115 [5]Département de sciences biologiques, Université de Montréal, Montréal, QC, Canada [6]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115 [7]Dana Farber Cancer Institute, Department of Bioinformatics and Computational Biology, Boston, MA, 02115 [8]Section of Infectious Diseases, Boston University School of Medicine, Boston, MA 02118 [9]DST/NRF Centre of Excellence for Biomedical TB Research/MRC Centre for Molecular and Cellular Biology, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg,

[*]Correspondence to: mrfarhat@partners.org, mmurray@hsph.harvard.edu.
[†]M. R. Farhat and B. J. Shapiro contributed equally.

Accession number: All sequences have been rendered publically available through the National Center for Biotechnology Information (NCBI). The Haarlem, C, 98_r168 and w-148 genome assemblies are available under the GenBank accession numbers of AASN00000000, AAKR00000000, ABVM00000000, and ACSX00000000 respectively. The 35 isolate raw sequences from Vancouver are available at the NCBI Sequence Read Archive under accession number SRA020129. The KZN-DS (KZN-4207), KZN_MDR (KZN-1435), KZN_XDR (KZN-605) raw read sequences are available under the number SRA009637. The isolates corresponding to our study identification numbers 51-73 raw sequence data are available at NCBI SRA009341. The rest of our isolate read sequences are available under the number SRA009458 under the project name XDR comparative. The public partially or completely finished genome sequences for MTB210, H37Ra, HN878, R1207, and X122 were accessed under the GenBank accession number of ADAB00000000, AAYK01000000, ADNF01000000, ADNH01000000, and ADNG01000000 respectively.

Authors contributions: This study was designed and conducted by M. R. Farhat and M. B. Murray. M. R. Farhat wrote the first drafts of the paper. B. J. Shapiro, P. C. Sabeti and E. S. Lander provided conceptual input into the evolutionary testing, analysis support, and key manuscript edits. K. J. Kieser and E. J. Rubin constructed the *ponA1* mutants and measured their minimium inhibitory concentrationss. R. Sultana provided bioinformatics support, and K. R. Jacobson helped with curation of the isolate phenotypes. R. M. Warren, E. M. Streicher, T. C. Victor, A. Calver conducted the molecular epidemiologic studies and performed the molecular characterization, drug susceptibility testing (DST) and selection of isolates from South Africa. A. Sloutsky and D. Kaur performed molecular and DST characterization and selected isolates from Peru and Russia. B. Plikaytis and J. E. Posey performed the molecular characterization, DST and selection of isolates from the Centers for Disease Control. M. R. Oggioni identified the patient and performed the molecular characterization and selection of the serial isolates from the patient in Italy. J. L. Gardy, J. C. Johnston, M. Rodrigues and P. K. C. Tang conducted the TB outbreak investigation in British Columbia and performed molecular characterization, drug susceptibility testing (DST), and sequencing of these isolates. M. Kato-Maeda conducted the epidemiologic study of TB transmission in San Francisco and M. L. Borowsky and B. Muddukrishna performed the molecular characterization and sequencing of these isolates. B. N. Kreisworth and N. Kurepina characterized the W-148, Haarlem, and C isolates. S. Gagneux collected the 24 drug sensitive MTB diversity strain set. J. Galagan and B. Birren provided oversight for sequencing and bioinformatics support.

Conflict of Interest: All authors declare no competing interests as defined by Nature Publishing Group or other interests that might be perceived to influence the results and/or discussion reported in this article.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

South Africa [10]Anglogold Ashanti Health West Vaal Hospital, Orkney, North West, South Africa [11]University of Massachusetts Medical School, Massachusetts Supranational TB Reference Laboratory, 305 South St., Boston MA 01230 [12]Division of Tuberculosis Elimination, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA, 30333 [13]Department of Genetics, University of Leicester, Leicester, LE1 7RH, UK [14]Communicable Disease Prevention and Control Services, British Columbia Centre for Disease Control, Vancouver V5Z 4R4, Canada [15]Clinical Prevention Services, British Columbia Centre for Disease Control, Vancouver V5Z 4R4, Canada [16]Mycobacteriology/TB Laboratory, BCCDC Public Health Microbiology and Reference Laboratory, Provincial Health Services Authority Laboratories, Vancouver V5Z 4R4, Canada [17] Division of Pulmonary and Critical Care, University of California, San Francisco, San Francisco CA 94043 [18]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114 [19]Department of Genetics, Harvard Medical School, Harvard University, Boston MA 02115 [20] Public Health Research Institute Tuberculosis Center, Rutgers, The State University of NJ, Newark, NJ 07103 [21]Departments of Biomedical Engineering, Boston University, Boston, MA 02215 [22]Department of Microbiology, Boston University, Boston, MA 02215 [23]Bioinformatics Program, Boston University, Boston, MA 02215 [24]Swiss Tropical and Public Health Institute, 4002 Basel, Switzerland [25]University of Basel, 4002 Basel, Switzerland [26]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115 [27]Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115

Abstract

Mycobacterium tuberculosis is successfully evolving antibiotic resistance, threatening attempts at tuberculosis epidemic control. Mechanisms of resistance, including the genetic changes favored by selection in resistant isolates, are incompletely understood. Using 116 newly and 7 previously sequenced *M. tuberculosis* genomes, we identified genomewide signatures of positive selection specific to the 47 resistant genomes. By searching for convergent evolution, the independent fixation of mutations at the same nucleotide site or gene, we recovered 100% of a set of known resistance markers. We also found evidence of positive selection in an additional 39 genomic regions in resistant isolates. These regions encode pathways of cell wall biosynthesis, transcriptional regulation and DNA repair. Mutations in these regions could directly confer resistance or compensate for fitness costs associated with resistance. Functional genetic analysis of mutations in one gene, *ponA1*, demonstrated an *in vitro* growth advantage in the presence of the drug rifampicin.

The evolution of antibiotic drug resistant bacteria is a major public health concern. To combat antibiotic resistant infections, we must not only develop new drugs, but also learn to use existing drugs more effectively. With some exceptions (*e.g.* in the case of phenotypic drug tolerance), resistance is encoded in the bacterial genome; therefore, resistance-associated mutations, whether they are directly causal of resistance or not, can serve as biomarkers which can be rapidly identified in the clinic by PCR or sequencing-based assays. These molecular biomarkers allow the determination of a bacterial infection's drug resistance profile in a matter of hours, instead of the days or weeks required for culture-based diagnostics. In some cases, this time lag can make the difference between a successful or unsuccessful treatment.

Here, we describe a method to identify biomarkers of drug resistance in a rapid and unbiased manner. It consists of sequencing the genomes of bacteria with different resistance phenotypes, and applying phylogenetic methods and statistical tests for positive selection to identify variants in the genome that are consistently associated with resistance. The method

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

is amenable to different microbes with different phenotypes of interest. Here, we apply it to identify biomarkers of drug resistance in *Mycobacterium tuberculosis* (MTB), the bacterium responsible for tuberculosis (TB).

The evolution and spread of drug resistant TB threaten to undermine the success of TB treatment and control programs worldwide. Multi-drug resistant (MDR) TB is defined as TB that is resistant to isoniazid and rifampicin, the two most effective anti-tubercular drugs. With a global estimate of 650,000 MDR cases in 2010[1] and a rising number of cases that are extensively drug-resistant (XDR; defined as MDR cases that are also resistant to fluoroquinolones and injectable agents), drug-resistant TB poses a major challenge, requiring advances in diagnostics, methods of surveillance, and therapeutics.

Resistance in MTB is thought to arise through the serial acquisition of point mutations in genes encoding drug activating enzymes or targets. Current molecular diagnostics amplify and detect known drug resistance mutations, and their performance depends on the inclusion of a comprehensive catalog of these mutations. Although known mutations explain much resistance in MTB, causative mutations have not been identified in 10-40% of clinically resistant isolates[2] and, even where causative mutations have been identified, there may be additional variants that contribute to drug resistance.

In addition to classical drug-resistance genes (encoding the protein target of the drug or a drug-metabolizing enzyme), mutations in three other classes of genes may confer a selective advantage in the presence of drugs. First, mutations that reduce cell wall permeability or increase the activity of drug efflux pumps are expected to increase the mean inhibitory concentrations of drugs, potentially providing an early step toward full-blown drug resistance[3]. Second, compensatory mutations that ameliorate the fitness costs of other resistance mutations can occur and be selected, as seen in both clinical and experimental evolution studies[4]. Third, mutator phenotypes can increase the rate at which rare beneficial mutations occur (though, at the expense of also accumulating deleterious mutations) and therefore provide a selective advantage in the presence of drug treatment[5].

To identify novel genomic regions associated with drug resistance, we performed next-generation whole genome sequencing of 116 MTB isolates from four categories: (i) eight epidemiologically-linked clusters of cases (epiclusters) with emergent drug resistance, (ii) two uniformly drug-sensitive epiclusters, (iii) 35 non epi-linked isolates sampled to represent the 6 major global lineages of MTB and (iv) 8 isolates from a single patient displaying emergent resistance (Figure 1). We combined these data with publicly available genomes of seven isolates. The full 123 MTB sample set include 47 isolates resistant to at least one TB drug, including nine isolates that are XDR-TB. The resulting dataset captured substantial genetic diversity, with 24,711 polymorphic sites relative to the H37Rv reference genome. A genome-wide phylogeny revealed significant population differentiation between epiclusters, confirmed by high fixation indices ($F_{ST} > 0.36$) between all epicluster pairs (Figure 1B,C).

We first determined whether drug resistance could be explained by previously known resistance mutations. To ensure that no resistance mutations were missed, we performed additional deep targeted sequencing of the known resistance genes in 35 resistant isolates (15 isolates with no apparent resistance mutations and 20 with at least one known resistance mutation). We detected mutations in known resistance determinants that had been missed by the initial whole genome sequencing in two isolates; the remaining 13 had confirmed resistance to at least one drug that could not be explained by known mutations (Supplementary Table 1). We did not miss any mutations in the remaining 20 isolates. The

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

isolates with 'unexplained' resistance likely harbor novel mutations that encode resistance and perhaps contribute more generally to the acquisition and maintenance of resistance.

Next, we reasoned that genomic variants (mutations or alleles) under selection in resistant strains could provide clues about the cellular mechanisms conferring resistance, while also serving as biomarkers of resistance. We therefore sought to identify genes harboring mutations that confer a selective advantage to drug resistant strains. Unfortunately, many commonly used tests to identify genes under positive selection are not well suited to bacteria such as MTB. Haplotype-based tests for positive selection, often used in humans and other eukaryotes, cannot be used as genetic diversity in MTB arises primarily by clonal expansion rather than by mating and homologous recombination among isolates[6,7]. The widely-used dN/dS is also not suited to MTB, as the method has low sensitivity in detecting positive selection in recently diverged sequences from a single species[8], and as strong purifying selection on synonymous mutations in MTB[6] can spuriously give rise to high dN/dS scores, resulting in low specificity. Indeed, dN/dS lacked power and likely specificity when applied to our MTB dataset, recovering only five of 11 known resistance determinants, while detecting 143 additional genes (Supplementary Table 2).

Instead, we sought to leverage evolutionary convergence – the repeated and independent emergence of resistance-associated mutations at specific loci or genes – to develop a test for selection in clonal bacterial species like MTB[7]. To identify independently arising mutations, we reconstructed a phylogenetic tree for the 123 isolates using *M. canetti* as an outgroup. Based on this tree, we inferred nonsynonymous and noncoding ancestral sequence changes and internal resistance states using parsimony. We focused on nonsynonymous mutations as these were more likely to encode functional protein changes than their synonymous counterparts. Nevertheless, due to emerging evidence that synonymous sites may also be under selection for adaptive changes in gene expression or mRNA stability[6,9] we also inferred synonymous mutations (in a secondary analysis reported only in the Supplementary Note).

We took several precautions to ensure that the reconstructed changes and resistance states in our analysis were not influenced by possible errors in the tree topology. First, we reconstructed the phylogeny in triplicate using different methodologies, and removed all mutations not inferred in all three trees. We also ignored ambiguous mutations from the ancestral reconstruction, and mutations occurring at branches with lower than 70% bootstrap support. Second, to remove local uncertainty in the tree topology, we counted 'close' changes only once. 'Close' changes were any changes that occurred in two isolates separated by less than the 98th percentile for within-epicluster genetic distance. Third, we implemented a simplified 'pairwise' convergence test in which we compared the most sensitive to the most resistant isolate in each of 8 epiclusters, ignoring the rest of the phylogeny entirely (Supplementary Figure 1).

Using the high confidence ancestral reconstruction, we designed a phylogenetic convergence test (phyC). We first looked for specific mutations with a higher frequency in the resistant branches compared with the sensitive branches as candidate targets of independent mutation (TIMs). To distinguish convergence due to positive selection in resistant branches from patterns expected by chance under neutral evolution[10], we assessed the significance of each candidate TIM relative to the expectation from the observed mutations across the phylogeny. Briefly, for each TIM with mutations in x resistant branches and y sensitive branches, we redistributed the $x + y$ hits onto the phylogeny, with branches receiving hits at random, in proportion to their length. We repeated this permutation 10,000 times to obtain an empirical P-value assessing the significance of the association of each TIM with resistance. This

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

procedure controls for the tree topology, including the distribution of resistance phenotypes across the tree, and for local mutation rate within each TIM.

We then expanded the PhyC test from individual nucleotide sites to encompass whole genes and intergenic regions as targets. Here, we looked for genes or intergenic regions with a higher frequency of independently arising mutations anywhere along their length, using the same framework as described above for the site-specific test.

We applied the PhyC test, for mutations, genes and intergenic regions, to the genomewide phylogeny of 123 TB isolates. For our analysis, we defined the resistance phenotype as resistance to any anti-TB drug by conventional drug susceptibility testing so that we would be able to identify mutations associated with multiple resistances as well as those that confer resistance to a single drug. We repeated these analyses to identify selection in isolates resistant to each of the five first-line anti-TB drugs (isoniazid, rifampin, pyrazinamide, ethambutol, and streptomycin).

As a proof of concept, we assessed the functional impact of the observed mutations within one of the TIMs identified by the PhyC test. We constructed two *ponA1* mutants (carrying 2 of the 3 SNPs (C123G and G1095T) that were most enriched in resistant strains) in an H37Rv MTB laboratory strain using recombineering and site directed mutagenesis. We then compared the survival of the two mutant strains to wild type cells and those that lacked the *A1* gene in increasing concentrations of the drugs rifampicin, isoniazid, streptomycinand ofloxacin.

Results

Targets of independent mutation

PhyC detected all eleven known resistance determinants as significant TIMs. Nine of these were also identified by a weaker but conservative phylogeny-independent ‘pairwise’ convergence test (Supplementary Tables 3 & 4) developed here. We further identified 39 novel TIMs not previously associated with resistance, consisting of 7 nonsynonymous coding sites, 2 non coding sites, 28 genes and 2 intergenic regions ($p < 0.05$) (Figure 2, Supplementary Table 5). All 9 individual nucleotide site TIMs fell within genes or intergenic regions also identified as TIMs. We observed that mutations in resistant branches cluster more closely in the genome than those in sensitive branches and that many of the TIMs fall in these regions of dense resistant-specific mutations (Supplementary Tables 6 & 7). Mutation 946T in the conserved membrane protein Rv0218 had the highest number of independent hits in resistant branches of any candidate mutation occurring in eight resistant branches and on no sensitive branch ($P < 0.00001$). However, because there were more sensitive than resistant isolates in our dataset, mutations in sensitive branches are far more prevalent than in resistant branches (compare red and blue histograms of Figure 2). Several of the TIMs significantly associated with resistance also occur in sensitive branches (Supplementary Table 5). This suggests that several TIMs may not cause resistance directly, but rather provide an incremental fitness advantage to resistant strains.

Functions of candidate selected loci

Among the 39 novel genomic regions identified by PhyC, 11 have annotated function (Fig. 4A), 16 belong to a family of genes (PE/PPE) of principally unknown function that is unique to mycobacteria and constitutes about 10% of the MTB genome (Supplementary Table 5), and the remaining 12 are of unknown function. We systematically mined the literature on these genes not previously associated with resistance, noting evidence that many closely interact with known DR genes (physically or genetically) or drug efflux pumps, alter

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

intrinsic drug resistance in MTB or non-tuberculous mycobacteria, are involved in DNA repair, replication or recombination or affect cell wall biogenesis.

Previously known resistance loci

Two novel TIMs were located nearby known resistance genes in the genome, suggesting that they modify or compensate for the phenotypes of the known genes. The first occurs in the promoter of the known resistance gene *rrs*, which encodes the 16S RNA component of the ribosome, a target of aminoglycoside drugs. The second, *rpoC* (Figure 3), is in the same operon as *rpoB*, which codes for the β subunit of RNA polymerase, the main target of the drug rifampicin. *RpoC* encodes the interacting β′ subunit, and has been identified as a target of compensatory mutations modifying the fitness of rifampicin-resistant isolates of both MTB and *S. typhimurium*[11-14]. Prior studies have shown that substitutions in *rpoC* are frequent in clinical rifampicin resistant isolates with known *rpoB* mutations and that the relative fitness of *rpoC/rpoB* mutants *in vitro* is higher than those with *rpoB* mutations alone[12-13]. Some *rpoC* mutations in *S. typhimurium* also confer low-level rifampicin resistance, suggesting that rifampin-resistance phenotypes may be the result of additive substitutions in different genes[13]. Among the 43 rifampicin resistant isolates in our collection, 13 (30%) harbored *rpoC* substitutions that did not appear in any rifampicin sensitive isolates.

Unexplained drug resistance

We determined whether nonsynonymous mutations in the TIMs could explain resistance in the 13 isolates without known drug resistance mutations. For each drug and isolate we identified all mutations in the candidate genes excluding mutations occurring in any isolate sensitive to that drug. Although no single candidate mutation or gene was found to account for all the unexplained drug-specific resistance, two of six isolates with unexplained kanamycin resistance harbored mutations in *PPE60* (Supplementary Table 8). Eight of the 13 (62%) isolates exhibited changes in at least one TIM, with four of the isolates exhibiting changes in two or more.

Drug-efflux pumps

Although no efflux pumps were identified among genes that met our statistical criteria for TIMs, we found that several efflux pumps, including the ABC transporters Rv0194 and Rv1463, have a larger number of independent mutations in resistant strains relative to sensitive strain.

DNA repair

Another of the novel TIMs, *dnaQ*, encodes a component of DNA polymerase III that provides “proof-reading” activity during DNA replication. Several *dnaQ* mutations in *E. coli* yield strong mutator phenotypes[15]. To date, no similar phenotype has been described in MTB, although *dnaQ* variants are not uncommon among clinical isolates[16].

The PE/PPE gene family as a target of independent mutation

Sixteen novel TIMs are members of the PE/PPE family, the majority of which are within the PGRS sub-family; this was the only gene group significantly enriched in the convergence analysis. Multiple members of this family are surface-exposed cell wall proteins; some affect cell wall structure and permeability and some have been shown to be antigens[17]. PE/PPE genes contain an extremely high density of substitutions, and were therefore excluded from the genomewide phylogeny. As a result, it is expected (although not guaranteed) that these genes would be enriched for conflicting phylogenetic signal (homoplasy), but not necessarily that homoplasic mutations be associated with resistance. The association of PE/

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

PPE genes with drug resistance is therefore noteworthy. Due to the high genetic diversity in these genes, the association may reflect random fixation of this diversity during population bottlenecks that occur during antibiotic treatment, rather than genuine positive selection on resistant isolates. In other words, resistant isolates may be descended from the survivors of severe bottlenecks, during which neutral mutations repeatedly become fixed, and these mutations are most readily observed in diverse loci like PE/PPE genes. However, we cannot rule out a possible functional role of PE/PPE genes in the evolution of resistance – for example, *PPE60* is one of the best candidates to account for some of the unexplained kanamycin resistant isolates (see below, and Supplementary Table 8).

Cell wall homeostasis

Five novel TIMs contribute to MTB cell wall biogenesis or remodeling. The structure of the mycobacterial cell wall is unique among prokaryotes in that in addition to the peptidoglycan layer typical of most bacteria, it contains several outer layers characterized by unusual complex lipids (Table 1 and Supplementary Figure 2). These layers contribute to the permeability barrier that underlies the intrinsic antibiotic resistance of most mycobacteria[18]. Multiple TB drugs target structures in the cell wall, and many of the known resistance genes code for enzymes in cell wall lipid pathways. Three of the five genes (*ppsA*, *pks12* and *pks3*) participate in the biosynthesis and translocation of the surface exposed lipids including phthiocerol dimycocerosate (PDIM) [19-21] while the remaining two (*murD* and *ponA1*) contribute to the biosynthesis and homeostasis of the cell wall component, peptidoglycan[22]. Deletion of *ppsA*, depletion of *pks12* and depletion of *ponA1* each affect susceptibility to antibiotics in non-tuberculous mycobacteria and/or MTB[23-25]. Deletion of *pks12* has been recently shown to increase drug resistance in *M. avium* through a cell wall remodeling pathway[26]. In addition, *pks12* had a synonymous site that was a significant TIM (Supplementary Table 9)

Functional Analysis of PonA1 Mutants

In the presence of the drug rifampicin, the strain carrying the *ponA1* G1095T mutation had a 4-6 fold survival advantage at a concentration of 0.00125μg/ml of drug (Figure 4). The MIC to rifampicin for this mutant is estimated at 0.0025 μg/ml or 2-fold higher than wildtype MIC (0.00125 μg/ml). In contrast, the *ponA1* C123G mutant strain showed no growth advantage in the presence of rifampicin, and neither mutant demonstrated a growth advantage when grown in the presence of isoniazid, streptomycin or ofloxacin. The 1095 site maps close to the *ponA1* transpeptidase domain catalytic site, raising the possibility that the SNP inactivates this enzymatic activity; this is supported by the finding that the *ponA1* deletion mutant demonstrates a similar rifampicin resistance phenotype (Figure 4).

Discussion

This work describes a comprehensive genomewide search for genes under selection in clinically resistant MTB isolates. Convergent evolution provides the basis for a highly sensitive test for selection that recovered all of a set of 11 known drug resistance determinants, and is easily generalizable to the study of other types of bacteria. Our method involves sequencing the genomes of related bacteria with different phenotypes of interest (in this case, antibiotic resistance), reconstructing a phylogeny, and identifying targets of convergent evolution using a simple statistical test. While the method relies on a genomewide phylogeny, it can accommodate recombination, provided that it is not so rampant as to completely obscure the phylogeny. Recombinant regions often conflict with the genomewide phylogeny, allowing them to be identified by our method as targets of convergent evolution if they are consistently associated with the phenotype of interest, but not if they are randomly distributed among genomes with different phenotypes. The method

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

is therefore amenable to bacteria with a range of recombination rates. It provides a rapid and unbiased means of identifying molecular biomarkers that are predictive of phenotypes of interest.

Applying this method to MTB, we identified 39 genes and intergenic regions newly associated with resistance. While several of these selected mutations occur in genes that are either nearby known DR loci in the genome or are mutator genes in other organisms, a preponderance were associated with cell wall permeability phenotypes. This suggests that stable drug resistance phenotypes may evolve through a complex step-wise process involving cell wall remodeling. Our finding that a *ponA1* mutation (G1095T) identified as a TIM conferred a fitness advantage in the presence of rifampicin is consistent with this model.

The relevance of cell wall remodeling pathways to drug resistance is also highlighted by two recent studies that compared resistant isolates to their drug sensitive precursors. In one, the investigators noted increased levels of PDIM and peptidoglycan precursors and up-regulation of the PDIM biosynthetic operon (including *ppsA*) in rifampicin resistant strains. Interestingly we identified *ppsA* as a TIM in strains resistant to rifampicin (95% of which had nonsynonymous mutations in rpoB, Supplementary Table 10) in addition to other drugs, raising the possibility that rifampin resistance-causing *rpoB* mutations may lead to alterations in cell wall metabolism possibly a result of altered transcription[27]. The second study documented the occurrence of eleven new non-synonymous mutations in serial MTB isolates from three patients who developed increasing levels of resistance during anti-TB therapy. Seven of the mutations occurred in genes involved in cell wall biosynthesis or transport including *fadD32* and Rv1739c, an ABC transporter. Although none of these overlapped with the TIMS identified in this study, the enrichment of mutations in genes associated with cell wall biosynthetic pathways among progressively drug resistant strains is consistent with our findings and further supports the hypothesis that these changes reflect accommodation to drug exposure[28].

The genomic TIMs we have identified here are associated with drug resistance and may represent changes that confer a selective advantage in the presence of drug. In aggregate, they provide promising new targets for molecular diagnostics and development of therapeutics against drug resistance in MTB.

Online Methods

Institutional review board

The study was evaluated by Institutional Review Boards at the Harvard School of Public Health, the Broad Institute of MIT and Harvard, and the Centers for Disease Control and Prevention where it was determined that it met criteria for exemption from human subject review.

Isolate selection

We assembled an archive of sensitive and resistant isolates that capture a range of *Mycobacterium tuberculosis* (MTB) lineages, geographic sources and resistance profiles. Building on our previous molecular epidemiological studies (Supplementary Table 11 and references within) we aimed to include sets of progressively resistant isolates, sampled either from community transmission chains or from individuals with chronic disease. These sets included isolates from 11 such micro-epidemics (‘epiclustered’ isolates), as defined by molecular fingerprinting, chosen to include the most sensitive and most resistant members of the cluster as well as 8 progressively resistant isolates from a single patient over time. To obtain a measure of background evolution restricted to sensitive isolates and avoid mis-

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

identifying highly variable loci as associated with drug resistance, we included 23 geographically diverse drug sensitive isolates and additional isolates from two drug sensitive epiclusters.

To increase the diversity of the sample, we included 11 additional resistant isolates even when a less resistant progenitor was not available. We also included 3 isolates that had been spontaneously evolved *in vitro* and manifested an aminoglycoside resistance phenotype that was unexplained by targeted sequencing of all previously known resistance genes. In all, 116 MTB isolates were selected for sequencing described in detail in Supplementary Table 11A.

We added 7 publicly available MTB genomes to our alignment, including two drug sensitive Beijing lineage isolates, the latter of which were missing from our sampled isolates thus far. Isolates obtained from public sources are detailed in Supplementary Table 11B. There is no established method to determine the power of genomic analyses performed with this sample size, but the strain set studied here is among the largest set of drug resistant MTB strains sequenced to date.

Of the 123 total isolates, 47 were resistant to one or more drugs (Supplementary Table 11). Eighty three isolates belonged to 14 distinct epiclusters, and the rest were isolates with a unique molecular fingerprint. Twelve clusters had one or more resistant isolates. One other publicly available genome from the species *Mycobacterium canetti* was included to serve as a phylogenetic outgroup.

Resistance phenotype

We defined a “broad resistance” phenotype as resistance to any anti-TB drug by conventional drug susceptibility testing. We used this “broad resistance” as our primary phenotype of interest as our goal was to identify genes and mutations associated with resistance to at least one drug, but potentially many. As a secondary, more specific phenotype of interest, we used resistance to each of the 5 first line anti-TB drugs (isoniazid, rifampin, pyrazinamide, ethambutol, and streptomycin). Resistance status to first line anti-TB drugs had much fewer missing data points than resistance to the other drugs (Supplementary Table 11).

Sequencing, Alignment, and SNP calling

DNA was extracted from all isolates using standard methods and sequenced on an Illumina GAIIx instrument using reads of 36bp length or more. Sequence reads were aligned to the reference genome sequence for H37Rv using MAQ[40]. Reads that aligned with more than 3 mismatches in the first 24bp or that aligned to multiple locations were discarded. SNPs were called with a minimum depth of 20X, and consensus quality score of 20. The required maximum mapping quality of reads covering the SNP was set at 50. SNPs within 5bp of an indel (insertion or deletion) or did not have an adjacent consensus quality of 20 were also discarded. Refer to the supplementary note for further details.

F_{ST} and dN/dS analyses

Fixation index (F_{ST}) and dN/dS rates were computed using standard methods detailed in the supplementary note.

Phylogeny construction

The phylogeny was constructed based on the whole MTB genome multiple alignment, as MTB populations are thought to be predominantly clonal, with most of the genome supporting a single consensus phylogeny not impacted significantly by recombination[6]. A

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

superset of SNPs relative to reference strain H37Rv[35] was created across all clinical isolates from the MAQ SNP reports. SNPs occurring in repetitive elements including transposases, PE/PPE/PGRS genes, and phiRV1 members (273 genes, 10% of genome) (genes listed in reference[41]) were excluded to avoid any concern about inaccuracies in the read alignment in those portions of the genome. Furthermore, SNPs in an additional 39 genes previously associated with drug resistance[38] were also removed to exclude the possibility that homoplasy of drug resistance mutations would significantly alter the phylogeny. After applying these filters to the initial set of 24,711 SNPs, the 23,393 remaining SNPs were concatenated and used to construct phylogenetic trees using three methods. Using the PHYLIP dnapars algorithm v3.68[42] we constructed a parsimony phylogenetic tree using default parameters with *M. canettii* as an outgroup root. We constructed a second phylogeny with Bayesian Markov chain Monte Carlo (MCMC) methods as implemented in the package MrBayes v3.2[42] using the GTR model and a stop criterion of standard deviation of split frequencies of <0.05. We constructed a maximum likelihood tree using PhyML v3.0[44] using the GTR model with 8 categories for the gamma model with and without *M. canetti* to determine the location of the root. One hundred bootstrap resamplings were performed for each tree, except for the Bayesian tree where posterior probabilities on the branches was used as a measure of confidence. A phylogeny was also constructed using the full SNP set (without excluding SNPs in repetitive elements or known drug resistance genes), and only minor differences in the terminal branches of the tree were found. We used the trees constructed with exclusion of SNPs in repetitive elements and known drug resistant genes for all subsequent analyses.

Phylogenetic convergence test for selection (PhyC)

The phylogenetic convergence test used sequences from all branches of the phylogeny. Ancestral nucleotide non-synonymous (or intergenic) substitutions were reconstructed along each branch using both parsimony and maximum likelihood criteria using the R v2.14.1 package ape v3.0.1[45]. Each branch was assigned a resistant or sensitive label using parsimony. The reconstruction was performed in triplicate for the three phylogenies (Bayesian, parsimony and maximum likelihood). We excluded all ambiguously reconstructed states (<90% probability for maximum likelihood reconstruction). We considered changes occurring along the terminal and deep branches of the phylogenetic tree, but excluded changes occurring at branches with bootstrap support or posterior probability of <70%. For each nucleotide site in the genome, we counted the number of convergent SNPs (to the same base) in resistant and sensitive branches. Given that some background convergence is expected due to neutral mutation and sequence error even without positive selection[10], we assessed the significance of each convergent SNP compared to the empirical background distribution (Supplementary Figure 3). For a SNP that converges in x resistant and y sensitive branches, we sampled x+y branches from the distribution of all SNPs in all branches genomewide, repeated this 10,000 times, and recorded the proportion of times substitutions were observed ≥x resistant and ≤y sensitive branches. This proportion serves as an empirical *p*-value for an unexpectedly high level of convergence among resistant branches, suggesting the action of selection. To be considered a candidate for positive selection, we required a SNP to have $p < 0.05$ across all phylogenetic and ancestral reconstruction methods.

As multiple different SNPs within the same gene might nevertheless code for similar resistance phenotypes, we expanded the convergence test beyond individual SNPs to include whole genes and intergenic regions. In this method, branches were defined as convergent if they contained a SNP in the same gene or region, even if the SNPs occurred at different nucleotide sites. For each gene and intergenic region in the genome we counted the number of SNPs occurring within the region boundaries, counting at most one SNP for each branch

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and counting SNPs in order of their frequency in the phylogeny. Using the same empirical resampling strategy as for SNP-based convergence, we generated a list of significant region-based convergence among resistant branches. Supplementary Table 5 details the genes found to be under selection using the phylogeny based convergence method. See the supplementary note for details on the pairwise convergence test and the analysis of the density of resistance-specific mutations.

Selection testing by first line drug phenotype

PhyC, and other supplementary tests for selection were performed similarly for resistance to each of the five first line TB drugs: isoniazid, rifampicin, ethambutol, pyrazinamide, and streptomycin (Supplementary Tables 10,12-15). The genes significant by the "broad resistance" phenotype and resistance to isoniazid, rifampicin, and streptomycin are highly similar with few exceptions. This is likely a result of how closely associated resistance to isoniazid, rifampicin, ethambutol, and streptomycin are (Supplementary Table 11, for example 82% of isolates resistant to either isoniazid or rifampin were resistant to both). The number of significant genes for pyrazinamide was significantly lower than the other drugs likely resulting from the larger number of isolates with a missing resistance phenotype to this drug, and a resultant low statistical power.

SNPs detected in genes under selection

Supplementary Table 16 lists all the SNPs seen in the TIMs in resistant isolates. SNPs were called relative to the preceding (ancestral) node for each isolate in the phylogenetic tree. Supplementary Table 17 provides a multiple alignment of the genetic sequence for all SNP sites in the TIMs (these include sites occurring in resistant strains only, sensitive strains only and SNP sites occurring in both types of strains).

Candidate genes variants in isolates with unexplained resistance

We identified nonsynonymous SNPs in the genes under positive selection in isolates with unexplained resistance. We filtered out SNPs in these genes that occurred in any isolates sensitive to each drug. The results are detailed in Supplementary Table 8.

M. tuberculosis mutant generation

Rv0050, encoding *ponA1*, was replaced with a hygromycin resistance cassette using mycobacterial recombineering in the H37Rv host strain. The *ponA1*: :*hyg* replacement was confirmed by PCR and whole genome sequencing. As we sought to identify mutants more likely to be independently causative of resistance we focused on *ponA1* SNP alleles C123G and G1095T as these occurred in mostly drug resistant clinical strains, and the third site (1891) alleles (C/T) were more prevalent in susceptible strains. SNP alleles in *ponA1* were generated by site directed mutagenesis and Sanger sequence confirmed. Wildtype or SNP alleles in p*onA1* were cloned under the control of a constitutive promoter and integrated in the *M. tuberculosis* genome as single copies at the L5 phage integration site.

MIC assays

All strains were grown in Middlebrook 7H9 medium supplemented with 0.25% glycerol, 10% oleic acid-albumin-dextrose-catalase, and 0.05% Tween-80. For MIC calculations, the strains Δ*ponA1*, Δ*ponA1* L5: :$ponA1_{wildtype}$, Δ*ponA1* L5: :$ponA1_{C123G}$, and Δ *ponA1* L5: : $ponA1_{G1095T}$ were grown until mid-log phase (0.5 – 0.8 spectroscopic optic density at 600nm (OD_{600})) and then diluted to a calculated starting OD_{600} of 0.006 and grown ± drug for 6 days at 37°C with shaking. All conditions were done in duplicate. Two sets of duplicate experiments were performed at slightly different drug concentrations. Percent survival is calculated by normalizing the OD_{600} measurements of each strain to its

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

respective untreated control. The MIC was defined as the drug concentration that inhibits growth to 1% or less of the untreated control.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by a Senior Ellison Foundation Award (MM.), a contact from the National Institute of Allergy and Infectious Diseases no. HHSN266200400001C (B.B.), the department of Pulmonary and Critical Care at Massachusetts General Hospital (M.R.F.) a postdoctoral fellowship from the Harvard MIDAS Center for Communicable Disease Dynamics (B.J.S.), and a Packard Foundation Fellowship (P.C.S.); S.G. was supported by the Swiss National Science Foundation (PP0033_119205).We thank the technical staff of the BCCDC PHMRL Mycobacteriology Laboratory in Vancouver, M. Bosman from the National Health Laboratory Service in Cape Town and Lanfranco Fattorini from the Istituto Superiore di Sanita in Rome.

References

1. World Health Organization. Global Tuberculosis Control 2011. WHO Press; Geneva: 2011.
2. Campbell PJ, et al. Molecular detection of mutations associated with first- and second-line drug resistance compared with conventional drug susceptibility testing of Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2011; 55:2032–2041. [PubMed: 21300839]
3. Nikaido H. Prevention of drug access to bacterial targets: permeability barriers and active efflux. Science. 1994; 264:382–388. [PubMed: 8153625]
4. Schrag SJ, Perrot V, Levin BR. Adaptation to the fitness costs of antibiotic resistance in Escherichia coli. Proc Biol Sci. 1997; 264:1287–1291. [PubMed: 9332013]
5. Denamur E, Matic I. Evolution of mutation rates in bacteria. Mol Microbiol. 2006; 60:820–827. [PubMed: 16677295]
6. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. Genome Res. 2012; 22:721–734. [PubMed: 22377718]
7. Shapiro BJ, David LA, Friedman J, Alm EJ. Looking for Darwin's footprints in the microbial world. Trends Microbiol. 2009; 17:196–204. [PubMed: 19375326]
8. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. 2008; 4:e1000304. [PubMed: 19081788]
9. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol. 2013; 30:549–560. [PubMed: 23223712]
10. Rokas A, Carroll SB. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol. 2008; 25:1943–1953. [PubMed: 18583353]
11. Casali N, et al. Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Res. 2012; 22:735–745. [PubMed: 22294518]
12. Comas I, et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nat Genet. 2012; 44:106–110. [PubMed: 22179134]
13. Brandis G, Wrande M, Liljas L, Hughes D. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. Mol Microbiol. 2012; 85:142–151. [PubMed: 22646234]
14. De Vos M, et al. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. Antimicrob Agents Chemother. 2013; 57:827–832. [PubMed: 23208709]
15. Tanabe K, Kondo T, Onodera Y, Furusawa M. A conspicuous adaptability to antibiotics in the Escherichia coli mutator strain, dnaQ49. FEMS Microbiol Lett. 1999; 176:191–196. [PubMed: 10418146]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

16. Dos Vultos T, Mestre O, Tonjum T, Gicquel B. DNA repair in Mycobacterium tuberculosis revisited. FEMS Microbiol Rev. 2009; 33:471–487. [PubMed: 19385996]
17. Soldini S, et al. PPE_MPTR genes are differentially expressed by Mycobacterium tuberculosis in vivo. Tuberc Edinb Scotl. 2011; 91:563–568.
18. Kaur D, Guerin ME, Skovierová H, Brennan PJ, Jackson M. Chapter 2: Biogenesis of the cell wall and other glycoconjugates of Mycobacterium tuberculosis. Adv Appl Microbiol. 2009; 69:23–78. [PubMed: 19729090]
19. Yu J, et al. Both phthiocerol dimycocerosates and phenolic glycolipids are required for virulence of Mycobacterium marinum. Infect Immun. 2012; 80:1381–1389. [PubMed: 22290144]
20. Matsunaga I, et al. Mycobacterium tuberculosis pks12 produces a novel polyketide presented by CD1c to T cells. J Exp Med. 2004; 200:1559–1569. [PubMed: 15611286]
21. Dubey VS, Sirakova TD, Kolattukudy PE. Disruption of msl3 abolishes the synthesis of mycolipanoic and mycolipenic acids required for polyacyltrehalose synthesis in Mycobacterium tuberculosis H37Rv and causes cell aggregation. Mol Microbiol. 2002; 45:1451–1459. [PubMed: 12207710]
22. Hett EC, Chao MC, Rubin EJ. Interaction and modulation of two antagonistic cell wall enzymes of mycobacteria. PLoS Pathog. 2010; 6:e1001020. [PubMed: 20686708]
23. Billman-Jacobe H, Haites RE, Coppel RL. Characterization of a Mycobacterium smegmatis mutant lacking penicillin binding protein 1. Antimicrob Agents Chemother. 1999; 43:3011–3013. [PubMed: 10582900]
24. Philalay JS, Palermo CO, Hauge KA, Rustad TR, Cangelosi GA. Genes required for intrinsic multidrug resistance in Mycobacterium avium. Antimicrob Agents Chemother. 2004; 48:3412–3418. [PubMed: 15328105]
25. Chavadi SS, et al. Inactivation of tesA reduces cell wall lipid production and increases drug susceptibility in mycobacteria. J Biol Chem. 2011; 286:24616–24625. [PubMed: 21592957]
26. Matsunaga I, Meda S, Nakata N, Fujiwara N. The polyketide synthase-associated multidrug tolerance in Mycobacterium intracellulare clinical isolates. Chemotherapy. 2012; 58:341–348. [PubMed: 23171694]
27. Bisson GP, et al. Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, rpoB mutant Mycobacterium tuberculosis. J Bacteriol. 2012; 194:6441–6452. [PubMed: 23002228]
28. Sun G, et al. Dynamic Population Changes in Mycobacterium tuberculosis During Acquisition and Fixation of Drug Resistance in Patients. J Infect Dis. 2012; 206:1724–1733. [PubMed: 22984115]
29. Krzywinski MI, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 200910.1101/gr.092759.109
30. Shigemura K, et al. Presence of a mutation in ponA1 of Neisseria gonorrhoeae in numerous clinical samples resistant to various beta-lactams and other, structurally unrelated, antimicrobials. J Infect Chemother Off J Jpn Soc Chemother. 2005; 11:226–230.
31. Zahrt TC, Deretic V. An essential two-component signal transduction system in Mycobacterium tuberculosis. J Bacteriol. 2000; 182:3832–3838. [PubMed: 10851001]
32. Nguyen HT, Wolff KA, Cartabuke RH, Ogwang S, Nguyen L. A lipoprotein modulates activity of the MtrAB two-component system to provide intrinsic multidrug resistance, cytokinetic control and cell wall homeostasis in Mycobacterium. Mol Microbiol. 2010; 76:348–364. [PubMed: 20233304]
33. Cangelosi GA, et al. The two-component regulatory system mtrAB is required for morphotypic multidrug resistance in Mycobacterium avium. Antimicrob Agents Chemother. 2006; 50:461–468. [PubMed: 16436697]
34. Möker N, et al. Deletion of the genes encoding the MtrA-MtrB two-component system of Corynebacterium glutamicum has a strong influence on cell morphology, antibiotics susceptibility and expression of genes involved in osmoprotection. Mol Microbiol. 2004; 54:420–438. [PubMed: 15469514]
35. Lew JM, Kapopoulou A, Jones LM, Cole ST. TubercuList--10 years after. Tuberc Edinb Scotl. 2011; 91:1–7.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

36. Jiang X, et al. Comparison of the proteome of isoniazid-resistant and -susceptible strains of Mycobacterium tuberculosis. Microb Drug Resist Larchmt N. 2006; 12:231–238.
37. Yang Q, Liu Y, Huang F, He ZG. Physical and functional interaction between D-ribokinase and topoisomerase I has opposite effects on their respective activity in Mycobacterium smegmatis and Mycobacterium tuberculosis. Arch Biochem Biophys. 2011; 512:135–142. [PubMed: 21683681]
38. Sandgren A, et al. Tuberculosis drug resistance mutation database. PLoS Med. 2009; 6:e2. [PubMed: 19209951]
39. Nessar R, Reyrat JM, Murray A, Gicquel B. Genetic analysis of new 16S rRNA mutations conferring aminoglycoside resistance in Mycobacterium abscessus. J Antimicrob Chemother. 2011; 66:1719–1724. [PubMed: 21652621]
40. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851–1858. [PubMed: 18714091]
41. Comas I, et al. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet. 2010; 42:498–503. [PubMed: 20495566]
42. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989; 5:164–166.
43. Ronquist F, et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol. 2012; 61:539–542. [PubMed: 22357727]
44. Guindon S, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59:307–321. [PubMed: 20525638]
45. Popescu AA, Huber KT, Paradis E. ape 3.0: new tools for distance based phylogenetics and evolutionary analysis in R. Bioinforma Oxf Engl. 201210.1093/bioinformatics/bts184

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

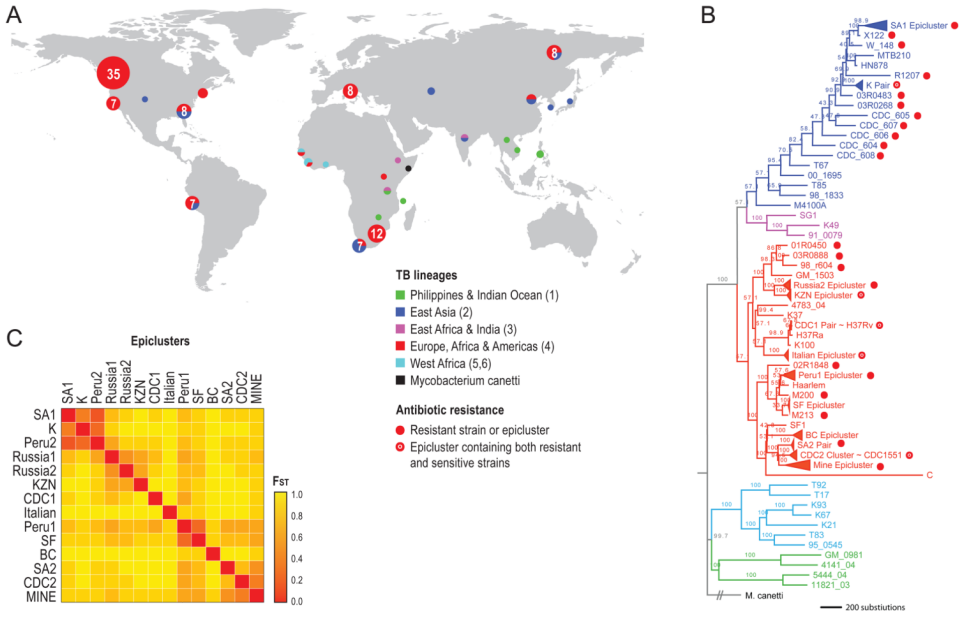

Figure 1.

Characteristics of sequenced TB isolates: (A) Geographic distribution of sampled isolates (circle size is proportional to the number of isolates sampled; circle color refers to TB lineage. (B) Parsimony phylogenetic tree with node bootstrap support. Root length not to scale. Epiclusters are merged into triangles for clarity, with the exception of two paraphyletic epiclusters: Peru2 and Russia1. (C) Genetic differentiation between the 14 epiclusters (higher F_{ST} reflects higher differentiation). F_{ST} values provided in Supplementary Table 19.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

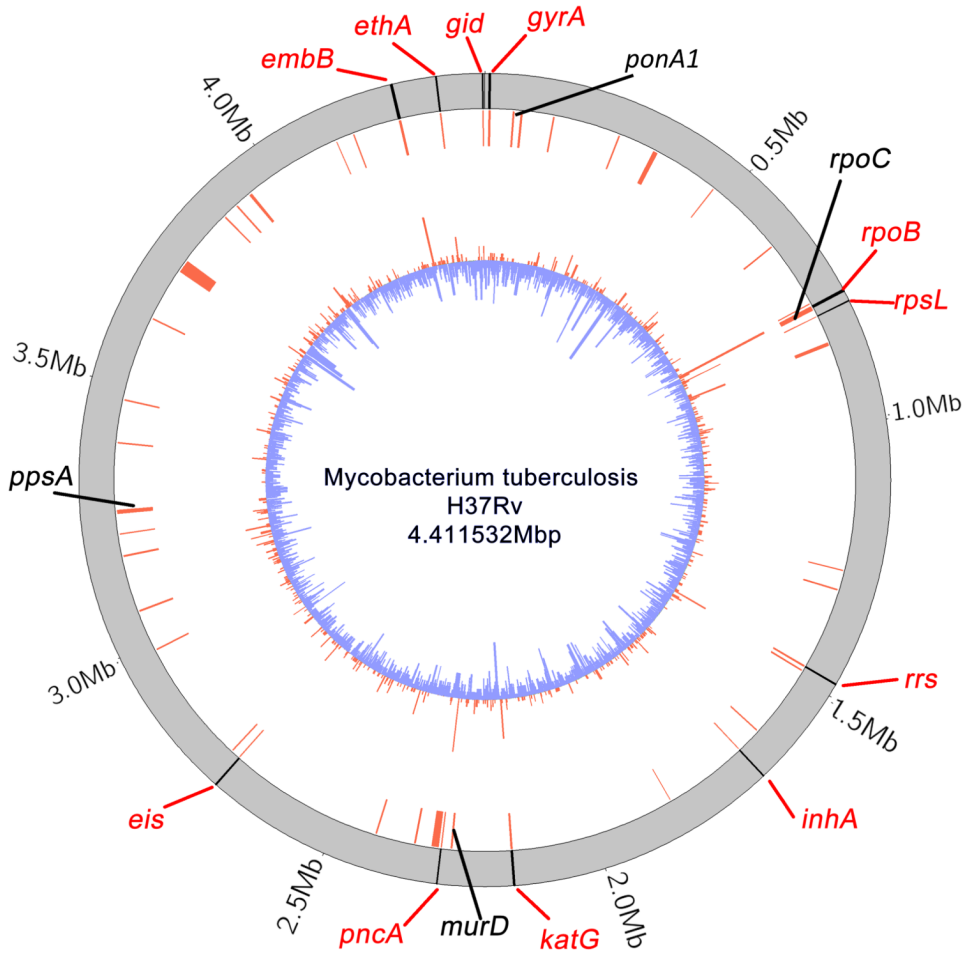

Figure 2.

Candidate genes under selection in resistant MTB. Circular plot of gene locations. Outer black lines represent: the 11 benchmark drug resistance genes in the H37Rv reference genome (red text). Inner red lines represent locations of targets of independent mutation (TIMs). Four novel TIMs of interest are named in black text. The innermost barplot shows the number of mutations per gene or intergenic region, in resistant (red) or sensitive (blue) isolates. Plotted using circos[29].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

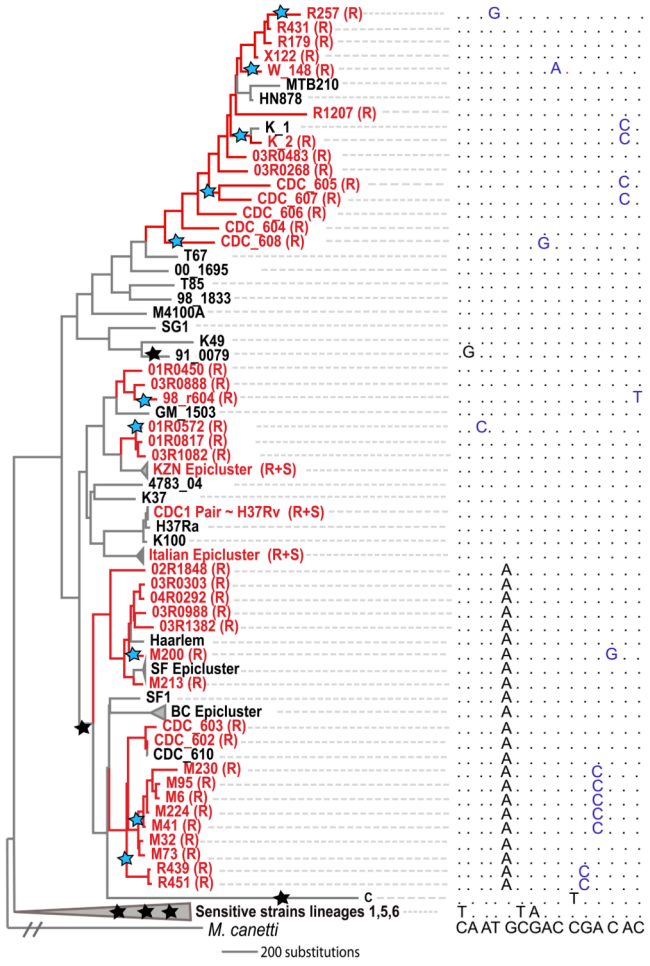

Figure 3.

Evolutionary convergence at the gene level in *rpoC*. Resistant branches (inferred by parsimony, and usually involving progressive resistance to several drugs) and strain names are colored red; sensitive branches in black. Stars on the phylogeny designate inferred sequence changes in *rpoC*: Blue stars denote changes in resistant branches (10 in total), black in sensitive branches (6 in total). Nucleotides in the multiple sequence alignment are also colored blue or black accordingly. Sites shown in the multiple alignment are (left to right) 763884, 764181, 764580, 764819, 765150, 765171, 765230, 765462, 765463, 765482, 765619, 766467, 766488, 766645, 767060 (H37Rv coordinates).

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

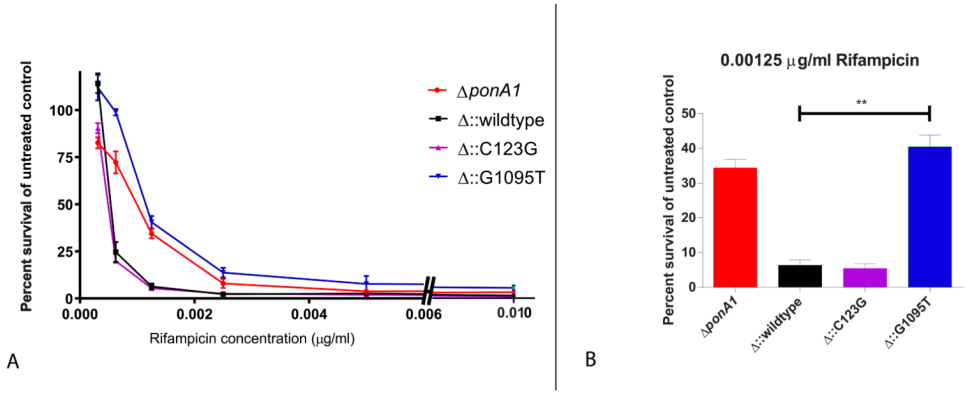

Figure 4.

MTB *ponA1* mutant survival in the presence of the drug rifampicin. (A) Bacterial survival (percent of untreated control OD_{600} absorbance) under increasing concentrations of rifampicin for Δ*ponA1* (*ponA1* deletion mutant), Δ: :wildtype (*ponA1* deletion mutant complemented with the wildtype *ponA1* gene), Δ: :G1095T (mutant complemented with *ponA1* carrying the G1095T allele), and Δ: :C123G (mutant complemented with *ponA1* carrying the C123G allele) (B) Bacterial survival (as in A) of strains cultured in the presence of 0.00125μg/ml of rifampicin. The two-sided t-test comparing survival between wildtype and Δ: :G1095T was significant with a P-value of 0.006. Error bars represent the standard deviation. Four replicate experiments were performed.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Targets of independent mutation of annotated function. Numbers in bold are literature references. Genes involved in cell wall biosynthesis are *ppsA*, *pks3*, *pks12*, *ponA1*, *murD*. Refer to Supplementary Table 5 for a complete list of TIMs.

		Cellular function					Resistance association		
		Synthesis or regulation of surface exposed lipids	**PG homeo-stasis**	**Transcription-regulation**	**DNA replication and repair**	**Glucose metabolism & anti-oxidation**	**Gene or pathway is resistance assoc. in MTB**	**Gene or pathway is resistance assoc. in NTM**	**Resistance assoc. in other bacteria**
ppsA	Rv2931	**19**					**27**	**25**	
pks3	Rv1180	**21**							
pks12	Rv2048c	**20**					**24**	**24,26**	
ponA1	Rv0050		**22**					**23**	**30**
murD	Rv2155c		**22**						
mtrB	Rv3245c			**31**				**32,33**	**34**
rpoC	Rv0668				**12**		**11,12**		**13**
dnaQ	Rv3711c				**35**				**15**
opcA	Rv1446c					**35**	**36**		
rbsK	Rv2436				**37**	**35**			
rrs promoter (pre-Rvnr01)							**38**	**39**	