Signatures of Selection at Drug Resistance Loci in 
*Mycobacterium tuberculosis*

Tatum D. Mortimer,*a,b* Alexandra M. Weber,*a,b* Caitlin S. Pepperell*ab*

*a*Division of Infectious Diseases, Department of Medicine, University of Wisconsin—Madison, Madison, Wisconsin, USA

*b*Department of Medical Microbiology and Immunology, University of Wisconsin—Madison, Madison, Wisconsin, USA

**ABSTRACT**

Tuberculosis (TB) is the leading cause of death by an infectious disease, and global TB control efforts are increasingly threatened by drug resistance in *Mycobacterium tuberculosis*. Unlike most bacteria, where lateral gene transfer is an important mechanism of resistance acquisition, resistant *M. tuberculosis* arises solely by *de novo* chromosomal mutation. Using whole-genome sequencing data from two natural populations of *M. tuberculosis*, we characterized the population genetics of known drug resistance loci using measures of diversity, population differentiation, and convergent evolution. We found resistant subpopulations to be less diverse than susceptible subpopulations, consistent with ongoing transmission of resistant *M. tuberculosis*. A subset of resistance genes ("sloppy targets") were characterized by high diversity and multiple rare variants; we posit that a large genetic target for resistance and relaxation of purifying selection contribute to high diversity at these loci. For "tight targets" of selection, the path to resistance appeared narrower, evidenced by single favored mutations that arose numerous times in the phylogeny and segregated at markedly different frequencies in resistant and susceptible subpopulations. These results suggest that diverse genetic architectures underlie drug resistance in *M. tuberculosis* and that combined approaches are needed to identify causal mutations. Extrapolating from patterns observed for well-characterized genes, we identified novel candidate variants involved in resistance. The approach outlined here can be extended to identify resistance variants for new drugs, to investigate the genetic architecture of resistance, and when phenotypic data are available, to find candidate genetic loci underlying other positively selected traits in clonal bacteria.

**IMPORTANCE**

*Mycobacterium tuberculosis*, the causative agent of tuberculosis (TB), is a significant burden on global health. Antibiotic treatment imposes strong selective pressure on *M. tuberculosis* populations. Identifying the mutations that cause drug resistance in *M. tuberculosis* is important for guiding TB treatment and halting the spread of drug resistance. Whole-genome sequencing (WGS) of *M. tuberculosis* isolates can be used to identify novel mutations mediating drug resistance and to predict resistance patterns faster than traditional methods of drug susceptibility testing. We have used WGS from natural populations of drug-resistant *M. tuberculosis* to characterize effects of selection for advantageous mutations on patterns of diversity at genes involved in drug resistance. The methods developed here can be used to identify novel advantageous mutations, including new resistance loci, in *M. tuberculosis* and other clonal pathogens.

**KEYWORDS**

*Mycobacterium tuberculosis*, antibiotic resistance, evolution, genomics, positive selection
Mycobacterium tuberculosis, the causative agent of tuberculosis (TB), is estimated to have caused 1.4 million deaths in 2015, making it the leading cause of death due to an infectious disease. The proportion of TB due to multidrug-resistant (MDR) M. tuberculosis resistant to the first-line antituberculosis drugs isoniazid (INH) and rifampin (RIF) is increasing (1), which poses a major threat to global public health. Unlike most bacteria, M. tuberculosis evolves clonally, so resistance cannot be transferred among strains or acquired from other species of bacteria: drug resistance in M. tuberculosis results from de novo mutation within patients and transmission of drug-resistant clones (2–4). The relative contributions of de novo emergence and transmitted drug resistance vary across sampling locations (5–9). Another potential variable in the emergence of drug-resistant TB is M. tuberculosis’s lineage structure: seven distinct lineages have been identified among globally extant populations of M. tuberculosis. Of these lineages, lineage 2 (L2) has been associated with relatively high rates of drug resistance, and it has been postulated that the acquisition of resistance is a result of higher rates of mutation in this lineage (10). Studies of M. tuberculosis evolution within hosts with TB have shown that the emergence of drug resistance is associated with clonal replacements that lead to reductions in genetic diversity of the bacterial population (11, 12).

Many of the methods developed to identify advantageous mutations, such as those conferring antibiotic resistance, depend on recombination to differentiate target loci from neutral variants (13). However, in clonal organisms like M. tuberculosis, neutral and deleterious mutations that are linked to advantageous variants will evolve in tandem with them. This linkage among sites can also cause competition between genetic backgrounds with beneficial mutations, decreasing the rate of fixation for beneficial alleles, while deleterious alleles are purged less efficiently (14–16).

While the majority of the M. tuberculosis genome is subject to purifying selection (i.e., selection against deleterious mutations) (3), antibiotic pressure exerts strong selection for advantageous variants that confer resistance. M. tuberculosis drug resistance has been the focus of extensive investigation, and a variety of resistance mutations have been characterized for commonly used antituberculosis drugs (17). Drug resistance mutations can be associated with fitness costs (18–20), and compensatory mutations that ameliorate these fitness costs have been identified in the context of rifampin resistance (21, 22). Resistance mutations found to have lower fitness costs in vitro—as measured by competition assays—are found at higher frequencies among M. tuberculosis clinical isolates and appear to be transmitted at higher rates relative to mutations with high in vitro fitness costs (18, 23). Candidate loci involved in resistance and compensation for its fitness effects have been identified previously by screening for homoplastic variants (i.e., mutations that emerge more than once in the phylogeny) that are significantly associated with drug-resistant phenotypes (24) and genes with higher dN/dS (ratio of nonsynonymous mutations to synonymous mutations) in resistant compared to sensitive isolates (25). Application of these methods to whole-genome sequence data from M. tuberculosis clinical isolates has recovered known drug resistance loci, as well as loci associated with cell surface lipids and biosynthesis, DNA replication, and metabolism.

The goal of the present study was to use patterns of genetic diversity at known drug resistance loci to identify the population genomic signatures of positive selection in natural populations of M. tuberculosis. Using whole-genome sequence data from two populations for which phenotypic resistance data were available, we have identified several distinct signatures associated with these loci under selection. On the basis of these results, we propose methods of identifying loci under positive selection, including novel drug resistance loci, in clonal bacteria such as M. tuberculosis.

RESULTS

We inferred the phylogeny of 1,161 M. tuberculosis isolates from Russia and South Africa (see Materials and Methods and see Table S1 in the supplemental material) using the approximate maximum likelihood method implemented in FastTree 2 (Fig. 1). The majority of the isolates belong to lineage 2 (L2) (n = 667) and L4 (n = 481). M.
tuberculosis nucleotide diversity was similar to previous estimates from a globally distributed sample (26). We identified lineage-specific patterns in overall diversity, with L4 having higher diversity than L2 (average number of pairwise differences of L2 [π\textsubscript{L2}], 3.6 × 10^{-5}, π\textsubscript{L4}, 1.5 × 10^{-4}). Previously published analyses of whole-genome sequence data from L2 indicate that the majority of L2 isolates worldwide belong to a sublineage that has undergone relatively recent expansion (27, 28). In this sample from Russia and South Africa, the majority of L2 isolates belong to this sublineage, while the L4 isolates are associated with deeper branching sublineages. This likely contributes to the observed differences in diversity.

The overall diversity of L2 was lower than that of L4 in our sample (Fig. 2, \( P < 2.2 \times 10^{-16} \)).
A total of 762 of the isolates in our sample (66%) are resistant to one or more antituberculosis drugs (Table 1).

Drug-resistant TB can be acquired as a result of de novo mutations within a patient or by infection with a resistant strain. When resistance is primarily mediated by de novo mutations, diversity should be similar in susceptible and resistant populations, as resistance will arise on multiple genetic backgrounds. In contrast, if resistance develops primarily via transmission of resistant strains, the resistant subpopulation should be less diverse than the susceptible subpopulation. We compared the nucleotide diversity of susceptible and resistant subpopulations and found genome-wide estimates of nucleotide diversity to be higher in isolates susceptible to a range of drugs for which phenotyping data were available (P = 0.029 by paired t-test). In comparisons of gene-wise diversity in susceptible and resistant populations, we found that resistant isolates had a greater number of genes with no diversity, but levels of diversity within genes in which it was measurable were similar between resistant and susceptible populations (Fig. 3).

Of the 3,162 genes included in our analyses, 109 (3%) were invariant across all isolates in our sample. This is likely due to strong purifying selection on these genes. An additional 136 genes harbored variation in the drug-susceptible populations but were invariant across all of the drug-resistant populations. We did not observe the converse, i.e., genes that were invariant in susceptible isolates specifically, which supports the conclusion from genome-wide diversity estimates that resistant isolates represent a subset of the diversity found in susceptible populations and suggests that there may be purifying selection that is specific to the setting of drug resistance. In order to evaluate

| Drug* | Frequency of the following characteristic in the data set: |
|-------|----------------------------------------------------------|
|       | Resistant | Susceptible | Unknown |
| INH   | 0.59      | 0.33        | 0.08    |
| STR   | 0.53      | 0.39        | 0.07    |
| RIF   | 0.50      | 0.43        | 0.07    |
| EMB   | 0.30      | 0.59        | 0.11    |
| PZA   | 0.21      | 0.67        | 0.12    |
| OFL   | 0.16      | 0.39        | 0.45    |
| PRO   | 0.15      | 0.22        | 0.62    |
| CAP   | 0.10      | 0.41        | 0.49    |
| MOX   | 0.09      | 0.41        | 0.49    |
| ETO   | 0.06      | 0.07        | 0.88    |
| AMI   | 0.05      | 0.34        | 0.61    |
| KAN   | 0.05      | 0.12        | 0.83    |

*AMI, amikacin; CAP, capreomycin; EMB, ethambutol; ETO, ethionamide; INH, isoniazid; KAN, kanamycin; MOX, moxifloxacin; OFL, ofloxacin; PRO, prothionamide; PZA, pyrazinamide; Rif, rifampin; STR, streptomycin.

FIG 3 Diversity of resistant and susceptible isolates. (A) Counts of genes with no nucleotide diversity in resistant and susceptible subpopulations. (B) Gene-wise nucleotide diversity (excluding invariant genes) in susceptible and resistant isolates. Among genes in which nucleotide diversity is measurable, it is similar between resistant and susceptible isolates even when drug resistance-associated genes and targets of independent mutation identified by Farhat et al. (24) are removed (P = 0.13). Abbreviations: AMI, amikacin; CAP, capreomycin; EMB, ethambutol; ETO, ethionamide; INH, isoniazid; KAN, kanamycin; MOX, moxifloxacin; OFL, ofloxacin; PRO, prothionamide; PZA, pyrazinamide; Rif, rifampin; STR, streptomycin.
were in the top fifth percentile of diversity (\(s\)ion, a selective sweep, or purifying selection. other genes in the genome. Excess rare variants can result from a population expan-

values). This indicates that these genes contain an excess of rare variants compared to

Tajima’s D values lower than expected based on their length (fifth percentile of residual

ethA are in the top fifth percentile of

target of multiple drug resistance mutations (\(rpsL\), and \(pncA\) were in the top fifth percentile of gene-wise \(\pi\) and/or \(\theta\) values. \(rrs\) and \(ethA\) are in the top fifth percentile of \(\theta\), but not \(\pi\). Surprisingly, despite \(katG\) being a target of multiple drug resistance mutations (Table 2), we did not identify extreme levels of diversity in \(katG\) (80th and 82nd percentiles of \(\pi\) and \(\theta\), respectively).

We also examined gene-wise diversity values within each lineage to look for

lineage-specific high-diversity genes. In both L2 and L4, \(gid\), \(rpsL\), \(pncA\), \(ethA\), and \(thyA\) were in the top fifth percentile of diversity (\(\pi\) and/or \(\theta\)). In L2, \(rpoB\), \(embB\), \(Rv1772\), and \(folC\) were additionally in the top fifth percentile of gene-wise \(\pi\) and/or \(\theta\) values. In L4, \(Rv0340\) was in the top fifth percentile of gene-wise \(\pi\) and/or \(\theta\). While \(rpoB\) and \(embB\) were not in the top 5th percentile of gene-wise \(\theta\) in L4, they still had high diversity (91st and 82nd percentiles, respectively). The lineage-specific differences in diversity of \(Rv1772\), \(folC\), and \(Rv0340\) suggest that there are interactions between these loci and loci that differentiate L2 and L4.

We used gene-wise estimates of Tajima’s D values to investigate gene-specific skews in the site frequency spectrum that could result from selection, where negative values indicate an excess of rare variants and positive values indicate an excess of intermediate frequency variants. We previously identified a relationship between gene length and gene-wise estimates of Tajima’s D values for \(M.\) \(tuberculosis\) (26), and this finding was corroborated here (\(R^2 = 0.3\) after log\(_2\) transformation). In order to identify genes with extreme values of Tajima’s D values—out of proportion with their length—we performed linear regression on log\(_2\)-transformed gene lengths and Tajima’s D values and identified genes with the largest residuals (Fig. 4). \(pncA\), \(ethA\), and \(embC\) all had Tajima’s D values lower than expected based on their length (fifth percentile of residual values). This indicates that these genes contain an excess of rare variants compared to other genes in the genome. Excess rare variants can result from a population expansion, a selective sweep, or purifying selection.

We calculated the ratio of \(\pi\) and \(\theta\) of resistance-associated genes in isolates susceptible and resistant to first-line drugs and identified genes with markedly different diversities in resistant and susceptible subpopulations (Fig. 5A). Among resistance genes in the top fifth percentile of gene-wise \(\pi\) and \(\theta\) overall, diversity of \(pncA\) and \(ethA\) is relatively high among resistant isolates, whereas diversity of \(gid\) is similar in resistant and susceptible populations. We also examined differences in this ratio between isolates in L2 and L4 (Fig. 5B). \(Rv1772\) and \(embR\) were more diverse in resistant isolates in L2, and \(kasA\) and \(thyA\) were more diverse in resistant isolates in L4.

We used fixation index (\(F_{ST}\)) outlier analysis to identify single nucleotide polymorphisms (SNPs) and indels that exhibited extreme differences in frequency between susceptible and resistant populations. Our \(a\) \(p\)riori expectation was that variants me-
diating resistance would be at markedly higher frequency in the drug-resistant sub-
population and that drug targets would be enriched among genes harboring variants 
with high FST. After removing SNPs in regions corresponding to indels and variants at 
sites missing data for greater than 5% of isolates, the highest FST SNPs in comparisons 
of resistant and susceptible subpopulations to first-line drugs are in
\[katG\] (2155168, FST/H11005 0.89, INH), \[rpoB\] (761155, FST/H11005 0.72, RIF), and 
\[rpsL\] (781687, FST/H11005 0.37, streptomycin [STR]). These SNPs were also FST outliers in the lineage-specific analyses.

We used a randomization procedure to assess the significance of observed FST values 
and found the maximum FST values after randomly assigning resistant and susceptible 
designations to be 0.023 for INH, 0.019 for RIF, and 0.018 for STR. In addition to SNPs 
within known drug resistance-associated genes, we identified FST outliers in genes that 
may be novel targets for drug resistance (Table 3).

Homoplastic SNPs—i.e., SNPs that evolve more than once in a phylogeny—are 
candidate loci under positive selection and have previously been used to identify

### Table 2: Signatures of selection in known drug resistance genes

| Gene | Locus tag | Drug(s) | TB Dream\(^a\) | \(\pi\) \(^b\) | \(\theta\) \(^b\) | TD\(^c\) | Homoplasy\(^d\) | FST\(^e\) | Type(s)\(^f\) |
|------|-----------|---------|----------------|--------|--------|------|------------|-------|----------|
| katG | Rv1908c   | INH     | 226            | 0.80   | 0.82   | 0.34 | Y          | Y     | Tight    |
| pncA | Rv2043c   | PZA     | 195            | 0.97   | 1.00   | 0.00 | Y          | N     | Sloppy   |
| embB | Rv3795    | EMB     | 117            | 0.77   | 0.89   | 0.08 | Y          | Y     | Hybrid   |
| ahpC | Rv2428    | INH     | 31             | 0.20   | 0.21   | 0.61 | Y          | N     |          |
| tlyA | Rv1694    | CAP     | 28             | 0.37   | 0.89   | 0.06 | N          | N     |          |
| embC | Rv3793    | EMB     | 28             | 0.59   | 0.74   | 0.01 | N          | N     |          |
| embR | Rv1267c   | EMB     | 25             | 0.46   | 0.49   | 0.28 | N          | N     |          |
| rsr  | Rvm01     | STR, KAN, CAP | 24 | 0.89 | 1.00 | 0.08 | N          | N     |          |
| ethA | Rv3854c   | ETO     | 23             | 0.72   | 1.00   | 0.00 | Y          | Y (IG)| Sloppy, tight (IG) |
| gid  | Rv3919c   | STR     | 22             | 1.00   | 1.00   | 0.07 | N          | Y     | Sloppy   |
| gyrB | Rv0005    | MOX, OFL| 15             | 0.58   | 0.91   | 0.00 | Y          | N     |          |
| fabG1| Rv1483    | INH, ETO| 13             | 0.60   | 0.66   | 0.30 | Y (IG)     | Y     | Tight    |
| inhA | Rv1484    | INH, ETO| 13             | 0.56   | 0.59   | 0.32 | Y          | N     |          |
| rpsL | Rv0682    | STR     | 13             | 0.99   | 0.95   | 0.80 | Y          | Y (IG)| Tight    |
| gyrA | Rv0006    | MOX, OFL| 12             | 0.81   | 0.94   | 0.10 | Y          | Y     |          |
| embA | Rv3794    | EMB     | 11             | 0.77   | 0.38   | 0.79 | N          | N     |          |
| kasA | Rv2245    | INH     | 7              | 0.73   | 0.18   | 0.86 | N          | N     |          |
| ndh  | Rv1854c   |         | 5              | 0.57   | 0.52   | 0.28 | N          | N     |          |
| iniA | Rv0342    | EMB, INH| 4              | 0.64   | 0.33   | 0.56 | N          | N     |          |
| Rv0340| Rv0340    | INH     | 3              | 0.89   | 0.88   | 0.57 | N          | N     |          |
| iniB | Rv0341    | EMB, INH| 3              | 0.07   | 0.07   | 0.79 | N          | N     |          |
| fabP | Rv0129c   | INH     | 3              | 0.78   | 0.19   | 0.89 | N          | N     |          |
| rID  | Rv3266c   | EMB     | 2              | 0.75   | 0.36   | 0.75 | N          | N     |          |
| iniC | Rv0343    | EMB, INH| 2              | 0.49   | 0.67   | 0.08 | N          | N     |          |
| thyA | Rv2764c   | PAS     | 2              | 0.84   | 0.94   | 0.28 | N          | N     |          |
| nat  | Rv3566c   | INH     | 2              | 0.76   | 0.55   | 0.63 | N          | N     |          |
| accD6| Rv2247    | INH     | 1              | 0.90   | 0.63   | 0.90 | N          | N     |          |
| furA | Rv1909c   | INH     | 1              | 0.80   | 0.63   | 0.62 | N          | N     |          |
| Rv1772| Rv1772    | INH     | 1              | 0.50   | 0.35   | 0.54 | N          | N     |          |
| fabD | Rv2243    | INH     | 1              | 0.26   | 0.28   | 0.54 | N          | N     |          |
| fabE24| Rv3139    | INH     | 1              | 0.36   | 0.58   | 0.12 | N          | N     |          |
| rpoB | Rv0667    | RIF     | 1              | 0.82   | 0.92   | 0.18 | Y          | Y     | Hybrid   |
| efpA | Rv2846c   | INH     | 1              | 0.10   | 0.11   | 0.65 | N          | N     |          |
| ethR | Rv3855    | ETO     | 1              | 0.58   | 0.77   | 0.22 | N          | N     |          |
| Rv0678| Rv0678    | BDQ     | 1              | 0.37   | 0.72   | 0.25 | N          | N     |          |
| eis  | Rv2416c   | KAN     | 1              | 0.51   | 0.28   | 0.54 | N          | N     |          |
| mshA | Rv0486    | ETO     | 1              | 0.86   | 0.48   | 0.87 | N          | N     |          |
| rpsA | Rv1630    | PZA     | 1              | 0.88   | 0.62   | 0.84 | N          | N     |          |
| fbc  | Rv2447c   | PAS     | 1              | 0.66   | 0.78   | 0.11 | Y          | N     |          |
| rplC | Rv0701    | LZD     | 1              | 0.57   | 0.77   | 0.21 | N          | N     |          |

\(^a\)The number of distinct entries in the TB Drug Resistance Mutation Database for each gene is reported in the TB Dream column.
\(^b\)\(\pi\) and \(\theta\) are the percentiles for each diversity value, respectively.
\(^c\)TD is the percentile of the residual after linear regression of Tajima's D values with gene length.
\(^d\)Genes with homoplastic SNPs are indicated with "Y" (for Yes) in the Homoplasy column. N, no.
\(^e\)If a homoplastic SNP was also an FST outlier, it is indicated with a "Y" in the FST column. N, no.
\(^f\)Genes are classified as tight, sloppy, or hybrid targets of selection based on diversity, homoplasy, and FST results. (IG) indicates an intergenic SNP.
resistance-associated mutations in \textit{M. tuberculosis} (24). Of the 235 genes with homoplastic SNPs that we identified in our sample, 13 are known to be associated with drug resistance (Fig. 6), and resistance genes were significantly enriched among genes with homoplastic SNPs ($P = 1.2 \times 10^{-4}$ by Fisher’s exact test). \textit{pncA} had the largest number of homoplastic SNPs of any gene in the genome (27 distinct SNPs that appear $\geq 1$ in the phylogeny). The SNPs identified in FST analysis were also identified as homoplastic (Fig. 6). Our results suggest that complementary approaches based on homoplasy and FST outlier analysis can be used to identify SNPs associated with a trait of interest (in this case drug resistance). In addition to genic SNPs, we observed homoplastic SNPs that are also FST outliers in intergenic regions upstream of drug resistance-associated genes (Table 3). These SNPs are candidate resistance and compensatory mutations with a regulatory mechanism of action.

In our analyses of indels, we controlled for the possibility that indels affecting the same gene may not be called in exactly the same position by considering indels within the same gene as identical. We identified four drug resistance-associated genes with homoplastic indels: \textit{gid}, \textit{ethA}, \textit{rpoB}, and \textit{pncA}. FST values for the deletion in \textit{gid} were in the top fifth percentile for populations resistant to capreomycin (CAP), ethambutol (EMB), ethionamide (ETO), kanamycin (KAN), ofloxacin (OFL), and pyrazinamide (PZA), but interestingly, the deletion was not associated with STR resistance ($F_{ST} = 0.04$). Unlike homoplastic SNPs, homoplastic indels were not significantly enriched for drug resistance-associated loci ($P = 1$).

We recovered 20 out of 40 known drug targets by identifying genes with extreme values of diversity, homoplastic SNPs, or SNPs that are FST outliers in comparisons of resistant and susceptible subpopulations. All genes with both extremely high diversity (top fifth percentile) and homoplastic mutations were associated with drug resistance (i.e., \textit{gid}, \textit{ethA}, \textit{pncA}, and \textit{rpsL}). We identified 67 genes with high diversity and Tajima’s D values more negative than expected based on gene length; only two of these were associated with drug resistance (i.e., \textit{ethA} and \textit{pncA}). Twenty out of 51 homoplastic SNPs that are also FST outliers fall within or upstream of known drug resistance-associated
genes. The remaining SNPs may be false-positive results or novel drug resistance mutations.

**DISCUSSION**

Highly virulent bacterial pathogens such as *M. tuberculosis*, *Yersinia pestis* (29), *Francisella tularensis* (30), and *Mycobacterium ulcerans* (31) appear to evolve clonally, i.e., with little to no evidence of lateral gene transfer. It is important to identify advantageous mutations in these and other organisms, as they are likely to be associated with phenotypes such as drug resistance, heightened transmissibility, or host adaptation. However, few methods are available for identifying loci under positive selection in the setting of clonal evolution. We adopted an empirical approach to this problem and used natural population data to characterize patterns of diversity at loci known to be under positive selection in *M. tuberculosis*.

In this analysis of clinical isolates from settings where drug resistance is endemic, we found genome-wide diversity to be higher in susceptible *M. tuberculosis* subpopulations than in the subpopulations resistant to first- and second-line drugs (with the exception of populations resistant to prothionamide [PRO] and moxifloxacin [MOX]). The observation of higher diversity in drug-susceptible populations is consistent with
a significant role for transmitted resistance in the propagation of drug-resistant *M. tuberculosis*. A recent study of extensively drug-resistant (XDR) *M. tuberculosis* infection in South Africa concluded that XDR cases result primarily from transmission of resistance, rather than *de novo* evolution of resistance mutations during infection (9). The primary studies for the sequence data analyzed here also identified clusters of drug-resistant isolates (5, 6), suggesting that resistant isolates were being transmitted. Our results, along with these previously published observations, suggest that the fitness of drug-resistant isolates can be high enough to allow them to circulate in regions where

### Table 3: Homoplastic F<sub>ST</sub> outliers

| Location | Gene<sup>a</sup> | Type | AMI | CAP | EMB | ETO | INH | KAN | MOX | OFL | PRO | PZA | RIF | STR | Known<sup>b</sup> | Lineage |
|----------|-----------------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|---------|
| 1821     | dnaN            | Intergenic | 0.07 | 0.12 | 0.33 | 0.06 | 0.11 | 0.46 | 0.66 | 0.29 | 0.18 | 0.10 | Y   | All  | N      | Y       |
| 7570     | gyrA            | Missense | 0.33 | 0.07 | 0.11 | 0.06 | 0.57 | 0.43 | 0.66 | 0.29 | 0.18 | 0.10 | Y   | All  | Y      | Y       |
| 49388    | hycC            | Synonymous | 0.05 | 0.10 | 0.10 | 0.05 | 0.11 | 0.46 | 0.66 | 0.29 | 0.18 | 0.10 | Y   | All  | Y      | Y       |
| 922004   | Rv0830          | Missense | 0.33 | 0.07 | 0.11 | 0.06 | 0.57 | 0.43 | 0.66 | 0.29 | 0.18 | 0.10 | Y   | All  | Y      | Y       |

<sup>a</sup>For intergenic SNPs, the closest gene is listed.

<sup>b</sup>Weir and Cockerham’s F<sub>ST</sub> (wcF<sub>ST</sub>) values in the top 1% of values genome-wide are reported for each drug.

<sup>c</sup>We identified mutations in genes previously associated with drug resistance (Y for Yes in the Known column) and novel putative resistance or compensatory mutations (N for Novel in the Known column).
drug-resistant *M. tuberculosis* is endemic. As discussed below, the fitness effects of *M. tuberculosis* drug resistance mutations appear to vary substantially; the finding of transmitted resistance in this and other studies suggests that the fitness of isolates harboring low-cost mutations is comparable to that of susceptible *M. tuberculosis*. The populations in our study have a high burden of drug-resistant TB, and the role of transmitted drug resistance may differ in other settings.

An alternative—but not mutually exclusive—explanation for the observation of higher diversity in susceptible populations is that drug-resistant *M. tuberculosis* is under distinct evolutionary constraints that reduce average genome-wide levels of diversity. In support of this hypothesis, we identified a specific subset of genes that were invariant across drug-resistant populations. Interestingly, while average diversity was lower for resistant subpopulations, the gene-wise diversity distributions had heavier tails, indicating there were more genes with extreme levels of diversity.

We found the genetic architecture of resistance to vary among targets, and resistance-associated genes tended to fall within categories that we term “sloppy,” “tight,” and “hybrid” targets of selection (the latter has a combination of tight and sloppy features and applies to *rpsL, embB, and rpoB*). “Sloppy” resistance genes are characterized by high levels of diversity. Genes associated with PZA, EMB, ETO, and STR resistance (i.e., *pncA, gid, rpsL, rrs*, and *ethA*) have high levels of diversity; some also had an excess of rare variants (*pncA, ethA*, and *embC*). The finding that these genes accumulate multiple, individually rare mutations implies that there is a large target for resistance and/or compensatory mutations within the gene: that is, resistance can result from multiple different variants acting individually or in concert.

In addition to its numerous rare mutations, *pncA* also contains the highest number of homoplasmic SNPs (27 SNPs emerged more than once in the phylogeny) of any gene in the data set. Among the 62 nonsynonymous *pncA* mutations in our data set, 55 have been previously reported in association with drug resistance (TB Drug Resistance Mutation Database [32]). The newly described SNPs may mediate drug resistance or compensation for the fitness effects of other variants. Relaxed purifying selection is likely to play a role in concert with selection for diverse advantageous resistance mutations in the accumulation of diversity in *pncA* and other sloppy targets. The fact that numerous mutations are segregating in a natural population suggests that alterations to these genes are generally associated with negligible fitness costs. An *M. tuberculosis* strain harboring a deletion in *pncA* conferring resistance to PZA was estimated to be endemic in Quebec, Canada, by 1800, long before the use of PZA for

![FIG 6 Homoplasmic SNPs in drug resistance-associated genes. SNPs with $F_{ST}$ in the top 1% of genome-wide values are labeled with the population (associated drug resistance) and the $F_{ST}$ value. *pncA* is remarkable for harboring diverse homoplasmic mutations, each of which occurs relatively infrequently ("sloppy target"). *embB, gyrA, katG, rpoB, and rpsL* harbor dominant mutations that occur frequently in the phylogeny and are strongly associated with resistant populations ("tight targets").](msystems.asm.org)
the treatment of TB (33–35). This supports the idea that purifying selection on pncA is relatively weak, which would contribute to its exceedingly high diversity and broaden the adaptive paths to resistance.

In contrast to pncA, gid, which is associated with low-level STR resistance (36), does not appear to have the signatures of a “sloppy” target for resistance despite its high diversity. We identified just three homoplastic SNPs within gid, and previous studies have found that STR-resistant isolates do not carry the same gid mutations (37). This could indicate that a multitude of mutations within gid confer resistance, but levels of diversity in the gene were similar in resistant and susceptible isolates. Previous studies of sequence polymorphism in gid have identified high diversity in this gene in both resistant and susceptible isolates (37–39): gid appears to be subject to relaxed purifying selection in the presence and absence of antibiotic pressure. Since gid mutations confer low-level resistance, it is also possible that misclassification of resistance phenotypes contributed to the lack of differentiation we and others have observed between putatively STR-resistant and -susceptible subpopulations. In addition, mutations in rpsL, which cause high-level resistance, could mask the contribution of gid to STR resistance.

We found some drug targets to be highly diverse in resistant subpopulations of either L2 or L4 (but not both), suggesting that resistance mutations in these genes interact with the genetic background; the fitness effects of mutations in these genes could, for example, vary on different genetic backgrounds. Lineage-specific FST outliers are another category of candidate locus with lineage-dependent roles in drug resistance (Table 3). Epistatic interactions between drug resistance mutations and M. tuberculosis lineage have been reported previously: for example, specific mutations in the inhA promoter have been associated with the L1 and M. africanum genetic backgrounds (40, 41).

In contrast to “sloppy” targets, we discovered individual homoplastic SNPs associated with drug-resistant subpopulations (i.e., with high FST) representing “tight” targets of selection in genes conferring resistance to INH, RIF, and STR. Numerous resistance mutations have been reported for katG, rpoB, rpsL, embB, and gyrA, but we find drug-resistant subpopulations to be defined by a specific subset of mutations in these genes. This suggests that certain mutations are strongly favored relative to others conferring resistance to the same drugs when M. tuberculosis is in its natural environment. Antibiotic resistance can impose fitness costs on M. tuberculosis during in vitro growth, with the range of fitness costs differing among mutations, even within the same gene (18). Mutations can also have different fitness effects depending on the genetic background, but the most fit mutants were the same across M. tuberculosis lineages in a study of RIF resistance (18).

In our analyses, we found the dominant INH resistance mutation in katG to affect the serine at position 315. This change reduces affinity to INH but preserves catalase activity (42) and is associated with lower fitness costs than other katG mutations, both in vitro and in a mouse model (43, 44). This mutation was recently shown to precede mutations conferring resistance to other drugs during accumulation of resistance in the evolution of multidrug-resistant M. tuberculosis (45). The dominant mutations we identified in rpoB (codon 450) and rpsL (codon 43) have also been found to have lower fitness costs in vitro compared to other mutations conferring resistance to RIF and STR in these genes (18, 44, 46). These results suggest that many of the findings regarding the relative fitness costs of M. tuberculosis resistance mutations in vitro and in animal models are relevant to the pathogen’s natural environment.

The fitness effects of mutations in gyrA (codon 94) and embB (codon 306) have not been measured; on the basis of our homoplasy and FST results, we propose that they have lower fitness costs than other mutations in these genes and that they represent “tight” targets of selection. Mutations at gyrA codon 94 were previously found to be the most prevalent in a survey of gyrA and gyrB mutations in fluoroquinolone-resistant M. tuberculosis clinical isolates (47). Interestingly, the mutation in embB codon 306 has been previously associated with acquisition of multiple resistances (48), and we find that this position is an FST outlier for all first-line drugs in L4. This mutation is not an FST
outlier in L2 (i.e., top fifth percentile), with the percentiles for $F_{ST}$ values ranging from 0.07 to 0.68 for first-line drugs in this lineage. These observations suggest that the genetic background affects interactions among resistance mutations and that $embB$ codon 306 is important for acquisition of multidrug resistance in L4, but not in L2.

We searched for indels with the signature of a “tight” target, i.e., homoplastic mutations segregating at markedly different frequencies in drug-susceptible and -resistant subpopulations. Unlike the pattern observed with SNPs, genes associated with drug resistance were not significantly enriched among the genes harboring homoplastic indels. We identified one homoplastic indel that was also an $F_{ST}$ outlier—a deletion in $gid$ that causes a frameshift. Patterns of variation in $gid$ are complex and suggest a role for relaxation of purifying selection (i.e., in the accumulation of excess SNPs in both resistant and susceptible isolates) and perhaps a tight target associated with multiresistance (i.e., this homoplastic/$F_{ST}$ outlier deletion that was associated with resistance to CAP, EMB, ETO, KAN, OFL, and PZA).

Our finding that, save for the frameshift mutation in $gid$, indels in resistance genes do not have the signature of “tight” targets suggests that they are generally associated with higher fitness costs than SNPs. Fifteen drug targets have been found in transposon mutagenesis experiments to be essential for $M. tuberculosis$ growth in vitro, including $rpoB$ and $rpsL$; deletions in these genes are likely to interrupt important functions (49). Deletions in nonessential genes could also have fitness costs. Deletions in $katG$, which is nonessential, can result in INH resistance, but they are not observed as frequently in clinical isolates as the KatG S315 SNP is, particularly among transmitted INH-resistant strains (23).

There are several limitations to our study. Resistance to multiple drugs was common in our sample, and in some cases, it was difficult to identify patterns of diversity and population differentiation that were specific to individual drugs. Our results are also limited by the accuracy with which drug resistance phenotypes were determined and a lack of phenotypic data for some drugs (particularly second-line drugs). Our sample was heavily skewed to lineages 2 and 4, and the results are not necessarily applicable to other $M. tuberculosis$ lineages. Finally, the data analyzed here were generated with short sequencing read technologies, and we were thus limited to characterizing diversity in regions of the $M. tuberculosis$ genome that can be resolved with these methods: regions that were masked from analysis (e.g., due to sequence repeats) may include unknown resistance targets. We also used an L4 genome (H37Rv) as a reference, and gene content specific to L2 may not have been identified.

We were not able to recover all drug resistance-associated genes using the analyses performed here. This is likely a result of limited phenotypic data for some drugs and their associated targets (e.g., $thyA$ and $folC$, which are associated with aminosalicylic acid [PAS] resistance). Our list of drug targets was dominated by genes associated with INH resistance, and signatures in genes that harbor rare resistance-associated alleles may be subtle compared to the KatG S315 mutation found at high frequency in drug-resistant populations.

We identified 31 SNPs that are not on the list of known drug resistance genes, which both emerged more than once in the phylogeny (homoplasies) and were segregating at markedly different frequencies in resistant and susceptible subpopulations ($F_{ST}$ outliers). These SNPs may be novel resistance determinants; notably, all nonsynonymous SNPs within this group are in genes linked with drug resistance in other studies (i.e., they are in genes encoding efflux pumps, genes differentially regulated in resistant isolates or in response to the presence of drug, potential drug targets, or genes in the same pathways as drug targets or resistance determinants) (50–54). In addition to a direct, previously unrecognized role in resistance, these SNPs could compensate for fitness costs of drug resistance. For example, we identified a homoplastic $F_{ST}$ outlier in $rpoC$, and mutations in $rpoC$ have been shown to compensate for RIF resistance in experimental evolution studies (22).

Intriguingly, we found lipid metabolism genes to be enriched among genes harboring homoplastic SNPs ($P = 0.013$). We have previously shown that these genes have
Selection at Drug Resistance Loci in M. tuberculosis

extreme values of diversity in a global sample of M. tuberculosis isolates and within individual hosts (26), suggesting that lipid metabolism genes may also be under positive selection in M. tuberculosis populations. The results presented here could be extended by phenotypic characterization of lipid profiles and identification of homoplastic variants that are at markedly different frequencies in isolates with distinct lipid profiles.

Here we have used drug resistance loci in M. tuberculosis to identify the signatures of positive selection in a clonal bacterium. We found these loci to be associated with distinct patterns of diversity that likely reflect differing genetic architectures underlying the traits under selection. The evolutionary path to resistance is broad for some drugs with “sloppy targets,” whereas for drugs with “tight targets,” the means of acquiring resistance appear more limited. This is likely due to fitness effects of resistance mutations in M. tuberculosis’s natural environment, as numerous resistance mutations have been identified in tight target genes. We also found evidence suggesting that there are important interactions among loci during the evolution of resistance. Our results suggest that purifying selection on a subset of genes intensifies in the setting of resistance, which could reflect epistatic interactions and/or a response to the metabolic milieu imposed by antimycobacterial agents. The results presented here can be used to create more realistic models of resistance evolution in M. tuberculosis and to develop novel strategies of preventing or mitigating the acquisition of resistance. For example, the narrow path to resistance for drugs with tight targets reveals potentially exploitable vulnerabilities, as does the finding of interdependencies among specific loci and the genetic background in the evolution of resistance and multiresistance. As new TB drugs become available for clinical use, the approach outlined here can be extended to investigate their architectures of resistance.

Efforts are under way to sequence and perform drug susceptibility testing on thousands of M. tuberculosis isolates with the goal of creating an exhaustive catalogue of drug resistance mutations and eventually using whole-genome sequencing (WGS) to diagnose drug resistance in clinical settings (CRyPTIC project, http://modmedmicro.nsms.ox.ac.uk/cryptic/, last accessed 24 May 2017). We found that loci under positive selection can be identified using relatively simple methods: “tight” targets are highly differentiated in their allele frequencies across phenotypic groups (i.e., FST outliers) and appear as homoplasies on the phylogeny; “sloppy” targets are characterized by high diversity and/or low Tajima’s D values, as well as homoplasies. Extrapolating from patterns observed among known resistance variants, we have discovered new candidate regulatory and genic resistance variants. The methods used in this study are widely available and should scale to analysis of the large collections of genomic and phenotypic data that are currently being generated. This approach can be extended to identify novel resistance loci in bacteria for which drug susceptibility phenotypes are defined, as well as other positively selected loci in clonal bacterial populations.

MATERIALS AND METHODS

Reference guided assembly. We downloaded sequencing read data from two large surveys of drug-resistant M. tuberculosis in Russia (5) and South Africa (6). We used FastQC (55) and TrimGalore! (56) for quality assessment and adapter trimming of the reads. Trimmed reads were mapped to M. tuberculosis H37Rv (accession no. NC_000962.3) using BWA-MEM v 0.7.12 (57). We used Samtools v 1.2 (58) and Picard Tools (https://broadinstitute.github.io/picard/) for sorting, format conversion, and addition of read group information. Variants were identified using Pilon v 1.16 (59). A detailed description of the reference guided assembly pipeline is available at https://github.com/pepperell-lab/RGBAPepPipe. We removed isolates with mean coverage less than 20×, isolates with percentage of the genome covered at 10× less than 90%, isolates where a majority of reads did not map to H37Rv, and isolates where greater than 10% of sites were unknown after mapping. The final data set contains 1,161 M. tuberculosis isolates (see Table S1 in the supplemental material). The alignment was masked to remove repetitive regions, including proline-glutamate (PE)/proline-proline-glutamate (PPE) genes.

Phylogenetic analysis. We estimated the approximately maximum likelihood phylogeny using the masked alignment from reference guided assembly with FastTree 2.1.9 (60). We compiled FastTree using the double precision option to accurately estimate branch lengths of closely related isolates. We used FigTree (http://tree.bio.ed.ac.uk/software/figtree/) for tree visualization.

SNP annotation. A VCF (variant call format) file of single nucleotide variants was created from the masked alignment using SNP-sites v 2.3.2 (61). Single nucleotide polymorphisms (SNPs) were annotated
using SnEff v 4.1j (62) to identify synonymous, nonsynonymous, and intergenic SNPs based on the annotation of M. tuberculosis H37Rv.

**Indel identification.** Insertions and deletions were identified during variant calling with Pilon. We used Emu (63) to normalize indels across multiple isolates. We used a presence/absence matrix for the normalized indels for further analyses of indel diversity.

**Population genetics statistics.** Whole-genome and gene-wise diversity (\(\pi\) and \(\theta\)) and neutrality (Tajima’s D) statistics were calculated using Egglib v 2.1.10 (64) for whole-genome alignments and gene-wise alignments. Isolates were further divided by lineage and drug resistance phenotype. Sites with missing data due to indels or low-quality base calls in more than 5% of isolates in the alignment were not included in calculation of statistics. Tajima’s D values showed a correlation with gene length in our sample. To find genes with extreme values of Tajima’s D, we performed linear regression in R (65) on log-transformed Tajima’s D values and gene length and identified genes with large residual values. To identify alleles with marked differences in frequency in resistant and susceptible isolates, Weir and Cockerham’s \(F_{ST}\) (66) was calculated using populations of resistant and susceptible isolates for each drug using vcflib v1.0.0-rc0-262-g50a3 (https://github.com/vcflib/vcflib). For nonbiallelic SNPs, we calculated \(F_{ST}\) for the two most common variants.

**Homoplasy.** We used TreeTime (67) to perform ancestral reconstruction and place SNPs and indels in the phylogeny. We identified homoplastic SNPs and indels as those arising multiple times in the phylogeny.

**Data availability.** Unless otherwise noted, all data and scripts associated with this study are available at https://github.com/pepperell-lab/mtbDrugResistance.

**SUPPLEMENTAL MATERIAL**
Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00108-17.

**TABLE S1**, TXT file, 0.02 MB.

**ACKNOWLEDGMENTS**

We thank members of the Pepperell laboratory for their input on analyses and data visualization.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under grant DGE-1256259 to T.D.M. T.D.M. is also supported by National Institutes of Health National Research Service award (T32 GM07215). C.S.P. is supported by National Institutes of Health (R01AI113287). Funding for this project was provided by the University of Wisconsin School of Medicine and Public Health from the Wisconsin Partnership Program.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or other funding agencies.

**REFERENCES**

1. World Health Organization. 2016. Global tuberculosis report 2016. World Health Organization, Geneva, Switzerland.

2. Supply P, Warren RM, Baruks A-L, Lesjean S, Van Der Spuy GD, Lewis L-A, Tibayrenc M, Van Helden PD, Locht C. 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area. Mol Microbiol 47: 529–538. https://doi.org/10.1046/j.1365-2958.2003.03315.x.

3. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, Birren BW, Bishai WR. 2015. Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of Mycobacterium tuberculosis isolates from KwaZulu-Natal. PLoS Med 12:e1001880. https://doi.org/10.1371/journal.pmed.1001880.}

4. Eldholm V, Balloux F. 2016. Antimicrobial resistance in Mycobacterium tuberculosis: the odd one out. Trends Microbiol 24:637–648. https://doi.org/10.1016/j.tim.2016.03.007.

5. Casali N, Nikolayevskyy V, Balabanova Y, Harris SR, Ignatyeva O, Kontsevaya I, Corander J, Bryant J, Parkhill J, Nejentsev S, Horstmann RD, Brown T, Drobniowski F. 2014. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet 46:279–286. https://doi.org/10.1038/ng.2878.

6. Cohen KA, Abeel T, Manson McGuire A, Desjardins CA, Munsamy V, Shea TP, Walker BJ, Sambutani N, Almeida DV, Alvarado L, Chapman SB, Mvelase NR, Duffy EF, 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 SK, Wortman J, Miliana KP, O’Donnell MR, Birren BW, Bishai WR, Pym AS, Earl AM. 2015. Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of Mycobacterium tuberculosis isolates from KwaZulu-Natal. PLoS Med 12:e1001880. https://doi.org/10.1371/journal.pmed.1001880.

7. Coscolla M, Barry PM, Oeltmann JE, Koshinsky H, Shaw T, Cilnis M, Posey J, Rose J, Weber T, Fofanov VY, Gagneux S, Kato-Maeda M, Metcalfe JZ. 2015. Genomic epidemiology of multidrug-resistant Mycobacterium tuberculosis during transmission on a global scale. MBio 6:e00659-15. https://doi.org/10.1128/mBio.00659-15.

8. Guerra-Assunção JA, Crampin AC, Houben R, Mzembe T, Mallard K, Coll F, Khan P, Band L, Chiwava A, Pereira RPA, McNerney R, Fine PEM, Parkhill J, Clark TG, Glynn JR. 2015. Large-scale whole genome sequencing of Mycobacterium tuberculosis provides insights into transmission in a high prevalence area. eLife 4:e04525. https://doi.org/10.7554/eLife.04525.

9. Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, Miliana K, Aliana S, Campbell A, Mthiyane T, Morris N, Mphangase P, van der Meulen H, Omar SV, Brown TS, Narechania A, Shaskina E, Kapwata T, Kreiswirth B, Gandhi NR. 2017. Transmission of extensively drug-resistant tuberculosis in South Africa. N Engl J Med 376:243–253. https://doi.org/10.1056/NEJMoa1604544.

10. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipshitz M, Fortune SM. 2013. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial dif-
ferences in the emergence of drug-resistant tuberculosis. Nat Genet 45:784–790. https://doi.org/10.1038/ng.2656.

11. Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, Caugant DA, Mannsäker T, Mengshoel AT, Dyhrol-Rise AM, Baloux F. 2014. Evolution of extensively drug-resistant Mycobacterium tuberculosis from a susceptible ancestor in a single patient. Genome Biol 15:490. https://doi.org/10.1186/s12864-015-0677-2.

12. Black PA, de Vos M, Louw GE, van der Merwe RG, Dippenaar A, Streicher EM, Pain A. 2015. Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in Mycobacterium tuberculosis isolates. BMC Genomics 16:857. https://doi.org/10.1186/s12866-015-0409-3.

13. Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annu Rev Genet 47:97–120. https://doi.org/10.1146/annurev-genet-111212-135326.

14. Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet Res 8:269–294. https://doi.org/10.1017/S0016632000010156.

15. Felsenstein J. 1974. The evolutionary advantage of recombination. Genetics 78:737–756.

16. Birky CW, Walsh JB. 1988. Effects on linkage on rates of molecular evolution. Proc Natl Acad Sci USA 85:6414–6418. https://doi.org/10.1073/pnas.85.17.6414.

17. Annedi da Silva PE, Palomino JC. 2011. Molecular basis and mechanisms of drug resistance in Mycobacterium tuberculosis: classical and new drugs. J Antimicrob Chemother 66:1417–1430. https://doi.org/10.1093/jac/dkr173.

18. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJM. 2006. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. Science 312:1944–1946. https://doi.org/10.1126/science.1124410.

19. Böttger EC, Springer B. 2008. Tuberculosis: drug resistance, fitness, and strategies for global control. Eur J Microbiol 167:141–148. https://doi.org/10.1186/s12866-015-0106-5.

20. Strauss OJ, Warren RM, Jordaen A, Streicher EM, Sampson SL, Victor TC, Dolby T, Simpson JA, van Heldens PD, Warren RM, Pain A. 2015. Whole-genome sequencing reveals genomic heterogeneity and antibiotic purification in Mycobacterium tuberculosis isolates. BMC Genomics 16:857. https://doi.org/10.1186/s12866-015-0677-2.

21. Brandis G, Wrande M, Liljas L, Hughes D. 2012. Fitness-compensatory mutations in rifampicin-resistant RNA polymerase. Mol Microbiol 85:2514–2516. https://doi.org/10.1111/j.1365-2958.2012.07787.x.

22. Brassard P, DeRiemer K, Enciso A, Muñoz S, Hopewell PC, Small PM, Pym AS. 2006. Impact of bacterial genetics on the transmissibility of isoniazid-resistant drug-resistant tuberculosis. PLoS Pathog 2:e61. https://doi.org/10.1371/journal.ppat.0020061.

23. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis DB, Coultoni MR, Gordy JL, Johnston JC, Rodriguez M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kurepina N, Keim P, Johansson A. 2009. Molecular evolutionary consequences of drug resistance in Mycobacterium tuberculosis family in a low-incidence setting. J Clin Microbiol 47:3101–3104. https://doi.org/10.1128/JCM.00168-11.

24. Gagneux S, Burgos MV, Galagan J, Ng, K, Menzies D, Behr MA. 2003. Widespread pyrazinamide-resistant Mycobacterium tuberculosis strains from a high-incidence setting. J Clin Microbiol 41:1266–1270. https://doi.org/10.1128/JCM.41.4.1266-1270.2003.

25. Spies FS, Ribeiro AW, Ramos DF, Ribeiro MO, Martin A, Palomino JC, Rossetti MLR, da Silva PEA, Zaka H. 2011. Streptomycin resistance and lineage-specific polymorphisms in Mycobacterium tuberculosis gIDB gene. J Clin Microbiol 49:2625–2630. https://doi.org/10.1128/JCM.00168-11.

26. Feuerriegel S, Oberhauser B, George AG, Dafae F, Richter E, Rüscher-Gerdes S, Niemann S. 2012. Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting. BMC Microbiol 12:90. https://doi.org/10.1186/1471-2180-12-90.

27. Jagielski T, Ignatowska H, Bakuła Z, Dziewit Ł, Napiórkowska A, Spies FS, Börger AW, Łukaszuk A, Płoski A, Czyżewska B, Cwyka R, Krupa M, Pfefferke K, Kurepina N, Chybowski T, Rysicki R, Gudowska E, Chauty A, Landier J, Fontanet A, Leirs H, Stinear T, da Silva PEA, Zaka H. 2011. Streptomycin resistance and lineage-specific polymorphisms in Mycobacterium tuberculosis gIDB gene. J Clin Microbiol 49:2625–2630. https://doi.org/10.1128/JCM.00168-11.

28. Ho, S, Meyer CG, Illehemann D, Owusu-Dabo E, Adjei O, Horstmann RD, Browne ENL, Chinhua B, Osei I, Gyapong J, Kurbica T, Ruesch-Gerdes S, Niemann S. 2010. Unequal distribution of resistance-conferring mutations among Mycobacterium tuberculosis and Mycobacterium africanum strains from Ghana. Int J Med Microbiol 300:489–495. https://doi.org/10.1016/j.ijmm.2010.04.019.

29. Fenniri F, Egger M, Bodmer T, Alpter T, Ewahlein M, Jaton K, Pfefferke G, Borrell S, Dubuis O, Bruderer T, Siegrist HH, Furrer H, Calmy A, Fehr J, Stalder JM, Ninet B, Böttger EC, Gagneux S, Swiss HIV Cohort Study, Swiss Molecular Epidemiology of Tuberculosis Study Group. 2012. Effect of selection and genetic diversity on drug resistance in Mycobacterium tuberculosis. J Med Microbiol 61:1120–1125. https://doi.org/10.1099/jmm.0.06460-11.

30. Wengenack NL, Uh JH, St Amand AL, Tomlinson AJ, Benson LM, Naylor S, Kline BC, Cockerill FR III, Ruskarn F. 1997. Reombinant Mycobacterium tuberculosis katG(S315T) is a competent catalase-peroxidase with reduced activity toward isoniazid. J Infect Dis 176:722–727. https://doi.org/10.1086/514096.

31. Pym AS, Saint-Joanis B, Cole ST. 2002. Effect of katG mutations on the virulence of Mycobacterium tuberculosis and the implication for trans-
mission in humans. Infect Immun 70:4955–4960. https://doi.org/10.1128/IAI.70.9.4955–4960.2002.

44. Spies FS, von Groll A, Ribeiro AW, Ramos DF, Ribeiro MO, Dalla Costa ER, Martin A, Palomino JC, Rossetti ML, Zaha A, da Silva PEA. 2013. Biological cost in Mycobacterium tuberculosis with mutations in the rplS, rrs, rpoB, and katG genes. Tuberculosis 93:150–154. https://doi.org/10.1016/j.tube.2012.11.004.

45. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, Ellerton B, Maruri F, Blackman A, Herrera M, Samuels DC, Sterling TR. 1999. Physiological cost of resistance in Mycobacterium tuberculosis revisited: a novel association with broad drug resistance in globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat Genet 49:395–402. https://doi.org/10.1038/ng.3767.

46. Billington OJ, McHugh TD, Gillespie SH. 2012. Role of embB codon 306 mutations in Mycobacterium tuberculosis. Antimicrob Agents Chemother 52:2503–2511.

47. Eilertson B, Maruri F, Blackman A, Herrera M, Samuels DC, Sterling TR. 2005. The Sequence Alignment/Map format and SAMtools. Bioinformatics 21:218–219. https://doi.org/10.1093/bioinformatics/bth342.

48. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, Ruben EJ, Schnappinger D, Ehrt S, Fortune SM, Sassetti CM, Loeger TR. 2017. Comprehensive essentiality analysis of the Mycobacterium tuberculosis genome via saturating transposon mutagenesis. mBio 8:e02133-16. https://doi.org/10.1128/mBio.02133-16.

49. Shekar S, Yeo ZX, Wong JCL, Chan MKL, Ong DCT, Tongyoo P, Wong S-Y, Lee ASG. 2014. Detecting novel genetic variants associated with isoniazid-resistant Mycobacterium tuberculosis. PLoS One 9:e102383. https://doi.org/10.1371/journal.pone.0102383.

50. Danilchanka O, Malaiender C, Niederweis M. 2008. Identification of a novel multidrug efflux pump of Mycobacterium tuberculosis. Antimicrob Agents Chemother 52:2503–2511. https://doi.org/10.1128/AAC.00298-08.

51. Fu LM, Shinnick TM. 2007. Genome-wide exploration of the drug action of capreomycin on Mycobacterium tuberculosis using Affymetrix oligo-nucleotide GeneChips. J Infect 54:277–284. https://doi.org/10.1016/j.jinf.2006.05.012.

52. Phong TQ, Ha DTT, Volker U, Hammer E. 2015. Using a label free quantitative proteomics approach to identify changes in protein abundance in multidrug-resistant Mycobacterium tuberculosis. Indian J Microbiol 55:219–230. https://doi.org/10.1007/s12088-015-0511-2.

53. Andrews S. 2012. FastQC. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

54. Kreuger F. 2013. TrimGalore! Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

55. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-bio.GN].

56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

57. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2– approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.

58. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-bio.GN].

59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

60. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056. https://doi.org/10.1099/mggom.0.000056.

61. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056. https://doi.org/10.1099/mggom.0.000056.

62. Cingolani P, Platts A, Wang LL, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

63. Salazar A, Earl A, Desjardins C, Abeel T. 2015. Normalizing alternate representations of large sequence variants across multiple bacterial genomes. BMC Bioinformatics 16:A8. https://doi.org/10.1186/s12859-015-0813-4.

64. De Mita S, Siol M. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC Genet 13:27. https://doi.org/10.1186/1471-2156-13-27.

65. Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x.

66. Sagulenko P, Puller V, Neher R. 2017. TreeTime: maximum likelihood phylogenetic analysis. bioRxiv https://doi.org/10.1101/153494.