PLOS BIOLOGY

SHORT REPORTS

Genome-wide association studies of global *Mycobacterium tuberculosis* resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms

The CRyPTIC Consortium[¶]*

University of Oxford, Oxford, United Kingdom

¶ Membership of The CRyPTIC Consortium is provided in Supporting information file S1 Acknowledgments.
* daniel.wilson@bdi.ox.ac.uk

Citation: The CRyPTIC Consortium (2022) Genome-wide association studies of global *Mycobacterium tuberculosis* resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. PLoS Biol 20(8): e3001755. https://doi.org/10.1371/journal.pbio.3001755

Academic Editor: Jason Ladner, Northern Arizona University, UNITED STATES

Received: March 25, 2022

Accepted: July 12, 2022

Published: August 9, 2022

Data Availability Statement: The log2 MIC phenotypes and short read genome sequence accession numbers used in the study are available in S2 Data. The phenotypes, genomes and analysis results are available from https://ftp.ebi.ac.uk/pub/databases/cryptic/release_june2022/pubs/gwas2022. The oligo-based GWAS pipeline is available from https://github.com/danny-wilson/kmer_pipeline.

Abstract

The emergence of drug-resistant tuberculosis is a major global public health concern that threatens the ability to control the disease. Whole-genome sequencing as a tool to rapidly diagnose resistant infections can transform patient treatment and clinical practice. While resistance mechanisms are well understood for some drugs, there are likely many mechanisms yet to be uncovered, particularly for new and repurposed drugs. We sequenced 10,228 *Mycobacterium tuberculosis* (MTB) isolates worldwide and determined the minimum inhibitory concentration (MIC) on a grid of 2-fold concentration dilutions for 13 antimicrobials using quantitative microtiter plate assays. We performed oligopeptide- and oligonucleotide-based genome-wide association studies using linear mixed models to discover resistance-conferring mechanisms not currently catalogued. Use of MIC over binary resistance phenotypes increased sample heritability for the new and repurposed drugs by 26% to 37%, increasing our ability to detect novel associations. For all drugs, we discovered uncatalogued variants associated with MIC, including in the *Rv1218c* promoter binding site of the transcriptional repressor *Rv1219c* (isoniazid), upstream of the *vapBC20* operon that cleaves 23S rRNA (linezolid) and in the region encoding an α-helix lining the active site of Cyp142 (clofazimine, all $p < 10^{-7.7}$). We observed that artefactual signals of cross-resistance could be unravelled based on the relative effect size on MIC. Our study demonstrates the ability of very large-scale studies to substantially improve our knowledge of genetic variants associated with antimicrobial resistance in *M. tuberculosis*.

Introduction

Tuberculosis (TB) continues to represent a major threat to global public health, with the World Health Organization (WHO) estimating 10 million cases and 1.4 million deaths in 2019 alone [1]. Multidrug resistance (MDR) poses a major challenge to tackling TB; it is estimated that there were 465,000 cases of rifampicin-resistant TB in 2019, of which 78% were resistant

Funding: This work was supported by Wellcome Trust/Newton Fund-MRC Collaborative Award (200205/Z/15/Z); and Bill & Melinda Gates Foundation Trust (OPP1133541). Oxford CRyPTIC consortium members are funded/supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health, and the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, a partnership between Public Health England and the University of Oxford, the views expressed are those of the authors and not necessarily those of the NIHR, Public Health England or the Department of Health and Social Care. J.M. is supported by the Wellcome Trust (203919/Z/16/Z). Z.Y. is supported by the National Science and Technology Major Project, China Grant No. 2018ZX10103001. K.M.M. is supported by EMBL's EIPOD3 programme funded by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska Curie Actions. T.C.R. is funded in part by funding from Unitaid Grant No. 2019-32-FIND MDR. R.S.O. is supported by FAPESP Grant No. 17/16082-7. L.F. received financial support from FAPESP Grant No. 2012/51756-5. B.Z. is supported by the National Natural Science Foundation of China (81991534) and the Beijing Municipal Science & Technology Commission (Z201100005520041). N.T.T.T. is supported by the Wellcome Trust International Intermediate Fellowship (206724/Z/17/Z). G.T. is funded by the Wellcome Trust. R.W. is supported by the South African Medical Research Council. J.C. is supported by the Rhodes Trust and Stanford Medical Scientist Training Program (T32 GM007365). A.L. is supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Respiratory Infections at Imperial College London. S.G.L. is supported by the Fonds de Recherche en Santé du Québec. C.N. is funded by Wellcome Trust Grant No. 203583/Z/16/Z. A.V.R. is supported by Research Foundation Flanders (FWO) under Grant No. G0F8316N (FWO Odysseus). G.M. was supported by the Wellcome Trust (098316, 214321/Z/18/Z, and 203135/Z/16/Z), and the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa (Grant No. 64787). The funders had no role in the study design, data collection, data analysis, data interpretation, or writing of this report. The opinions, findings and conclusions expressed in this manuscript reflect those of the authors alone.

to the first-line drugs rifampicin and isoniazid—called MDR-TB [1]. While treatment is 85% successful overall, that drops to 57% for rifampicin-resistant and MDR-TB [1]; underdiagnosis and treatment failures then amplify the problem by encouraging onward transmission of MDR-TB [2]. New treatment regimens for MDR-TB are therefore an important focus, introducing new and repurposed drugs such as bedaquiline, clofazimine, delamanid, and linezolid [3,4]; however, resistance is already emerging [5–7].

Understanding mechanisms of resistance in TB is important for developing rapid susceptibility tests that improve individual patient treatment, recommending drug regimens that reduce the development of MDR and developing new and improved drugs that expand treatment options [8,9]. Genomics can accelerate drug susceptibility testing, replacing slower culture-based methods by predicting resistance from the sequenced genome rather than directly phenotyping the bacteria [10]. Genome sequencing-based susceptibility testing for first-line drugs has achieved sensitivities of 91.3% to 97.5% and specificities of 93.6% to 99.0% [11], surpassing the thresholds for clinical accreditation, motivating its adoption by multiple public health authorities [12]. In low-resource settings, molecular tests such as Cepheid GeneXpert and other line probe assays offer rapid and more economical susceptibility testing by genotyping a panel of known resistance-conferring genetic variants [13], with performance close to that achieved by whole-genome sequencing [14,15]. However, the limited number of resistance-conferring mutations that can be included in such tests can lead to missed MDR diagnoses and incorrect treatment [11,16]. Both approaches rely on the development and maintenance of resistance catalogues of genetic variants [11,17].

In the discovery of resistance-conferring variants, traditional molecular approaches have been replaced by high-throughput, large-scale whole-genome sequencing studies of hundreds to thousands of resistant and susceptible clinical isolates [18–23]. Despite the strong performance of genome-based resistance prediction for first-line drugs, knowledge gaps remain, especially for second-line drugs [17,24,25]. There are numerous challenges in the pursuit of previously uncatalogued resistance mechanisms. Very large sample sizes are needed to identify rarer resistance mechanisms with confidence. The lack of recombination in *Mycobacterium tuberculosis* makes it difficult to pinpoint resistance variants unless they arise on multiple genetic backgrounds, reiterating the need for large sample sizes. Sophisticated analyses are required that attempt to disentangle genetic causation from correlation [26]. A reliance on a binary resistance/sensitivity classification paradigm has hindered reproducibility for some drugs by failing to mirror the continuous nature of resistance [27–29].

The aim of Comprehensive Resistance Prediction for Tuberculosis: An International Consortium (CRyPTIC) was to address these challenges by assembling a global collection of over 10,000 *M. tuberculosis* isolates from 27 countries followed by whole-genome sequencing and semiquantitative determination of minimum inhibitory concentration (MIC) to 13 first- and second-line drugs using a bespoke 96-well broth micodilution plate assay. The development of novel, inexpensive, high-throughput drug susceptibility testing assays allowed us to conduct the project at scale, while investigating MIC on a grid of 2-fold concentration dilutions [30,31]. Here, we report the identification of previously uncatalogued resistance-conferring variants through 13 genome-wide association studies (GWAS) investigating MIC values in 10,228 *M. tuberculosis* isolates. We identify these discoveries relative to specific catalogues for the sake of concreteness and tractability, while acknowledging the catalogues do not include all credible resistance mechanisms known from the literature; we expand on this limitation in the Discussion. Our analyses employed a linear mixed model (LMM) to identify putative causal variants while controlling for confounding and genome-wide linkage disequilibrium (LD) [20,32]. We developed a novel approach to testing associations at both 10,510,261 oligopeptides (11-mers) and 5,530,210 oligonucleotides (31-mers) to detect relevant genetic

L.G. was supported by the Wellcome Trust (201470/Z/16/Z), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number 1R01AI146338, the GOSH Charity (VC0921) and the GOSH/ICH Biomedical Research Centre (www.nihr.ac.uk). A.B. is funded by the NDM Prize Studentship from the Oxford Medical Research Council Doctoral Training Partnership and the Nuffield Department of Clinical Medicine. D.J.W. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant No. 101237/Z/13/B) and by the Robertson Foundation. A.S.W. is an NIHR Senior Investigator. T.M.W. is a Wellcome Trust Clinical Career Development Fellow (214560/Z/18/Z). A.S.L. is supported by the Rhodes Trust. R.J.W. receives funding from the Francis Crick Institute which is supported by Wellcome Trust, (FC0010218), UKRI (FC0010218), and CRUK (FC0010218). T.C. has received grant funding and salary support from US NIH, CDC, USAID and Bill and Melinda Gates Foundation. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. Parts of the work were funded by the German Center of Infection Research (DZIF). The Scottish Mycobacteria Reference Laboratory is funded through National Services Scotland. The Wadsworth Center contributions were supported in part by Cooperative Agreement No. U60OE000103 funded by the Centers for Disease Control and Prevention through the Association of Public Health Laboratories and NIH/NIAID grant AI-117312. Additional support for sequencing and analysis was contributed by the Wadsworth Center Applied Genomic Technologies Core Facility and the Wadsworth Center Bioinformatics Core. SYNLAB Holding Germany GmbH for its direct and indirect support of research activities in the Institute of Microbiology and Laboratory Medicine Gauting. N.R. thanks the Programme National de Lutte contre la Tuberculose de Madagascar. The funders had no role in study design (except through the normal peer review feedback process), data collection and analysis, decision to publish (except that it is understood that results of the study would be published in the normal way), or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: E.R. is employed by Public Health England and holds an honorary contract with Imperial College London. I.F.L. is Director of the Scottish Mycobacteria Reference Laboratory. S.N. receives funding from German Center for Infection Research, Excellenz Cluster

variation in both coding and non-coding sequences and to avoid a reference-based mapping approach that can inadvertently miss significant variation. We report previously uncatalogued variants associated with MIC for all 13 drugs, focusing on variants in the 20 most significant genes per drug. We highlight notable discoveries for each drug and demonstrate the ability of large-scale studies to improve our knowledge of genetic variants associated with antimicrobial resistance in *M. tuberculosis*.

Results

CRyPTIC collected isolates from 27 countries worldwide, oversampling for drug resistance [31]. A total of 10,228 genomes were included in total across the GWAS analyses: 533 were lineage 1; 3,581 were lineage 2; 805 were lineage 3; and 5,309 were lineage 4. Due to rigorous quality control, we dropped samples for each drug as detailed in the methods, resulting in a range of 6,388 to 9,418 genomes used in each GWAS, for which we constructed a phylogeny (**Fig 1A**). MICs were determined on a grid of 2-fold concentration dilutions for 13 antimicrobials using quantitative microtiter plate assays: first-line drugs: ethambutol, isoniazid, and rifampicin; second-line drugs: amikacin, ethionamide, kanamycin, levofloxacin, moxifloxacin, and rifabutin; and the new and repurposed drugs: bedaquiline, clofazimine, delamanid, and linezolid. The phenotype distributions differed between the drugs, with low numbers of sampled resistant isolates for the new and repurposed drugs that have not yet been widely used in TB treatment (**Figs 1B and S1**). Applying epidemiological cutoffs (ECOFFs) to the MIC [31], the GWAS featured 66 isolates resistant to bedaquiline, 97 resistant to clofazimine, 77 resistant to delamanid, and 67 resistant to linezolid. We performed oligopeptide- and oligonucleotide-based GWAS analyses, controlling for population structure using LMMs. We focused initially on oligopeptides, interpreting oligonucleotides only where necessary for clarifying results.

Estimates of sample heritability (variance in the phenotype explained by additive genetic effects) were higher for MIC compared to binary resistant versus sensitive phenotypes for the new and repurposed drugs bedaquiline, clofazimine, delamanid, and linezolid by at least 26%. Across drugs, binary heritability ranged from 0% to 94.7% and MIC heritability from 36.0% to 95.6%, focusing on oligopeptides (**Figs 2 and S2 and S1 Table**). For delamanid, binary heritability was not significantly different from zero (2.99×10^{-6}; 95% confidence interval (CI) 0.0% to 0.5%), while MIC heritability was 36.0% (95% CI 28.9% to 43.1%). Estimates of sample heritability were more similar between binary and MIC phenotypes for the remaining drugs, differing by −3.6% to +5.2%.

GWAS identified oligopeptide variants associated with changes in MIC for all 13 drugs after controlling for population structure (**Table 1 and Figs 3, S3 and S4**). In total, across the drugs, we tested for associations at 10,510,261 variably present oligopeptides and 5,530,210 oligonucleotides; these captured substitutions, insertions, and deletions. The drugs differed in the number of genes or intergenic regions that were significant, the drugs with fewest significant genes being isoniazid (12), levofloxacin (13), and moxifloxacin (6). We defined the significance of a gene or intergenic region by the most significant oligopeptide within it and assessed all significant variants above a 0.1% minor allele frequency (MAF) threshold for the top 20 significant genes. The top 20 genes for each drug are detailed in **Table 1**. Some variants were identified in novel genes, some were novel variants in known genes, and some were known variants. We highlight examples of these (in reverse order) in the following sections. Highlighted examples have been chosen to exclude genes or variants in LD with other regions where possible; some are in LD with other less significant variants.

We assessed whether the top genes for each drug were in either of 2 previously described resistance catalogues [11,17]; we describe variants in genes not in these catalogues as

Precision Medicine in Chronic Inflammation, Leibniz Science Campus Evolutionary Medicine of the LUNG (EvoLUNG)tion EXC 2167. P.S. is a consultant at Genoscreen. T.R. is funded by NIH and DoD and receives salary support from the non-profit organization FIND. T.R. is a co-founder, board member and shareholder of Verus Diagnostics Inc, a company that was founded with the intent of developing diagnostic assays. Verus Diagnostics was not involved in any way with data collection, analysis or publication of the results. T. R. has not received any financial support from Verus Diagnostics. UCSD Conflict of Interest office has reviewed and approved T.R.'s role in Verus Diagnostics Inc. T.R. is a co-inventor of a provisional patent for a TB diagnostic assay (provisional patent #: 63/048.989). T.R. is a co-inventor on a patent associated with the processing of TB sequencing data (European Patent Application No. 14840432.0 & USSN 14/912,918). T.R. has agreed to "donate all present and future interest in and rights to royalties from this patent" to UCSD to ensure that he does not receive any financial benefits from this patent. S.S. is working and holding ESOPs at HaystackAnalytics Pvt. Ltd. (Product: Using whole genome sequencing for drug susceptibility testing for Mycobacterium tuberculosis). G.F.G. is listed as an inventor on patent applications for RBD-dimer-based CoV vaccines. The patents for RBD-dimers as protein subunit vaccines for SARS-CoV-2 have been licensed to Anhui Zhifei Longcom Biopharmaceutical Co. Ltd, China.

Abbreviations: CI, confidence interval; ECOFF, epidemiological cutoff; FWER, family-wide error rate; GWAS, genome-wide association studies; LD, linkage disequilibrium; LMM, linear mixed model; MAF, minor allele frequency; MDR, multidrug resistance; MIC, minimum inhibitory concentration; MPTR, major polymorphic tandem repeat; NO, nitric oxide; TB, tuberculosis; WHO, World Health Organization.

uncatalogued (**Table 1**). The interpretation of oligopeptides and oligonucleotides required manual curation to determine the underlying variants they tagged; the most significant oligopeptide or oligonucleotide for each allele captured by the significant signals are described in **S1 Data**. For 8/13 drugs with previously catalogued resistance determinants, the most significant GWAS signal in CRyPTIC was a previously catalogued variant, consistent with previous GWAS in *M. tuberculosis* [18–23]. The most significant catalogued variants for each drug were (lowercase for nucleotides, uppercase for amino acids): *rrs* a1401g (amikacin, kanamycin), *embB* M306V (ethambutol), *fabG1* c−15t (ethionamide), *katG* S315T (isoniazid), *gyrA* D94G (levofloxacin, moxifloxacin), and *rpoB* S450L (rifampicin) [11,17]. For the remaining drugs, which had no resistance determinants in the catalogues to which we referred [11,17], the genes identified by the top signals were: *Rv0678* (bedaquiline, clofazimine), *ddn* (delamanid), *rplC* (linezolid), and *rpoB* (rifabutin). However, for all these associations, there does exist credible evidence elsewhere in the literation (e.g., [85,86]). The top variants identified for each drug were all significant at $p < 1.04 \times 10^{-15}$.

For many drugs, the most significant oligopeptide was high frequency, and the direction of effect was to decrease MIC relative to alternative alleles (**S5 Fig**). This implies that oligopeptides and oligonucleotides associated with lower MIC are more likely to be genetically identical across strains than those associated with higher MIC. This would be consistent with the independent evolution of increased MIC from a shared, low-MIC TB ancestor. Often there were multiple low-frequency oligopeptides mapping to the same positions, supporting this idea.

Uncatalogued variants significantly associated with MIC are important because they could improve resistance prediction and shed light on underlying resistance mechanisms; they may be novel or previously implicated in resistance but not to a standard of evidence sufficient to be catalogued. We discuss the choice of catalogues in the Discussion [11,17].

We next looked at uncatalogued variants in known resistance-conferring genes. We identified uncatalogued variants in *gyrB* associated with levofloxacin and moxifloxacin MIC (minimum *p*-value levofloxacin: $p < 10^{-15.6}$, moxifloxacin: $p < 10^{-11.6}$). The primary mechanisms of resistance to the fluoroquinolones levofloxacin and moxifloxacin are mutations in *gyrA* or *gyrB*, the subunits of DNA gyrase. The *gyrB* Manhattan plots for levofloxacin and moxifloxacin both contained 2 adjacent peaks within the gene, but for each drug just 1 of the 2 peaks was significant, and these differed between the drugs (**Fig 4**). Interpretation of oligopeptides and oligonucleotides requires an understanding of the variants that they capture, which we visualised by aligning them to H37Rv and interpreting the variable sites (e.g., **Fig 4C and 4D**). For levofloxacin, the peak centred around amino acid 461. Significant oligopeptides captured amino acids 461 and 457, which are both uncatalogued [11,17] with 457 falling just outside of the *gyrB* quinolone resistance-determining region (QRDR-B) [33]. Oligopeptides capturing 461N were associated with increased MIC (e.g., **N**SAGGSAKSGR, $-\log_{10}p = 15.65$, effect size $\beta = 2.46$, present in 15/7,300 genomes). Oligopeptides capturing the reference alleles at codons 461 and 457 were significantly associated with lower MIC (e.g., 461D: **D**SAGGSAKSGR, $-\log_{10}p = 13.47$, $\beta = -2.14$, present in 7,278/7,300 genomes; 457V/461D: SELY**V**VEG**D**SA, $-\log_{10}p = 12.51$, $\beta = -1.96$, present in 7,272/7,300 genomes). For moxifloxacin, the peak centred around amino acid 501. Significant oligopeptides captured amino acids 499 and 501. Oligopeptides capturing 501D were associated with increased MIC (e.g., NT**D**VQAIITAL, $-\log_{10}p = 10.64$, $\beta = 1.86$, present in 23/6,388 genomes). Oligopeptides capturing the reference allele at codons 499 and 501 were associated with lower MIC (e.g., **N**T**E**VQAIITAL, $-\log_{10}p = 11.63$, $\beta = -1.33$, present in 6,332/6,388 genomes). Amino acids 461 and 501 are at the interface between *gyrB* and the bound fluroquinolone [34]. *gyrB* is included in the reference catalogues for predicting levofloxacin (including D461N) but not moxifloxacin resistance;

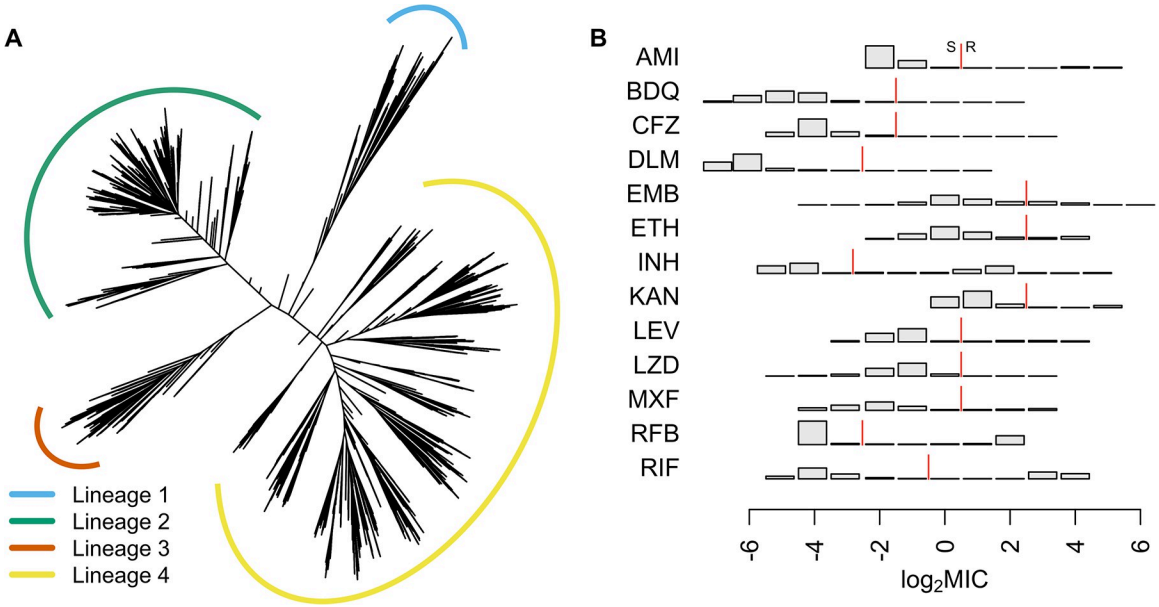

Fig 1. (**A**) Phylogeny of 10,228 isolates sampled globally by CRyPTIC used in the GWAS analyses. Lineages are coloured blue (lineage 1), green (2), orange (3), and yellow (4). Branch lengths have been square root transformed to visualise the detail at the tips. (**B**) Distributions of the $\log_2$ MIC measurements for all 13 drugs in the GWAS analyses, AMI, BDQ, CFZ, DLM, EMB, ETH, INH, KAN, LEV, LZD, MXF, RFB, and RIF. The red line indicates the ECOFF breakpoint for binary resistance versus sensitivity calls [31]. AMI, amikacin; BDQ, bedaquiline; CFZ, clofazimine; DLM, delamanid; ECOFF, epidemiological cutoff; EMB, ethambutol; ETH, ethionamide; INH, isoniazid; KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MIC, minimum inhibitory concentration; MXF, moxifloxacin; RFB, rifabutin; RIF, rifampicin.

https://doi.org/10.1371/journal.pbio.3001755.g001

therefore, our results support inclusion of *gyrB* (in particular E501D) in future moxifloxacin catalogues [11,17].

Next, we looked at specific examples of significant associations identified by GWAS in genes not catalogued by [11,17] for each of the drugs. A well-recognized challenge in GWAS for antimicrobial resistance is the presence of artefactual cross-resistance. To mitigate this risk, we preferentially highlight variants significantly associated with a single drug. However, many catalogued resistance variants demonstrated artefactual cross-resistance. For example, variants in the rifampicin resistance-determining region were in the top 20 significant associations for all drugs except for delamanid (**Table 1**). Interestingly, we observed that the magnitude of effect sizes was often larger on MIC of the drug to which catalogued variants truly confer resistance (**S6 Fig**). For example, the effect sizes for significant oligopeptides in *rpoB* were greater for rifampicin and rifabutin than for all other drugs. This suggests that the β estimates could help to prioritise drugs for follow up when genes are significantly associated with multiple drugs.

First-line drugs

Ethambutol and rifampicin. Oligonucleotides downstream of *spoU* (*Rv3366*) were significantly associated with ethambutol and rifampicin MIC (minimum p-value $p < 10^{-10.0}$, **S7 Fig**). SpoU is a tRNA/rRNA methylase, shown to have DNA methylation activity [35]. As the association was outside of the coding region, we interpreted oligonucleotides for this association. Oligonucleotides associated with increased MIC captured the relatively common adenine 20 nucleotides downstream of the stop codon (e.g., C**A**AACCAGCCGGTATGCGCACAACGAAGCTC, RIF: $-\log_{10}p = 12.82$, $\beta = 3.19$, present in 159/8,394 genomes; EMB:

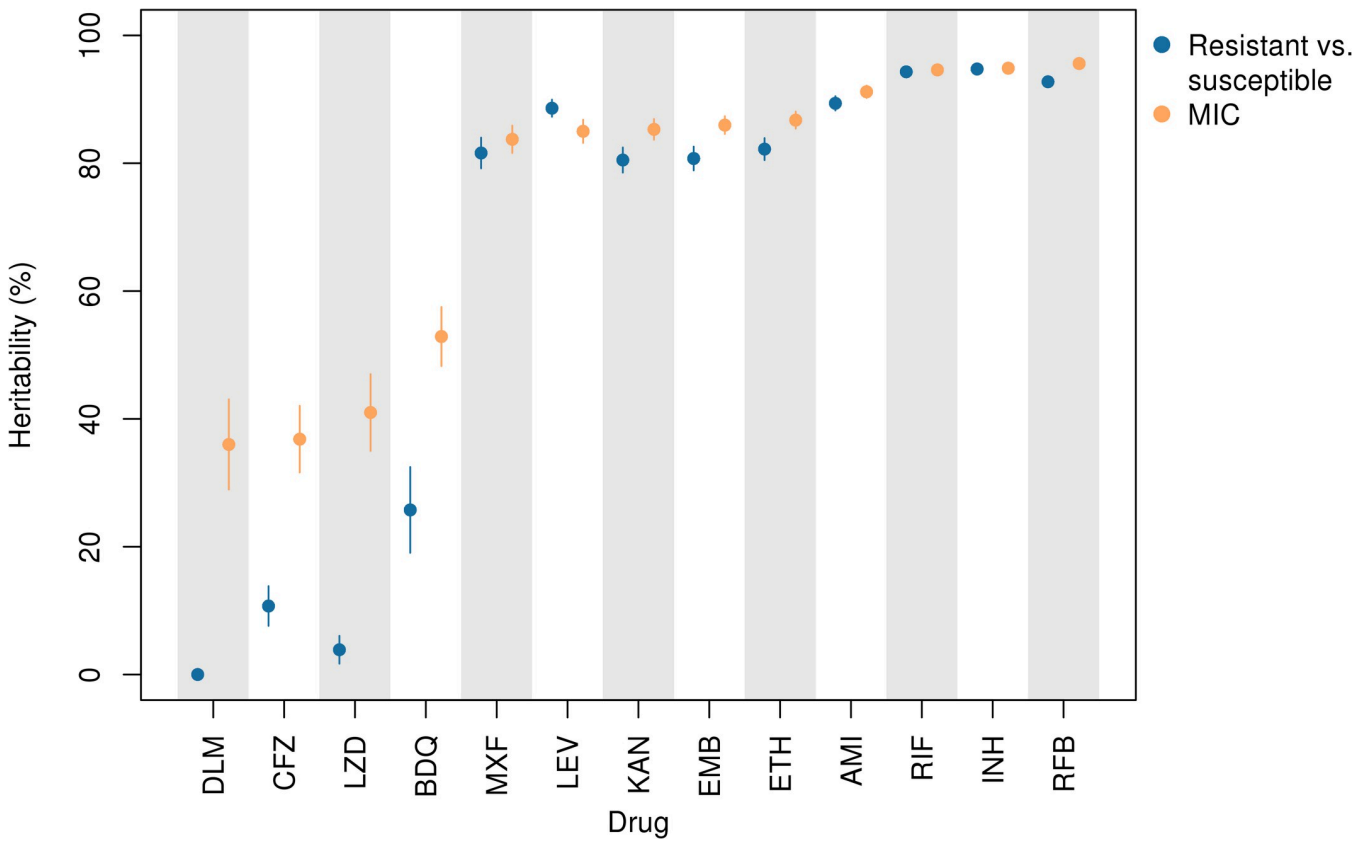

Fig 2. Sample heritability for MIC (orange) versus binary resistance/sensitivity (blue) assuming additive genetic variation in oligopeptide presence/absence across 13 drugs, DLM, CFZ, LZD, BDQ, MXF, LEV, KAN, EMB, ETH, AMI, RIF, INH, and RFB. Lines depict 95% CIs. MIC heritability was at least 26% higher than binary heritability for the new and repurposed drugs BDQ, CFZ, DLM, and LZD. AMI, amikacin; BDQ, bedaquiline; CFZ, clofazimine; CI, confidence interval; DLM, delamanid; EMB, ethambutol; ETH, ethionamide; INH, isoniazid; KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MIC, minimum inhibitory concentration; MXF, moxifloxacin; RFB, rifabutin; RIF, rifampicin.

https://doi.org/10.1371/journal.pbio.3001755.g002

$-\log_{10}p = 10.86$, $\beta = 1.36$, present in 163/7,081 genomes). This mutation has been identified in previous association studies as associated with rifampicin and ethambutol resistance [36,37] but has not been catalogued. The new evidence provided by CRyPTIC supports reevaluation of this putative resistance-conferring variant. The simultaneous association of *spoU* with rifampicin and ethambutol may be an example of artefactual cross-resistance. The effect sizes on MIC for rifampicin ($\beta = 3.19$) were larger than for ethambutol ($\beta = 1.36$), suggesting prioritisation of the rifampicin association over the ethambutol association reported here.

Isoniazid. Oligopeptides in *Rv1219c* were significantly associated with isoniazid MIC (minimum *p*-value $p < 10^{-8.5}$, **S8 Fig**). Rv1219c represses transcription of the Rv1217c-Rv1218c multidrug efflux transport system [38]. It binds 2 motifs, a high-affinity intergenic sequence in the operon's promoter and a low-affinity intergenic sequence immediately upstream of *Rv1218c* [38]. The peak signal of association coincides with the C-terminal amino acids 188 to 189 in the low-affinity binding domain of Rv1219c. Multiple extremely low-frequency oligopeptides were associated with increased MIC, present in just 1 or 2 genomes. In contrast, oligopeptides containing the reference alleles at codons 188 to 189 were present in 8,919/8,929 genomes and strongly associated with decreased MIC (e.g., EVYT**EG**LLADR, $-\log_{10}p = 8.46$, $\beta = -3.63$, present in 8,919/8,929 genomes). Substitutions at these positions may

Table 1. The top genes or intergenic regions ranked by their most significant oligopeptides per drug, up to a maximum of 20 (more only when the 20th was tied). Genes are highlighted in bold if they were catalogued for that drug by [11,17]. Gene names separated by colons indicate intergenic regions. Genes or intergenic regions capturing repeat regions are highlighted with the superscript ^{R}. Alphabetic characters following gene names are used to cross-reference with the corresponding Manhattan plots in Fig 3.

Drug	Top significant genes and intergenic regions
First-line	
Ethambutol	***embB***, *rpoB* (A), *katG* (B), ***embA***, *pncA* (C), *gyrA* (D), *rpsL* (E), *Rv1565c* (F), *Rv2478c:Rv2481c* (G), *Rv1752* (H), *Rv3183:Rv3188* ^{R} (I), *dxs2:Rv3382c* ^{R} (J), *rpsA/coaE* (K), *ctpI* (L), *guaA* (M), *moaC3:Rv3327* ^{R} (N), *lprF:Rv1371* ^{R} (O), *fabG1* (P), *spoU* (Q), *glpK* (R)
Isoniazid	***katG***, ***proA:ahpC***, ***fabG1***, *rpoB* (A), ***inhA***, *embB* (B), *Rv1139c:Rv1140* (C), *Rv1158c* (D), *rpsL* (E), *Rv1219c* (F), *ftsK/Rv2749* (G), *gid* (H)
Rifampicin	***rpoB***, *katG* (A), *embB* (B), *Rv1565c* (C), *guaA* (D), *ctpI* (E), *spoU* (F), *dxs2:Rv3382c* ^{R} (G), *Rv3183:Rv3188* ^{R} (H), *relA* (I), *proA:ahpC* (J), *fabG1* (K), *moaC3:Rv3327* ^{R} (L), *Rv0810c* (M), *fadD9* (N), *Rv3779* (O), *rpsL* (P), *rpoC* (Q), *Rv2190c:Rv2191* (R)
Second-line	
Amikacin	***rrs***, *gyrA* (A), *rpoB* (B), *echA8* (C), *Rv2896c* (D), *Rv0078A* (E), *Rv1830* (F), *Rv0792c/Rv0793* (G), *PPE54* (H), *Rv2041c* (I), *PPE42* (J), *cyp141:Rv3122* (K), *Rv1765c* ^{R} (L), *lprF:Rv1371* ^{R} (M), *espA:ephA* (N), *narU* (O), *rne* (P), *Rv1393c* (Q), *Rv1362c* (R), *Rv0579* (S), *glnE* (T), *ethA* (U), *Rv0208c:Rv0209* (V)
Ethionamide	***fabG1***, *ethA* (A), *rpoB* (B), *gyrA* (C), *inhA* (D), *whiB7* (E), *PPE3* (F), *mpt53* (G), *embB* (H), *eccA1* (I), *embA* (J), *Rv0565c* (K), *fadB4* (L), *plsC* (M), *Rv0920c* (N), *Rv3698* (O), *rrs* (P), *pncA* (Q), *PPE56* (R), *Rv2019* (S), *lprF:Rv1371* ^{R} (T)
Kanamycin	***rrs***, ***eis***, *gyrA* (A), *rpoB* (B), *ethA* (C), *fabG1* (D), *Rv1830* (E), *ptbB* (F), *PPE42* (G), *echA8* (H), *lprF:Rv1371* ^{R} (I), *Rv2348c:plcC* (J), *narU* (K), *pgi* (L), *mmaA4* (M), *pncA* (N), *viuB* (O), *lprC* (P), *murA* (Q), *Rv1393c* (R), *Rv0579* (S), *glnE* (T), *rne* (U), *Rv1362c* (V), *Rv0208c:Rv0209* (W)
Levofloxacin	***gyrA***, *rrs* (A), ***gyrB***, *embB* (B), *rpoB* (C), *vapC36* (D), *mce2F* (E), *fabG1* (F), *katG* (G), *folC* (H), *tlyA* (I), *ethA* (J), *Rv0228* (K)
Moxifloxacin	***gyrA***, *rrs* (A), *rpoB* (B), *gyrB* (C), *embB* (D), *katG* (E)
Rifabutin	*rpoB*, *embB* (A), *katG* (B), *rpoC* (C), *Rv0810c* (D), *Rv2478c:Rv2481c* ^{R} (E), *Rv2647:Rv2650c* ^{R} (F), *rplP* (G), *Rv2797c* (H), *cpsY* (I), *lysA* (J), *mprB* (K), *mprA* (L), *Rv3228* (M), *Rv1290c* (N), *pncA* (O), *Rv2277c:pitB* ^{R} (P), *Rv0726c* (Q), *cysA3/cysA2* (R), *Rv0914c* (S)
New and repurposed	
Bedaquiline	*Rv0678*, *rpoB* (A), *rrs* (B), *atpE* (C), *pgi* (D), *mmaA4* (E), *rplC* (F), *Rv0078A* (G), *era/amiA2* (H), *viuB* (I), *pncA* (J), *murA* (K), *Rv0792c/Rv0793* (L), *dnaB* (M), *Rv2665:clpC2* (N), *PPE54* (O), *Rv0332* (P), *Rv2019* (Q), *vapC22* (R), *Rv2896c* (S)
Clofazimine	*Rv0678*, *fabG1* (A), *cyp142* (B), *Rv3183:Rv3188* ^{R} (C), *moaC3:Rv3327* ^{R} (D), *dxs2:Rv3382c* ^{R} (E), *mmsA* (F), *Rv3723:Rv3725* (G), *gid* (H), *rpoB* (I), *pks1* (J), *mmaA2:mmaA1* (K), *Rv3273* (L), *mce3R/yrbE3A* (M), *Rv3796* (N), *mez* (O), *Rv2390c* (P), *yrbE3B* (Q), *Rv0207c* (R), *argS* (S)
Delamanid	*ddn*, *fadE22* (A), *fba* (B), *Rv2180c* (C), *gap* (D), *lprF:Rv1371* ^{R} (E), *Rv0914c* (F), *Rv1200* (G), *fadE10* (H), *dinP* (I), *mmpL8* (J), *cut1* ^{R} (K), *PPE39* ^{R} (L), *Rv3430a:gadB* (M), *Rv1429* (N), *Rv3847* (O), *pknH* (P), *plsC* (Q), *agpS* (R), *Rv3263* (S)
Linezolid	*rplC*, *rpoB* (A), *emrB* (B), *Rv3552* (C), *add* (D), *vapC33* (E), *ppgK* (F), *pncB1:Rv1331* (G), *lprA* (H), *pafA* (I), *PE_PGRS6* (J), *vapB20* (K), *Rv0061c* (L), *PE_PGRS4* (M), *Rv1049* (N), *lprF:Rv1371* ^{R} (O), *Rv3183:Rv3188* ^{R} (P), *dxs2:Rv3382c* ^{R} (Q), *Rv0556* (R), *Rv0514* (S)

https://doi.org/10.1371/journal.pbio.3001755.t001

therefore derepress the multidrug efflux transport system. Indeed, overexpression of Rv1218c has been observed to correlate with higher isoniazid MIC in vitro [39].

Second-line drugs

Amikacin and kanamycin. Oligopeptides in *PPE42* (*Rv2608*) were significantly associated with aminoglycoside MIC, for both amikacin and kanamycin (minimum p-value $p < 10^{-12.8}$, **S9 Fig**). PPE42 is an outer membrane-associated PPE-motif family protein and potential B cell antigen. It elicits a high humoral and low T-cell response [40] and is 1 of 4 antigens in the vaccine candidate ID93 [41]. The C-terminal major polymorphic tandem repeats (MPTRs)

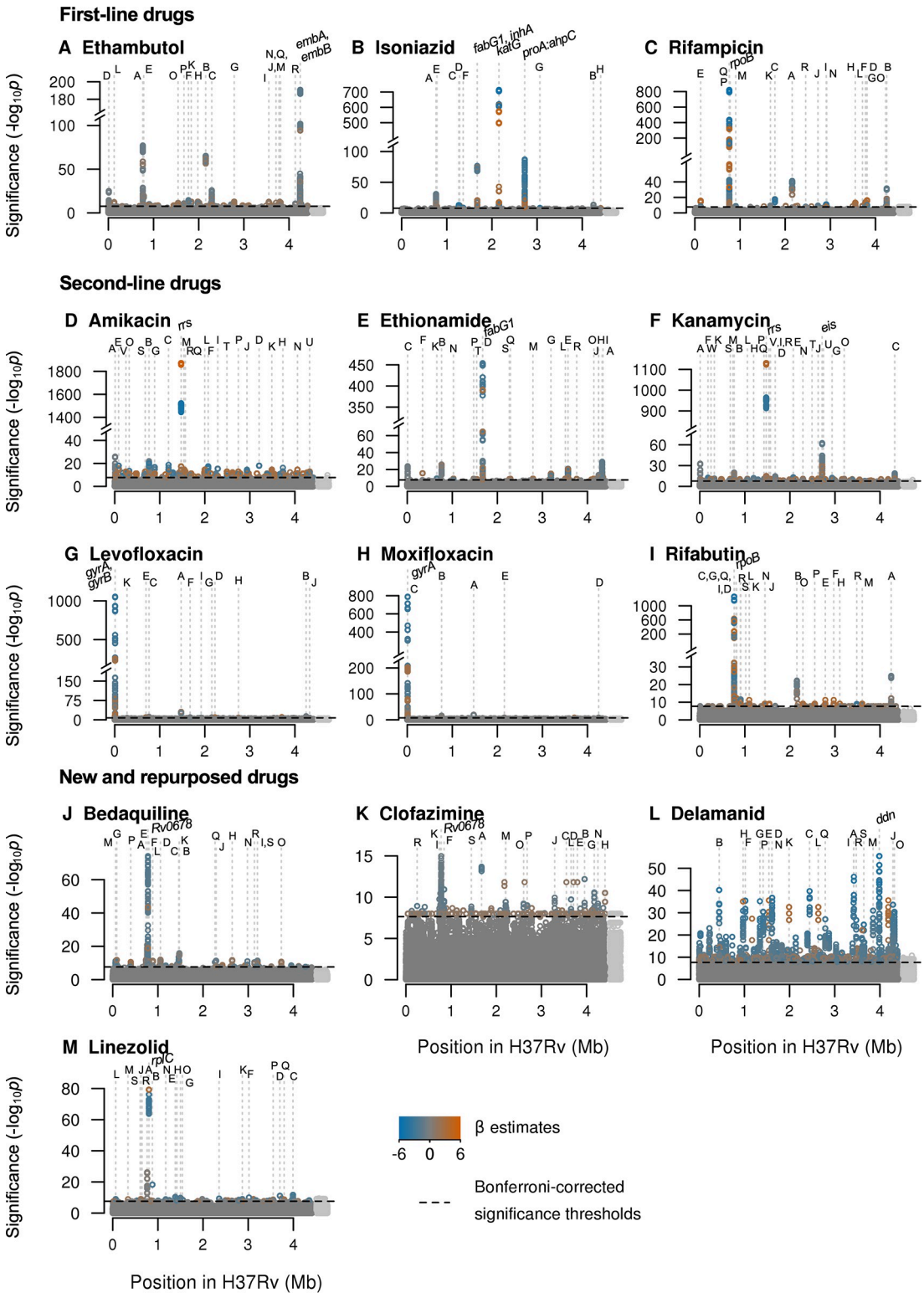

Fig 3. Manhattan plots of regions containing oligopeptide variants associated with MIC across 13 drugs. Significant oligopeptides are coloured by the direction (orange = increase, blue = decrease) and magnitude of their effect size on MIC, estimated by LMM [32]. Bonferroni-corrected significance thresholds are shown by the black dashed lines. The top 20 genes ranked by their most significant oligopeptides are annotated alphabetically. Gene names separated by colons indicate intergenic regions. Gene names for those annotated with letters can be found in Table 1. Oligopeptides were aligned to the H37Rv reference;

unaligned oligopeptides are plotted to the right in light grey. LMM, linear mixed model; MIC, minimum inhibitory concentration.

https://doi.org/10.1371/journal.pbio.3001755.g003

contain a region of high antigenicity [40]. The peak association with MIC occurred halfway along the coding sequence. The oligopeptides most associated with higher MIC captured a premature stop codon at position 290 (e.g., PLLE*AARFIT, amikacin $-\log_{10}p = 11.25$, $\beta = 3.12$, present in 38/8,430 genomes; kanamycin $-\log_{10}p = 10.25$, $\beta = 2.33$, present in 40/8,748 genomes). A nearby premature stop codon at amino acid 484 was previously identified in a multidrug-resistant strain [42], supporting the proposition that truncation of PPE42 enhances aminoglycoside resistance.

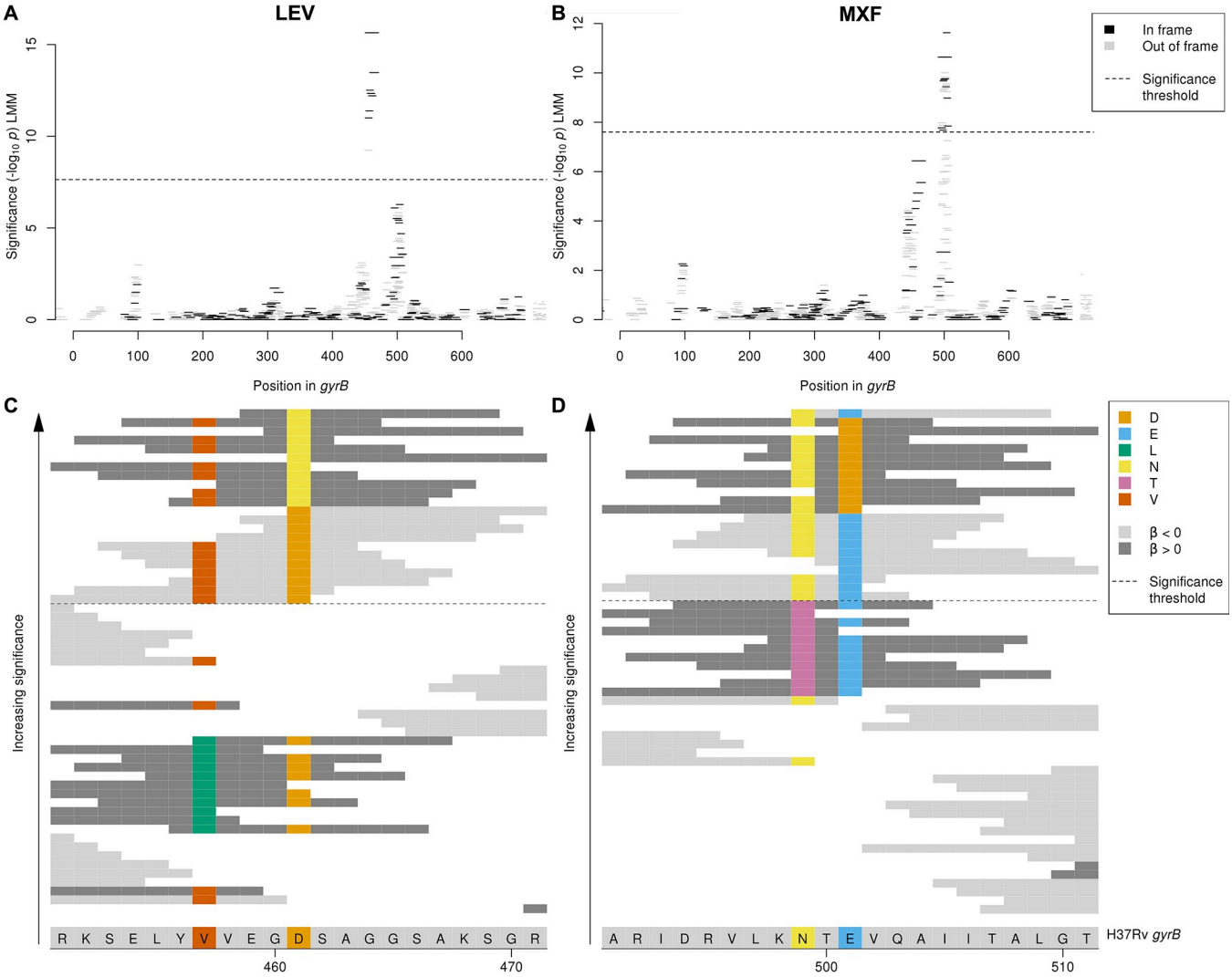

Fig 4. Interpreting significant oligopeptide variants for levofloxacin and moxifloxacin MIC in *gyrB*. Oligopeptide Manhattan plots are shown for (**A**) levofloxacin and (**B**) moxifloxacin. Oligopeptides are coloured by the reading frame that they align to, black for in frame and grey for out of frame in *gyrB*. Oligopeptides aligned to the region by nucmer but not realigned by BLAST are shown in grey on the right-hand side of the plots. The black dashed lines indicate the Bonferroni-corrected significance thresholds—all oligopeptides above the line are genome-wide significant. Alignment is shown of oligopeptides significantly associated with (**C**) levofloxacin and (**D**) moxifloxacin. The H37Rv reference codons are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The background colour of the oligopeptides represents the direction of the β estimate, light grey when $\beta < 0$ (associated with lower MIC), dark grey when $\beta > 0$ (associated with higher MIC). Oligopeptides are coloured by their amino acid residue at variant positions only. MIC, minimum inhibitory concentration.

https://doi.org/10.1371/journal.pbio.3001755.g004

Ethionamide. Oligopeptides and oligonucleotides upstream and within the transcriptional regulator *whiB7* (*Rv3197A*) were significantly associated with ethionamide MIC (minimum *p*-value $p < 10^{-18.2}$, **S10 Fig**). Oligonucleotides associated with higher MIC captured a single-base guanine deletion 177 bases upstream of *whiB7*, within the 5′ untranslated region [43] (e.g., AACCGTGTCGCCGCCGCGACTGACGAGTCCT, $-\log_{10}p = 18.18$, $\beta = 2.16$, present in 46/8,287 genomes), while oligopeptides associated with higher MIC captured multiple substitutions within the AT-hook motif known to bind AT-rich sequences [44,45] (e.g., DQGSIVSQQHP, $-\log_{10}p = 10.85$, $\beta = 1.96$, present in 22/8,287 genomes). Substitutions in the AT-hook motif may disrupt the binding with the *whiB7* promoter sequence, while deletions upstream of *whiB7* have been shown to result in overexpression of WhiB7 [46]. WhiB7 is induced by antibiotic treatment and other stress conditions and activates its own expression along with other drug resistance genes, for example, *tap* and *erm* [45]. Variants in and upstream of another *whiB*-like transcriptional regulator, *whiB6*, were previously found to be associated with resistance to ethionamide [19,47], capreomycin, amikacin, kanamycin, and ethambutol [22,23]. WhiB7 has been implicated in cross-resistance to multiple drugs, including macrolides, tetracyclines, and aminoglycosides [45,46]; however, activation of WhiB7 is not induced by all antibiotics, for example, isoniazid [43]. Interestingly, oligopeptides and oligonucleotides in or upstream of *whiB7* were not found to be significantly associated with any of the other 12 antimicrobials. This could indicate yet another mechanism by which *whiB7* is involved in resistance to anti-TB drugs.

Levofloxacin. Oligopeptides in *tlyA* (*Rv1694*) were significantly associated with MIC of the fluoroquinolone levofloxacin (minimum *p*-value $p < 10^{-7.8}$, **S11 Fig**). *tlyA* encodes a methyltransferase that methylates ribosomal RNA. Variants in *tlyA*, including loss-of-function mutations, confer resistance to the aminoglycosides viomycin and capreomycin [48] by knocking out its methyltransferase activity [49].

An extremely low-frequency oligopeptide was associated with increased MIC and captured a 1-nucleotide adenosine insertion between positions 590 and 591 in codon 198 in a conserved region [50]. In contrast, oligopeptides containing the reference alleles in this region were associated with decreased MIC (e.g., GKGQVGPGGV**V**, $-\log_{10}p = 7.83$, $\beta = -1.86$, present in 7,281/7,300 genomes). The resulting frameshift likely mimics the knockout effect of deleting the 27 C-terminal residues of TlyA, which ablates methyltransferase activity [51]. While loss-of-function mutations conferring antimicrobial resistance were previously reported to specifically increase aminoglycoside MIC, fluoroquinolones were not investigated [52]. The signal in *tlyA* may therefore reveal genuine, previously unidentified cross-resistance.

Rifabutin. Oligonucleotides in *cysA2* (*Rv0815c*) and *cysA3* (*Rv3117*) were significantly associated with rifabutin MIC (minimum *p*-value $p < 10^{-7.7}$, **S12 Fig**). They encode identical proteins, which are putative uncharacterised thiosulfate: cyanide sulfurtransferases, known as rhodaneses, belonging to the essential sulfur assimilation pathway, secreted during infection [53]. No genome-wide significant signals associated specific oligopeptides or oligonucleotides with higher MIC. Significant oligonucleotides that aligned to *cysA2* and *cysA3* were associated with lower MIC. They captured 2 variants: a synonymous nucleotide substitution, a thymine at position 117 in codon 39, and a non-synonymous nucleotide substitution, a guanine at position 103 inducing amino acid substitution 35D (e.g., CATAT**G**ACCGTGACCA-TAT**T**GCCGGCGCGAT, $-\log_{10}p = 7.74$, $\beta = -2.65$, present in 9,396/9,418 genomes). These positions coincide with the rhodanese characteristic signature in the N-terminal region, important for rhodanese stability [54]. However, the mechanism of resistance against rifabutin remains to be elucidated.

New and repurposed drugs

Bedaquiline. Oligonucleotides situated in the region of overlap at the 3′ ends of *amiA2* (*Rv2363*) and *era* (*Rv2364c*) were significantly associated with bedaquiline MIC (minimum *p*-value $p < 10^{-10.5}$, S13 Fig). These genes encode an amidase and a GTPase, respectively, on opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in *era*, 7 to 19 nucleotides upstream of the stop codon, and 3′ noncoding in *amiA2*, 4 to 16 nucleotides downstream of the stop codon (e.g., CCCCAAACAGCT**T**GGCCGACTGGG**G**TTTTAG, $-\log_{10}p = 10.47$, $\beta = 1.26$, present in 7,919/8,009 genomes). The second additionally captures a variant that induces a non-synonymous guanine substitution at position 1,451 in *amiA2*, and is 3′ intergenic in *era*, 1 nucleotide downstream of the stop codon (e.g., CAAACAGCT**T**GGCCGACTGGG**G**TTTTAG**C**TC, $-\log_{10}p = 7.87$, $\beta = 0.88$, present in 7,898/8,009 genomes). Interestingly, AmiA2 has previously been identified at lower abundance in MDR compared to sensitive isolates [55], and Era (but not AmiA2) has been shown to be required for optimal growth of H37Rv [56]. These variants may therefore enhance tolerance to bedaquiline.

Clofazimine. Oligopeptides in *cyp142* (*Rv3518c*), which encodes a cytochrome P450 enzyme with substrates of cholesterol/cholest-4-en-3-one, were significantly associated with clofazimine MIC (minimum *p*-value $p < 10^{-12.2}$, S14 Fig). Oligopeptides associated with higher MIC captured the amino acid residue 176I (e.g., EDFQIT**I**DAFA, $-\log_{10}p = 7.99$, $\beta = 1.14$, present in 100/7,297 genomes). The association signal falls within the F α-helix of CYP142, which lines the entrance to the active site with largely hydrophobic residues, forming part of the substrate binding pocket [57,58]. Homology with CYP125 suggests that residue 176 captured by the GWAS is within 5 Å of the binding substrate [58]. The potential for cytochrome P450 enzymes as targets for anti-TB drugs has been highlighted [59]; CYP142 is inhibited by azole drugs [59] and has been found to form a tight complex with nitric oxide (NO) [60]. The anti-mycobacterial activity of clofazimine has been shown to produce reactive oxygen species [61]; therefore, the substitution identified by the GWAS may disrupt the binding of NO to CYP142. Methionine and isoleucine are both hydrophobic residues, so the mechanism for how this would disrupt binding is unknown.

Delamanid. Oligonucleotides in *pknH* (*Rv1266c*), which encodes a serine/threonine-protein kinase, were significantly associated with delamanid MIC (minimum *p*-value $p < 10^{-30.2}$, S15 Fig). Delamanid is a prodrug activated by deazaflavin-dependent nitroreductase that inhibits cell wall synthesis. PknH phosphorylates the adjacent gene product EmbR [62], enhancing its binding of the promoter regions of the *embCAB* operon [63]. Mutations in *embAB* are responsible for ethambutol resistance [29]. The peak GWAS signal localized to the C-terminal periplasmic domain of PknH [62]. Oligonucleotides below our MAF threshold captured extremely low-frequency triplet deletions of either ACG at nucleotides 1645–7 or GAC at nucleotides 1644–6. In contrast, oligonucleotides containing the reference alleles in this region were associated with decreased MIC (e.g., CAAGACGGTCACCGTCACGAATAAGGCCAAG, $-\log_{10}p = 30.21$, $\beta = -3.29$, present in 7,555/7,564 genomes). These variants likely disrupt intramolecular disulphide binding linking the 2 highly conserved alpha helices that form the V-shaped cleft of the C-terminal sensor domain [64]. Since NO is released upon activation of DLM, and deletion of PknH alters sensitivity to nitrosative and oxidative stresses [65], these rare variants may alter tolerance to delamanid mediated by NO.

Linezolid. Oligonucleotides in *vapB20* (*Rv2550c*) were significantly associated with linezolid MIC (minimum *p*-value $p < 10^{-8.6}$, S16 Fig). VapB20 is an antitoxin cotranscribed with its complementary toxin VapC20 [66]. The latter modifies 23S rRNA [67], the target of linezolid that inhibits protein synthesis by competitively binding 23S rRNA. The peak signal in

vapB20 occurred just upstream of the promotor and VapB20 binding sites, 21 nucleotides upstream of the −35 region [67]. Oligonucleotides below our MAF threshold associated with increased MIC shared a cytosine 33 nucleotides upstream of *vapB20*, replacing the reference nucleotide thymine that was associated with decreased MIC (e.g., GAATCGG**A**TGCTT GCCGCTGGCTGCCGAGTT, $-\log_{10}p = 8.60$, $\beta = -2.02$, present in 6,724/6,732 genomes). This substitution may derepress the toxin, which could interrupt linezolid binding by cleaving the sarcin-ricin loop of 23S rRNA.

Discussion

In this study, we tested oligopeptides and oligonucleotides for association with semiquantitative MIC measurements for 13 antimicrobials to identify novel resistance determinants. Analysing MIC rather than binary resistance phenotypes enabled identification of variants that cause subtle changes in MIC. This is important, on the one hand, because higher rifampicin and isoniazid MIC in sensitive isolates are associated with increased risk of relapse after treatment [68]. Conversely, low-level resistance among isolates resistant to rifampicin and isoniazid mediated by particular mutations may sometimes be overcome by increasing the drug dose, or replacing rifampicin with rifabutin, rather than changing to less desirable drugs with worse side effects [69–73]. The investigation of MIC was particularly effective at increasing sample heritability for the new and repurposed drugs.

The MICs were positively correlated between many drugs, particularly among first-line drugs (see Figure 4A of [78]). Consequently, many of the 10,228 isolates we studied were MDR and XDR. In GWAS, this generates artefactual cross-resistance, in which variants that cause resistance to one drug appear associated with other drugs to which they do not confer resistance. In practice, it is difficult to distinguish between associations that are causal versus artefactual without experimental evidence. Nevertheless, we found frequent evidence of artefactual cross-resistance: several genes and intergenic regions featured among the top 20 strongest signals of association to multiple drugs, including *rpoB* (12 drugs), *embB* (7), *fabG1* (7), *rrs* (6), *gyrA* (6), *katG* (6), *lprF*:*Rv1371* (6), *pncA* (5), *ethA* (4), *Rv3183*:*Rv3188* (4) *dxs2*:*Rv3382c* (4), *rpsL* (3), and *moaC3*:*Rv3327* (3). Among previously catalogued variants, we observed that the estimated effect sizes were usually larger in magnitude for significant true associations than significant artefactual associations (**S4 Fig**). In future GWAS, this relationship could help tease apart true versus artefactual associations when an uncatalogued variant is associated with multiple drugs.

We focused on variants in the top 20 most significant genes identified by GWAS for each of the 13 drugs, classifying significant oligopeptides and oligonucleotides according to whether the variants they tagged were previously catalogued among known resistance determinants or not. While the interpretation of oligopeptides and oligonucleotides required manual curation to determine the underlying variants they tagged, the approach had the advantage of avoiding reference-based variant calling that can miss important signals, particularly at difficult-to-map regions. Exhausitvely and concisely calling variants relative to a reference genome in a large dataset of over 10,000 genomes is not trivial [80]. An additional benefit of analysing oligopeptides (versus oligonucleotides) is to pool signals of association across nucleotides that encode the same amino acid. Since we expected most resistance-conferring mutations to affect coding sequences, we anticipated that this should improve power and interpretability for most associations. A combination of strong linkage disequilbrium and lower diversity at the protein versus nucleotide level meant that, despite testing all 6 reading frames, the Bonferroni thresholds for the oligopeptide analyses were 2-fold less stringent than for the oligonucleotide analyses.

For 8/13 drugs with previously catalogued resistance determinants, the most significant GWAS signal in CRyPTIC was a previously catalogued variant. Among the uncatalogued

variants there are promising signals of association, including in the *Rv1218c* promoter binding site of the transcriptional repressor *Rv1219c* (associated with MIC for isoniazid) upstream of the *vapBC20* operon that cleaves 23S rRNA (linezolid) and in the region encoding a helix lining the active site of *cyp142* (clofazimine). These variants would benefit from further investigation via replication studies in independent populations, experimental exploration of proposed resistance mechanisms, or both.

We elected to classify significant variants as catalogued versus uncatalogued, rather than known versus novel, for several reasons. The catalogues represent a concrete, preexisting knowledgebase collated by expert groups for use in a clinical context [11,17]. We chose [11,17] as they were the most recent and up-to-date catalogues available for the drugs we investigated. The inclusion criteria for variants to be considered catalogued are therefore stringent; it follows that a class of variants exist that have been reported in the literature but not assimilated into the catalogues [11,17]. The literature is vast and heterogenous, with evidence originating from molecular, clinical, and GWAS. Inevitably, some uncatalogued variants in the literature will be false positives, while others will be real but did not meet the standard of evidence or clinical relevance for cataloguing. Evidence from CRyPTIC that supports uncatalogued variants in the latter group is of equal or greater value than the discovery of completely novel variants, because it contributes to a body of independent data supporting their involvement. This seems to apply to the most significant signals of association for BDQ, CLO, DLM, LZD, and RFB (Table 1), none of which appeared in our reference catalogues, but all of which appear in credible reports in the literature (e.g., [85,86]). To take another example, *gyrB* did not appear in the catalogues we used for moxifloxacin [11,17]. Yet our rediscovery of *gyrB* 501D complements published reports associating the substitution with moxifloxacin resistance [74–76], strongly enhancing the evidence in favour of inclusion in future catalogues. Indeed, the recent WHO prediction catalogue, published after the completion of this study and which draws on the CRyPTIC data analysed here includes the E501D resistance-associated variant [77]. Moreover, of the 5 new genes added to the forthcoming WHO catalogue [77] but not featuring in the catalogues [11,17] used here—*eis* (amikacin), *ethA* (ethionamide), *inhA* (ethionamide), *rplC* (linezolid), *gyrB* (moxifloxacin)—we identify all as containing significant variants by GWAS except one, *eis* (amikacin).

The combination of a very large dataset exceeding 10,000 isolates and quantification of resistance via MIC enabled the CRyPTIC study to attribute a large proportion of fine-grained variability in antimicrobial resistance in *M*. *tuberculosis* to genetic variation. Compared to a parallel analysis of binary resistance phenotypes in the same samples, we observed an increase in sample heritability of 26.1% to 37.1% for the new and repurposed drugs bedaquiline, clofazimine, delamanid, and linezolid. The improvement was most striking for delamanid, whose heritability was not significantly different to zero for the binary resistance phenotype. In the case of delamanid, the MIC analysis detected a surprisingly large number of signals. Since few isolates were strongly delaminid resistant, this indicates we were picking apart fine-grained differences in MIC, a phenotype which may be more polygenic than binary resistance/sensitivity. In contrast, the scope for improvement was marginal for the better-studied drugs isoniazid and rifampicin, where MIC heritabilities of 94.6% to 94.9% were achieved. This demonstrates the ability of additive genetic variation to explain almost all the phenotypic variability in MIC for these drugs. Nevertheless, we were still able to find uncatalogued hits for these drugs. The very large sample size also contributed to increased sample heritability compared to previous pioneering studies. Compared to Farhat and colleagues [22] who estimated the heritability of MIC phenotypes in 1,452 isolates, we observed increases in heritability of 2.0% (kanamycin), 3.3% (amikacin), 14.0% (isoniazid), 10.8% (rifampicin), 11.2% (ethambutol), and 19.4% (moxifloxacin), although these sample heritabilities depend on the idiosyncrasies of sampling.

Furthermore, many of the uncatalogued signals we report here as significant detected rare variants at below 1% MAF, underlining the ability of very large-scale studies to improve our understanding of antimicrobial resistance not only quantitatively, but to tap otherwise unseen rare variants that reveal new candidate resistance mechanisms.

Materials and methods

Ethics statement

Approval for CRyPTIC study was obtained by Taiwan Centers for Disease Control IRB No. 106209, University of KwaZulu Natal Biomedical Research Ethics Committee (UKZN BREC) (reference BE022/13) and University of Liverpool Central University Research Ethics Committees (reference 2286), Institutional Research Ethics Committee (IREC) of The Foundation for Medical Research, Mumbai (Ref nos. FMR/IEC/TB/01a/2015 and FMR/IEC/TB/01b/2015), Institutional Review Board of P.D. Hinduja Hospital and Medical Research Centre, Mumbai (Ref no. 915-15-CR [MRC]), scientific committee of the Adolfo Lutz Institute (CTC-IAL 47-J/2017), and in the Ethics Committee (CAAE: 81452517.1.0000.0059) and Ethics Committee review by Universidad Peruana Cayetano Heredia (Lima, Peru) and LSHTM (London, UK).

Sampling frames

CRyPTIC collected isolates from 27 countries worldwide, oversampling for drug resistance, as described in detail in [31]. Clinical isolates were subcultured for 14 days before inoculation onto 1 of 2 CRyPTIC designed 96-well microtiter plates manufactured by Thermo Fisher. The first plate used (termed UKMYC5) contained doubling-dilution ranges for 14 different antibiotics, the second (UKMYC6) removed para-aminosalicylic acid due to poor results on the plate [30] and changed the concentration of some drugs. Para-aminosalicylic acid was therefore not included in the GWAS analyses. Phenotype measurements were determined to be high quality, and included in the GWAS analyses, if 3 independent methods (Vizion, AMyGDA, and BashTheBug) agreed on the value [31]. Sequencing pipelines differed slightly between the CRyPTIC sites, but all sequencing was performed using Illumina, providing an input of matched pair FASTQ files containing the short reads.

A total of 15,211 isolates were included in the initial CRyPTIC dataset with both genomes and phenotype measurements after passing genome quality control filters [31,78]; however, some plates were later removed due to problems identified at some laboratories with inoculating the plates [31]. Genomes were also excluded if they met any of the following criteria, determined by removing samples at the outliers of the distributions: (i) no high-quality phenotypes for any drugs; (ii) total number of contigs $> 3{,}000$; (iii) total bases in contigs $< 3.5 \times 10^6$ or $> 5 \times 10^6$; (iv) number of unique oligonucleotides $< 3.5 \times 10^6$ or $> 5 \times 10^6$; and (v) sequencing read length not 150/151 bases long. This gave a GWAS dataset of 10,422 genomes used to create the variant presence/absence matrices. We used Mykrobe [78–80] to identify *Mycobacterium* genomes not belonging to lineages 1 to 4 or representing mixtures of lineages. This led to the exclusion of 193 genomes, which were removed from GWAS by setting the phenotypes to NA. The number of genomes with a high-quality phenotype for at least 1 of the 13 drugs was therefore 10,228. Of these, 533 were lineage 1; 3,581 were lineage 2; 805 were lineage 3; and 5,309 were lineage 4. Due to rigorous quality control described above, only samples with high-quality phenotypes were tested for each drug, resulting in a range of 6,388 to 9,418 genomes used in each GWAS. The $\log_2$ MIC phenotypes used in the study are available in **S2 Data**, and the data is publicly available for download via an FTP site at the European Bioinformatics Institute (https://ftp.ebi.ac.uk/pub/databases/cryptic/release_june2022/pubs/gwas2022).

Phylogenetic inference

A pairwise distance matrix was constructed for the full CRyPTIC dataset based on variant calls [78]. For visualisation of the dataset, a neighbour-joining tree was built from the distance matrix using the ape package in R and subset to the GWAS dataset. Negative branch lengths were set to zero, and the length was added to the adjacent branch. The branch lengths were square rooted and the tree annotated by lineages assigned by Mykrobe [79].

Oligonucleotide/oligopeptide counting

To capture SNP-based variation, indels, and combinations of SNPs and indels, we pursued oligonucleotide and oligopeptide-based approaches, focusing primarily on oligopeptides. Where helpful for clarifying results, we interpreted significant associations using oligonucleotides. Sequence reads were assembled de novo using Velvet Optimiser [81] with a starting lower hash value of half the read length, and a higher hash value of the read length minus one; if these were even numbers, they were lowered by one. If the total sequence length of the reads in the FASTQ file was greater than 1×10^9, then the reads were randomly subsampled prior to assembly down to a sequence length of 1×10^9, which is around 227× mean coverage. For the oligopeptide analysis, each assembly contig was translated into the 6 possible reading frames in order to be agnostic to the correct reading frame. A total of 11 amino acid long oligopeptides were counted in a 1 amino acid sliding window from these translated contigs. The 31-bp nucleotide oligonucleotides were also counted from the assembled contigs using dsk [82]. For both oligonucleotide and oligopeptide analyses, a unique set of variants across the dataset was created, with the presence or absence of each unique variant determined per genome. An oligonucleotide/oligopeptide was counted as present within a genome if it was present at least once. This resulted in 60,103,864 oligopeptides and 34,669,796 oligonucleotides. Of these, 10,510,261 oligopeptides and 5,530,210 oligonucleotides were variably present in the GWAS dataset of 10,228 genomes.

Oligonucleotide/oligopeptide alignment

We used the surrounding context of the contigs that the oligopeptides/oligonucleotides were identified in to assist with their alignment. First, we aligned the contigs of each genome to the H37Rv reference genome [83] using nucmer [84], keeping alignments above 90% identity, assigning an H37Rv position to each base in the contig. Version 3 of the H37Rv strain (NC_000962.3) was used as the reference genome throughout the analysis. All numbering refers to the start positions in the H37Rv version 3 GenBank file. This gave a position for each oligonucleotide identified in the contigs, and after translating the 6 possible reading frames of the contig, each oligopeptide too. Each oligonucleotide/oligopeptide was assigned a gene or intergenic region (IR) or both in each genome. These variant/gene combinations were then merged across all genomes into unique variant/gene combinations, where a variant could be assigned to multiple genes or intergenic regions. Variant/gene combinations were then kept if seen in 5 or more genomes. In some specific regions where significant oligonucleotides or oligopeptides appeared to be capturing an invariant region, a threshold of just 1 genome was used to visualise low-frequency variants in the region. This was used only for interpretation of the signal in the region and not for the main analyses. To improve alignment for the most significant genes and intergenic regions, all oligonucleotides/oligopeptides in the gene/IR plus those that aligned to a gene/IR within 1 kb were realigned to the region using BLAST. Alignments were kept if above 70% identity, recalculated along the whole length of the oligonucleotide/oligopeptide assuming the whole oligonucleotide/oligopeptide aligned. Oligopeptides were aligned to all 6 possible reading frames and only the correct reading frame was

interpreted. An oligonucleotide/oligopeptide was interpreted as unaligned if it did not align to any of the 6 possible reading frames. A region was determined to be significant if it contained significant oligopeptides above an MAF of 0.1% that were assigned to the region that also aligned using BLAST. If no significant oligopeptides aligned to the correct reading frame of a protein, or if the significant region was intergenic, then oligonucleotides were assessed.

Covariates

Isolates were sampled from 9 sites and MICs were measured on 2 versions of the quantitative microtiter plate assays, UKMYC5 and UKMYC6 [31]. UKMYC6 contained adjusted concentrations for some drugs. Therefore, in order to account for possible batch effects, we controlled for site plus plate type in the LMM by coding them as binary variables. These plus an intercept were included as covariates in the GWAS analyses.

Testing for locus effects

We performed association testing using LMM analyses implemented in the software GEMMA to control for population structure [32]. Significance was calculated using likelihood ratio tests. We computed the relatedness matrix from the presence/absence matrix using Java code that calculates the centred relatedness matrix. GEMMA was run using no MAF cutoff to include all variants. When assessing the most significant regions for each drug, we excluded oligopeptides below 0.1% MAF. To understand the full signal at these regions, oligopeptides and nucleotides were visualised in alignment figures to interpret the variants captured. When assessing the gene highlighted for each drug, we assessed the LD (r^2) of the most significant oligopeptide or nucleotide in the gene with all other top oligopeptides or nucleotides for the top 20 genes for the drug. The top variants in the genes noted were not in high LD with known causal variants, in some cases they were in LD with other top 20 gene hits that were less significant.

Correcting for multiple testing

Multiple testing was accounted for by applying a Bonferroni correction calculated for each drug. The unit of correction for all studies was the number of unique "phylopatterns," i.e., the number of unique partitions of individuals according to variant presence/absence for the phenotype tested. An oligopeptide/oligonucleotide was considered to be significant if its p-value was smaller than α/n_p, where we took $\alpha = 0.05$ to be the genome-wide false positive rate (i.e., family-wide error rate (FWER)) and n_p to be the number of unique phylopatterns above 0.1% MAF in the genomes tested for the particular drug. The $-\log_{10}p$ significance thresholds for the oligopeptide analyses were: 7.69 (amikacin, kanamycin), 7.65 (bedaquiline), 7.64 (clofazimine, levofloxacin), 7.67 (delamanid, ethionamide), 7.62 (ethambutol, linezolid), 7.70 (isoniazid), 7.60 (moxifloxacin), 7.71 (rifabutin), and 7.68 (rifampicin). The $-\log_{10}p$ significance thresholds for the oligonucleotide analyses were: 7.38 (amikacin, kanamycin), 7.34 (bedaquiline, clofazimine, levofloxacin), 7.36 (delamanid, ethionamide), 7.32 (ethambutol), 7.39 (isoniazid, rifabutin), 7.33 (linezolid), 7.31 (moxifloxacin), and 7.37 (rifampicin).

Estimating sample heritability

Sample heritability is the proportion of the phenotypic variation that can be explained by the bacterial genotype assuming additive effects. This was estimated using the LMM null model in GEMMA [32] from the presence versus absence matrices for both oligopeptides and oligonucleotides separately. Sample heritability was estimated for the MIC phenotype as well as for the binary sensitive versus resistant phenotype. The binary phenotypes were determined using the

ECOFF, defined as the MIC that encompasses 99% of wild-type isolates [31], all those below the ECOFF were considered susceptible, and those above the ECOFF were considered to be resistant.

Author contributions

Conceptualisation: Camilla Rodrigues, David Moore, Derrick W. Crook, Daniela M. Cirillo, Philip W Fowler, Zamin Iqbal, Nazir A. Ismail, Nerges Mistry, Stefan Niemann, Tim E.A. Peto, Guy Thwaites, A. Sarah Walker, Timothy M Walker, Daniel J. Wilson

Methodology: Sarah G. Earle, Daniel J. Wilson, Clara Grazian, A Sarah Walker

Software: Sarah G. Earle, Daniel J. Wilson

Formal analysis: Sarah G. Earle, Daniel J. Wilson, Clara Grazian, A Sarah Walker

Investigation: The CRyPTIC Consortium

Resources: The CRyPTIC Consortium

Data curation: Martin Hunt, Jeff Knaggs, Zamin Iqbal, Philip W Fowler, Zamin Iqbal

Writing – original draft preparation: Sarah G. Earle, Daniel J. Wilson, A Sarah Walker, Philip W Fowler

Writing – review & editing: The CRyPTIC Consortium

Visualization: Sarah G. Earle

Supervision: Daniela M Cirillo, Derrick W. Crook, Tim E.A. Peto, Daniel J. Wilson, A Sarah Walker, Zamin Iqbal, Philip W Fowler

Project administration: The CRyPTIC Consortium;

Funding acquisition: Camilla Rodrigues, David Moore, Derrick W. Crook, , Daniela M. Cirillo, Zamin Iqbal, Nazir A. Ismail, Nerges Mistry, Stefan Niemann, Tim E.A. Peto, Guy Thwaites, A. Sarah Walker, Timothy M Walker, Daniel J. Wilson

Supporting information

S1 Data. The interpretation of oligopeptides and oligonucleotides required manual curation to determine the underlying variants they tagged; the most significant oligopeptide or oligonucleotide for each allele captured by the significant signals are described here.
(XLSX)

S2 Data. The sample identifiers, log2 MIC phenotypes, and European Nucleotide Archive accession numbers of the genomes analysed in this study.
(XLSX)

S1 Acknowledgements. CRyPTIC Consortium memberlist.
(DOCX)

S1 Fig. Distributions of the log2 MIC measurements for all 13 drugs in the GWAS analyses, AMI, BDQ, CFZ, DLM, EMB, ETH, INH, KAN, LEV, LZD, MXF, RFB, and RIF.
The red dashed line indicates the ECOFF, measurements to the left of the ECOFF are considered sensitive, and those to the right are considered resistant. AMI, amikacin; BDQ, bedaquiline; CFZ, clofazimine; DLM, delamanid; ECOFF, epidemiological cutoff; EMB, ethambutol; ETH, ethionamide; GWAS, genome-wide association studies; INH, isoniazid; KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MIC, minimum inhibitory concentration; MXF, moxifloxacin; RFB, rifabutin; RIF, rifampicin.
(PDF)

S2 Fig. Oligopeptide and oligonucleotide sample heritability estimates for binary resistant vs. sensitive phenotypes compared to semiquantitiative MIC phenotypes. Sample heritability estimates and 95% CIs are shown for the 13 drugs, DLM, CFZ, LZD), BDQ, MXF, LEV, KAN, EMB, ETH, AMI, RIF, INH, and RFB. When estimating heritability of the same phenotype, the oligopeptide and oligonucleotide estimates are very similar. AMI, amikacin; BDQ, bedaquiline; CI, confidence interval; CFZ, clofazimine; DLM, delamanid; EMB, ethambutol; ETH, ethionamide; INH, isoniazid; KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MIC, minimum inhibitory concentration; MXF, moxifloxacin; RFB, rifabutin; RIF, rifampicin.
(PDF)

S3 Fig. QQ plots for the oligopeptide analyses, part A. Comparing the empirical distribution of *p*-values to the expected distribution under the null hypothesis for the drugs AMI, BDQ, CFZ, DLM, EMB, ETH, INH, and KAN. Oligopeptides in the orange (MAF $< 0.1\%$) were not initially analysed, only used for signal interpretation. AMI, amikacin; BDQ, bedaquiline; CFZ, clofazimine; DLM, delamanid; EMB, ethambutol; ETH, ethionamide; INH, isoniazid; KAN, kanamycin; MAF, minor allele frequency.
(PDF)

S4 Fig. QQ plots for the oligopeptide analyses, part B. Comparing the empirical distribution of *p*-values to the expected distribution under the null hypothesis for the KAN, LEV, LZD, MXF, RFB, and RIF. Oligopeptides in the orange (MAF $< 0.1\%$) were not initially analysed, only used for signal interpretation. KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MAF, minor allele frequency; MXF, moxifloxacin; RFB, rifabutin; RIF, rifampicin.
(PDF)

S5 Fig. Effect size (beta) estimates and −log10 *p*-values for all significant oligopeptide variants for each drug, AMI, BDQ, CFZ, DLM, EMB, ETH, INH, KAN, LEV, LZD, MXF, RFB, and RIF. For many of the drugs, the most significant oligopeptides were associated with lower MIC. AMI, amikacin; BDQ, bedaquiline; CFZ, clofazimine; DLM, delamanid; EMB, ethambutol; ETH, ethionamide; INH, isoniazid; KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MIC, minimum inhibitory concentration; MXF, moxifloxacin; RFB, rifabutin; RIF, rifampicin.
(PDF)

S6 Fig. Significant oligopeptide (rpoB, katG, gyrA, embB) and oligonucleotide (rrs) effect size (beta) estimates for known resistance genes plus the flanking 33 amino acids (oligopeptides) or 100 bases (oligonucelotides). On the left, the beta estimates are shown for all significant oligopeptides for the drugs the gene is causal for, on the right, the beta estimates are shown for the same gene, but for the drugs they are artefactually associated to. For many drugs, the beta estimate is lower when the gene is significant due to artefactual cross-resistance. Drug name abbreviations are as follows: AMI, BDQ, CFZ, DLM, EMB, ETH, INH, KAN, LEV, LZD, MXF, RFB, and RIF. AMI, amikacin; BDQ, bedaquiline; CFZ, clofazimine; DLM, delamanid; EMB, ethambutol; ETH, ethionamide; INH, isoniazid; KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MXF, moxifloxacin; RFB, rifabutin; RIF, rifampicin.
(PDF)

S7 Fig. Variants in *spoU* associated with EMB and RIF MIC. Manhattan plots showing the oligopeptide association results for the *spoU* coding region **A** ethambutol and **B** rifampicin, and oligonucleotide alignment plots showing close-ups of the significant region just downstream of *spoU* for **C** ethambutol and **D** rifampicin. The black dashed lines indicate the

Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *spoU*. Oligopeptides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the oligonucleotide alignment plots, the H37Rv reference codons are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligonucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligonucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligonucleotides are coloured by their amino acid residue at all variant positions. Oligonucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. The *spoU* stop codon is highlighted in red in the alignment plots. EMB, ethambutol; MAF, minor allele frequency; MIC, minimum inhibitory concentration; RIF, rifampicin.
(PDF)

S8 Fig. Variants in *Rv1219c* associated with isoniazid MIC. Manhattan plots showing the association results for the *Rv1219c* coding region for the **A** oligopeptides and **B** oligonucleotides, and oligopeptide alignment plots showing close-ups of the significant region in *Rv1219c* for **C** oligopeptides present in 5 or more genomes in the full GWAS dataset and **D** oligopeptides present in at least 1 genome in the full GWAS dataset. The black dashed lines indicate the Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *Rv1219c*. Oligopeptides and nucleotides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the oligopeptide alignment plots, the H37Rv reference codons are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligopeptides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligopeptides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligopeptides are coloured by their amino acid residue at all variant positions. Oligopeptides and nucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. GWAS, genome-wide association studies; MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S9 Fig. Variants in *PPE42* associated with AMI and KAN MIC. Manhattan plots showing the oligopeptide association results for the *PPE42* coding region **A** amikacin and **B** kanamycin, and oligopeptide alignment plots showing close-ups of the significant region in *PPE42* for **C** amikacin and **D** kanamycin. Black dashed lines indicate the Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *PPE42*. Oligopeptides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the oligopeptide alignment plots, the H37Rv reference codons are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligopeptides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligopeptides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligopeptides are coloured by their amino acid residue at all variant positions. Oligopeptides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. AMI, amikacin; KAN, kanamycin; MAF, minor allele frequency;

MIC, minimum inhibitory concentration.
(PDF)

S10 Fig. Variants in and upstream of *whiB7* associated with ethionamide MIC. Manhattan plots showing the association results for *whiB7* for the **A** oligopeptides and **B** oligonucleotides, and alignment plots showing close-ups of significant regions in *whib7* for **C** oligopeptides in the C-terminal end of the coding region and **D** oligonucleotides in the upstream intergenic region. The black dashed lines indicate the Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *whiB7*. Oligopeptides and nucleotides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the alignment plots, the H37Rv reference codons are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligopeptides and nucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligopeptides and nucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligopeptides and nucleotides are coloured by their allele at all variant positions. Oligopeptides and nucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S11 Fig. Variants in *tlyA* associated with levofloxacin MIC. Manhattan plots showing the association results for the *tlyA* coding region for the **A** oligopeptides and **B** oligonucleotides, and alignment plots showing close-ups of the significant region in *tlyA* for the **C** oligopeptides and **D** oligonucleotides. The black dashed lines indicate the Bonferroni-corrected significance thresholds; no oligonucleotides were significant in this region. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *tlyA*. Oligopeptides and nucleotides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the alignment plots, the H37Rv reference alleles are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligopeptides and nucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligopeptides and nucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligopeptides and nucleotides are coloured by their allele at all variant positions. Oligopeptides and nucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S12 Fig. Variants in *cysA2* and *cysA3* associated with rifabutin MIC. Manhattan plots showing the association results for the coding region for **A** *cysA2* and **B** cysA3, and oligonucleotide alignment plots showing close-ups of the significant region for **C** *cysA2* and **D** *cysA3*. The black dashed lines indicate the Bonferroni-corrected significance thresholds. The significant oligonucleotides that align to *cysA2* and *cysA3* are the same. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *cysA2* or *cysA3*. Oligopeptides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plot. In the oligonucleotide alignment plots, the H37Rv reference alleles are shown at the bottom of the figure, grey for an invariant site,

coloured at variant site positions. The oligonucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligonucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligonucleotides are coloured by their allele at all variant positions. Oligonucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. The region that encodes the rhodanese characteristic signature in the N-terminal region is highlighted in red in the alignment plots. MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S13 Fig. Variants in *amiA2* and *era* associated with bedaquiline MIC. Manhattan plots showing the association results for the *amiA2/era* coding region for the **A** oligopeptides and **B** oligonucleotides, and oligonucleotide alignment plots showing close-ups of the significant region in *amiA2/era* in the correct reading frame for **C** *amiA2* and **D** *era*. The black dashed lines indicate the Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *amiA2*. Oligopeptides and nucleotides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the oligonucleotide alignment plots, the H37Rv reference alleles are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligonucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligonucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligonucleotides are coloured by their allele at all variant positions. Oligonucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S14 Fig. Variants in *cyp142* associated with clofazimine MIC. Manhattan plots showing the association results for the *cyp142* coding region for the **A** oligopeptides and **B** oligonucleotides, and alignment plots showing close-ups of the significant region in *cyp142* for the **C** oligopeptides and **D** oligonucleotides. The black dashed lines indicate the Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, red for the correct reading frame for *cyp142*. Oligopeptides and nucleotides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the alignment plots, the H37Rv reference alleles are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligopeptides and nucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligopeptides and nucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligopeptides and nucleotides are coloured by their allele at all variant positions. Oligopeptides and nucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S15 Fig. Variants in *pknH* associated with delamanid MIC. Manhattan plots showing the association results for the *pknH* coding region for the **A** oligopeptides and **B** oligonucleotides, and alignment plots showing close-ups of the significant region in *pknH* for the **C**

oligopeptides and **D** oligonucleotides. The black dashed lines indicate the Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *pknH*. Oligopeptides and nucleotides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the alignment plots, the H37Rv reference alleles are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligopeptides and nucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligopeptides and nucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligopeptides and nucleotides are coloured by their allele at all variant positions. Oligopeptides and nucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S16 Fig. Variants in *vapB20* associated with linezolid MIC. Manhattan plots showing the association results for *vapB20* for the **A** oligopeptides and **B** oligonucleotides, and **C** oligonucleotide alignment plot showing a close-up of the significant region just upstream of *vapB20*. The black dashed lines indicate the Bonferroni-corrected significance thresholds. In the Manhattan plots, oligopeptides are coloured by the reading frame that they align to, black for the correct reading frame for *amiA2*. Oligopeptides and nucleotides assigned to the region but did not align using BLAST are shown in grey on the right-hand side of the plots. In the oligonucleotide alignment plot, the H37Rv reference alleles are shown at the bottom of the figure, grey for an invariant site, coloured at variant site positions. The oligonucleotides that aligned to the region are plotted from least significant at the bottom to most significant at the top. The background colour of the oligonucleotides represents the direction of the b estimate, light grey when $b < 0$ (associated with lower MIC), dark grey when $b > 0$ (associated with higher MIC). Oligonucleotides are coloured by their allele at all variant positions. Oligopeptides and nucleotides below the MAF threshold and not included in the analysis, but visualised here for signal interpretation, are marked by *s. MAF, minor allele frequency; MIC, minimum inhibitory concentration.
(PDF)

S1 Table. Oligopeptide and oligonucleotide sample heritability estimates for binary resistant vs. sensitive phenotypes compared to semiquantitiative MIC phenotypes. Sample heritability estimates and 95% CIs are shown for the 13 drugs. CI, confidence interval; MIC, minimum inhibitory concentration.
(PDF)

Acknowledgments

We thank Faisal Masood Khanzada and Alamdar Hussain Rizvi (NTRL, Islamabad, Pakistan), Angela Starks and James Posey (Centers for Disease Control and Prevention, Atlanta, USA), Juan Carlos Toro and Solomon Ghebremichael (Public Health Agency of Sweden, Solna, Sweden), and Iñaki Comas and Álvaro Chiner-Oms (Instituto de Biología Integrativa de Sistemas, Valencia, Spain; CIBER en Epidemiología y Salud Pública, Valencia, Spain; Instituto de Biomedicina de Valencia, IBV-CSIC, Valencia, Spain). Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre.

References

1. World Health Organization. Global Tuberculosis. Report 2020.
2. Shah NS, Auld SC, Brust JCM, Mathema B, Ismail N, Moodley P, et al. Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. N Engl J Med. 2017; 376(3):243–253. https://doi.org/10.1056/NEJMoa1604544 PMID: 28099825
3. World Health Organization. WHO Consolidated Guidelines on Tuberculosis, Module 4: Treatment -. Drug-Resistant Tuberculosis Treatment 2020.
4. World Health Organization. Rapid Communication: Key changes to the treatment of drug-resistant tuberculosis. 2019.
5. Kranzer K, Kalsdorf B, Heyckendorf J, Andres S, Merker M, Hofmann-Thiel S, et al. New World Health Organization Treatment Recommendations for Multidrug-Resistant Tuberculosis: Are We Well Enough Prepared? Am J Respir Crit Care Med. 2019; 200(4). https://doi.org/10.1164/rccm.201902-0260LE PMID: 31026398
6. Andres S, Merker M, Heyckendorf J, Kalsdorf B, Rumetshofer R, Indra A, et al. Bedaquiline-Resistant Tuberculosis: Dark Clouds on the Horizon. Am J Respir Crit Care Med. 2020; 201(12).
7. Polsfuss S, Hofmann-Thiel S, Merker M, Krieger D, Niemann S, Rüssmann H, et al. Emergence of Low-level Delamanid and Bedaquiline Resistance During Extremely Drug-resistant Tuberculosis Treatment. Clin Infect Dis. 2019; 69(7):1229–1231. https://doi.org/10.1093/cid/ciz074 PMID: 30933266
8. Islam M, Hameed H, Mugweru J, Chhotaray C, Wang C, Tan Y, et al. Drug resistance mechanisms and novel drug targets for tuberculosis therapy. J Genet Genomics. 2016; 44(1):21–37. https://doi.org/10.1016/j.jgg.2016.10.002 PMID: 28117224
9. Goossens S, Sampson S, Van Rie A. Mechanisms of Drug-Induced Tolerance in Mycobacterium tuberculosis. Clin Microbiol Rev. 2020; 34(1):e00141–e00120.
10. Pankhurst L, Del Ojo EC, Votintseva A, Walker T, Cole K, Davies J, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. Lancet Respir Med. 2016; 4(1):49–58. https://doi.org/10.1016/S2213-2600(15)00466-X PMID: 26669893
11. The CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018; 379(15):1403–1415. https://doi.org/10.1056/NEJMoa1800474 PMID: 30280646
12. Walker TM, Gibertoni Cruz AL, Tim E, Smith EG, Esmail H, Crook DW. Tuberculosis is changing. Lancet Infect Dis. 2017; 17(4):359–361. https://doi.org/10.1016/S1473-3099(17)30123-8 PMID: 28298254
13. Makhado NA, Matabane E, Faccin M, Pinçon C, Jouet A, Boutachkourt F, et al. Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study. Lancet Infect Dis. 2018; 18(12):1350–1359. https://doi.org/10.1016/S1473-3099(18)30496-1 PMID: 30342828
14. Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, et al. Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. N Engl J Med. 2010; 363(11):1005–1015. https://doi.org/10.1056/NEJMoa0907847 PMID: 20825313
15. Boehme C, Nicol M, Nabeta P, Michael J, Gotuzzo E, Tahirli R, et al. Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study. Lancet. 2011; 377(9776):1495–1505. https://doi.org/10.1016/S0140-6736(11)60438-8 PMID: 21507477
16. Sanchez-Padilla E, Merker M, Beckert P, Jochims F, Dlamini T, Kahn P, et al. Detection of Drug-Resistant Tuberculosis by Xpert MTB/RIF in Swaziland. N Engl J Med. 2015; 372(12):1181–1182. https://doi.org/10.1056/NEJMc1413930 PMID: 25785984
17. Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis. Eur Respir J. 2017; 50(6):1701354. https://doi.org/10.1183/13993003.01354-2017 PMID: 29284687
18. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013; 45(10):1183–1189. https://doi.org/10.1038/ng.2747 PMID: 23995135
19. Zhang H, Li D, Zhao L, Fleming J, Lin NWT, Liu Z, et al. Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013; 45(10):1255–1260. https://doi.org/10.1038/ng.2735 PMID: 23995137
20. Earle SG, Wu C, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016; 1(5):16041. https://doi.org/10.1038/nmicrobiol.2016.41 PMID: 27572646

21. Nair MB, Mallard K, Ali S, Abdallah AM, Alghamdi S, Alsomali M, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018; 50(2):307–316. https://doi.org/10.1038/s41588-017-0029-0 PMID: 29358649

22. Farhat M, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan C, et al. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. Nat Commun. 2019; 10(2128). https://doi.org/10.1038/s41467-019-10110-6 PMID: 31086182

23. Oppong YEA, Phelan J, Perdigão J, Machado D, Miranda A, Portugal I, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC Genom. 2019; 20(1):252. https://doi.org/10.1186/s12864-019-5615-3 PMID: 30922221

24. Farhat M, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, et al. Genetic Determinants of Drug Resistance in Mycobacterium tuberculosis and Their Diagnostic Value. Am J Respir Crit Care Med. 2016; 194(5):621–630. https://doi.org/10.1164/rccm.201510-2091OC PMID: 26910495

25. Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo EC, Bradley P, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study. Lancet Infect Dis. 2015; 15(10):1193–1202. https://doi.org/10.1016/S1473-3099(15)00062-6 PMID: 26116186

26. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010; 11(7):459–463. https://doi.org/10.1038/nrg2813 PMID: 20548291

27. World Health Organization. Technical report on critical concentrations for TB drug susceptibility testing of medicines used in the treatment of drug-resistant TB. 2018.

28. Schön T, Miotto P, Köser CU, Viveiros M, Böttger E, Cambau E. Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. Clin Microbiol Infect. 2017; 23 (3):154–160. https://doi.org/10.1016/j.cmi.2016.10.022 PMID: 27810467

29. Sreevatsan S, Stockbauer K, Pan X, Kreiswirth B, Moghazeh S, Jacobs WJ, et al. Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. Antimicrob Agents Chemother. 1997; 41(8):1677–1681. https://doi.org/10.1128/AAC.41.8.1677 PMID: 9257740

30. Rancoita P, Cugnata F, Gibertoni Cruz A, Borroni E, Hoosdally S, Walker T, et al. Validating a 14-Drug Microtiter Plate Containing Bedaquiline and Delamanid for Large-Scale Research Susceptibility Testing of Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2018; 62(9):e00344–e00318. https://doi.org/10.1128/AAC.00344-18 PMID: 29941636

31. The CRyPTIC Consortium. Epidemiological cutoff values for a 96-well broth microdilution plate for high-throughput research antibiotic susceptibility testing of M. tuberculosis. medRxiv https://doi.org/101101/2021022421252386 2021.

32. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44(7):821–824. https://doi.org/10.1038/ng.2310 PMID: 22706312

33. Pantel A, Petrella S, Veziris N, Brossier F, Bastian S, Jarlier V, et al. Extending the definition of the GyrB quinolone resistance-determining region in Mycobacterium tuberculosis DNA gyrase for assessing fluoroquinolone resistance in M. tuberculosis. Antimicrob Agents Chemother. 2012; 56(4):1990–1996. https://doi.org/10.1128/AAC.06272-11 PMID: 22290942

34. Blower TR, Williamson BH, Kerns RJ, Berger JM. Structure of tuberculosis quinolone–gyrase complex. Proc Natl Acad Sci U S A. 2016; 113(7):1706–1713.

35. Sharma G, Upadhyay S, Srilalitha M, Nandicoori V, Khosla S. The interaction of mycobacterial protein Rv2966c with host chromatin is mediated through non-CpG methylation and histone H3/H4 binding. Nucleic Acids Res. 2015; 43(8):3922–3937. https://doi.org/10.1093/nar/gkv261 PMID: 25824946

36. Lai YP, Ioerger T. Exploiting Homoplasy in Genome-Wide Association Studies to Enhance Identification of Antibiotic-Resistance Mutations in Bacterial Genomes. Evol Bioinform Online. 2020. https://doi.org/10.1177/1176934320944932 PMID: 32782426

37. Dixit A, Freschi L, Vargas R, Calderon R, Sacchettini J, Drobniewski F, et al. Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting. Sci Rep. 2019; 9(5602).

38. Kumar N, Radhakrishnan A, Wright C, Chou T, Lei H, Bolla J, et al. Crystal structure of the transcriptional regulator Rv1219c of Mycobacterium tuberculosis. Protein Sci. 2014; 23(4):423–432. https://doi.org/10.1002/pro.2424 PMID: 24424575

39. Wang K, Pei H, Huang B, Zhu X, Zhang J, Zhou B, et al. The expression of ABC efflux pump, Rv1217c-Rv1218c, and its association with multidrug resistance of Mycobacterium tuberculosis in China. Curr Microbiol. 2013; 66(3):222–226. https://doi.org/10.1007/s00284-012-0215-3 PMID: 23143285

40. Chakhaiyar P, Nagalakshmi Y, Aruna B, Murthy K, Katoch V, Hasnain S. Regions of high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608, show a

differential humoral response and a low T cell response in various categories of patients with tuberculosis. J Infect Dis. 2004; 190(7):1237–1244. https://doi.org/10.1086/423938 PMID: 15346333

41. Coler R, Day T, Ellis R, Piazza F, Beckmann A, Vergara J, et al. The TLR-4 agonist adjuvant, GLA-SE, improves magnitude and quality of immune responses elicited by the ID93 tuberculosis vaccine: first-in-human trial. NPJ Vaccines. 2018; 3(34). https://doi.org/10.1038/s41541-018-0057-5 PMID: 30210819

42. Bhattacharyya K, Nemaysh V, Joon M, Pratap R, Varma-Basil M, et al. Correlation of drug resistance with single nucleotide variations through genome analysis and experimental validation in a multi-drug resistant clinical isolate of M. tuberculosis. BMC Microbiol. 2020; 20(223).

43. Burian J, Ramón-García S, Sweet G, Gómez-Velasco A, Av-Gay Y, Thompson C. The mycobacterial transcriptional regulator whiB7 gene links redox homeostasis and intrinsic antibiotic resistance. J Biol Chem. 2012; 287(1):299–310. https://doi.org/10.1074/jbc.M111.302588 PMID: 22069311

44. Ramón-García S, Ng C, Jensen P, Dosanjh M, Burian J, Morris R, et al. WhiB7, an Fe-S-dependent transcription factor that activates species-specific repertoires of drug resistance determinants in actinobacteria. J Biol Chem. 2013; 288(48):34514–34528. https://doi.org/10.1074/jbc.M113.516385 PMID: 24126912

45. Morris R, Nguyen L, Gatfield J, Visconti K, Nguyen K, Schnappinger D, et al. Ancestral antibiotic resistance in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2005; 102(34):12200–12205. https://doi.org/10.1073/pnas.0505446102 PMID: 16103351

46. Reeves A, Campbell P, Sultana R, Malik S, Murray M, Plikaytis B, et al. Aminoglycoside cross-resistance in Mycobacterium tuberculosis due to mutations in the 5' untranslated region of whiB7. Antimicrob Agents Chemother. 2013; 57(4):1857–1865. https://doi.org/10.1128/AAC.02191-12 PMID: 23380727

47. Hicks N, Carey A, Yang J, Zhao Y, Fortune S. Bacterial Genome-Wide Association Identifies Novel Factors That Contribute to Ethionamide and Prothionamide Susceptibility in Mycobacterium tuberculosis. MBio. 2019; 10(2):e00616–e00619. https://doi.org/10.1128/mBio.00616-19 PMID: 31015328

48. Maus C, Plikaytis B, Shinnick T. Mutation of tlyA Confers Capreomycin Resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2005; 49 (2):571–577. https://doi.org/10.1128/AAC.49.2.571-577.2005 PMID: 15673735

49. Johansen S, Maus C, Plikaytis B, Douthwaite S. Capreomycin Binds across the Ribosomal Subunit Interface Using tlyA-Encoded 2′-O-Methylations in 16S and 23S rRNAs. Mol Cell. 2006; 23(2):173–182. https://doi.org/10.1016/j.molcel.2006.05.044 PMID: 16857584

50. Arenas NE, Salazar LM, Soto CY, Vizcaíno C, Patarroyo ME, Patarroyo MA, et al. Molecular modeling and in silico characterization of Mycobacterium tuberculosis TlyA: Possible misannotation of this tubercle bacilli-hemolysin. BMC Struct Biol. 2011; 11(16). https://doi.org/10.1186/1472-6807-11-16 PMID: 21443791

51. Monshupanee T, Johansen SK, Dahlberg AE, Douthwaite S. Capreomycin susceptibility is increased by TlyA-directed 2′-O-methylation on both ribosomal subunits. Mol Microbiol. 2012; 85:1194–1203. https://doi.org/10.1111/j.1365-2958.2012.08168.x PMID: 22779429

52. Zhao J, Wei W, Yan H, Zhou Y, Li Z, Chen Y, et al. Assessing capreomycin resistance on tlyA deficient and point mutation (G695A) Mycobacterium tuberculosis strains using multi-omics analysis. Int J Med Microbiol. 2019; 309(7). https://doi.org/10.1016/j.ijmm.2019.06.003 PMID: 31279617

53. Meza AN, Cambui CCN, Moreno ACR, Fessel MR, Balan A. Mycobacterium tuberculosis CysA2 is a dual sulfurtransferase with activity against thiosulfate and 3-mercaptopyruvate and interacts with mammalian cells. Sci Rep. 2019; 9(16791). https://doi.org/10.1038/s41598-019-53069-6 PMID: 31727914

54. Cipollone R, Ascenzi P, Visca P. Common themes and variations in the rhodanese superfamily. IUBMB Life. 2007; 59:51–59. https://doi.org/10.1080/15216540701206859 PMID: 17454295

55. Phong T, Ha do T, Volker U, Hammer E. Using a Label Free Quantitative Proteomics Approach to Identify Changes in Protein Abundance in Multidrug-Resistant Mycobacterium tuberculosis. Indian. J Microbiol. 2015; 55(2):219–230. https://doi.org/10.1007/s12088-015-0511-2 PMID: 25805910

56. Sassetti CM, Boyd DH, Rubin EJ. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 2003; 48:77–84. https://doi.org/10.1046/j.1365-2958.2003.03425.x PMID: 12657046

57. Driscoll M, McLean K, Levy C, Mast N, Pikuleva I, Lafite P, et al. Structural and biochemical characterization of Mycobacterium tuberculosis CYP142: evidence for multiple cholesterol 27-hydroxylase activities in a human pathogen. J Biol Chem. 2010; 285(49):38270–38282. https://doi.org/10.1074/jbc.M110.164293 PMID: 20889498

58. García-Fernández E, Frank D, Galán B, Kells P, Podust L, García J, et al. A highly conserved mycobacterial cholesterol catabolic pathway. Environ Microbiol. 2013; 15(8):2342–2359. https://doi.org/10.1111/1462-2920.12108 PMID: 23489718

59. Ortiz de Montellano P. Potential drug targets in the Mycobacterium tuberculosis cytochrome P450 system. J Inorg Biochem. 2018; 180(235–245). https://doi.org/10.1016/j.jinorgbio.2018.01.010 PMID: 29352597

60. Ouellet H, Lang J, Couture M, Ortiz de Montellano P. Reaction of Mycobacterium tuberculosis cytochrome P450 enzymes with nitric oxide. Biochemistry. 2009; 48(5):863–872. https://doi.org/10.1021/bi801595t PMID: 19146393

61. Yano T, Kassovska-Bratinova S, Teh J, Winkler J, Sullivan K, Isaacs A, et al. Reduction of clofazimine by mycobacterial type 2 NADH:quinone oxidoreductase: a pathway for the generation of bactericidal levels of reactive oxygen species. J Biol Chem. 2011; 286(12):10276–10287. https://doi.org/10.1074/jbc.M110.200501 PMID: 21193400

62. Molle V, Kremer L, Girard-Blanc C, Besra G, Cozzone A, Prost JF. An FHA Phosphoprotein Recognition Domain Mediates Protein EmbR Phosphorylation by PknH, a Ser/Thr Protein Kinase from Mycobacterium tuberculosis. Biochemistry. 2003; 42(51):15300–15309. https://doi.org/10.1021/bi035150b PMID: 14690440

63. Sharma K, Gupta M, Pathak M, Gupta N, Koul A, Sarangi S, et al. Transcriptional control of the mycobacterial embCAB operon by PknH through a regulatory protein, EmbR, in vivo. J Bacteriol. 2006; 188 (8):2936–2944. https://doi.org/10.1128/JB.188.8.2936-2944.2006 PMID: 16585755

64. Cavazos A, Prigozhin DM, Alber T. Structure of the Sensor Domain of Mycobacterium tuberculosis PknH Receptor Kinase Reveals a Conserved Binding Cleft. J Mol Biol. 2012; 422(4):488–494. https://doi.org/10.1016/j.jmb.2012.06.011 PMID: 22727744

65. Papavinasasundaram KG, Chan B, Chung JH, Colston MJ, Davis EO, Av-Gay Y. Deletion of the Mycobacterium tuberculosis pknH Gene Confers a Higher Bacillary Load during the Chronic Phase of Infection in BALB/c Mice. J Bacteriol. 2005; 187(16):5751–5760. https://doi.org/10.1128/JB.187.16.5751-5760.2005 PMID: 16077122

66. Deep A, Kaundal S, Agarwal S, Singh R, Thakur KG. Crystal structure of Mycobacterium tuberculosis VapC20 toxin and its interactions with cognate antitoxin, VapB20, suggest a model for toxin–antitoxin assembly. FEBS J. 2017; 284:4066–4082. https://doi.org/10.1111/febs.14289 PMID: 28986943

67. Winther K, Brodersen D, Brown A, Gerdes K. VapC20 of Mycobacterium tuberculosis cleaves the Sarcin–Ricin loop of 23S rRNA. Nat Commun. 2013; 4(2796). https://doi.org/10.1038/ncomms3796 PMID: 24225902

68. Colangeli R, Jedrey H, Kim S, Connell R, Ma S, Chippada Venkata UD, et al. Bacterial Factors That Predict Relapse after Tuberculosis Therapy. N Engl J Med. 2018; 379(9):823–833. https://doi.org/10.1056/NEJMoa1715849 PMID: 30157391

69. Walsh KF, Vilbrun SC, Souroutzidis A, Delva S, Joissaint G, Mathurin L, et al. Improved Outcomes With High-dose Isoniazid in Multidrug-resistant Tuberculosis Treatment in Haiti. Clin Infect Dis. 2019; 69 (4):717–719. https://doi.org/10.1093/cid/ciz039 PMID: 30698688

70. Dooley KE, Miyahara S, von Groote-Bidlingmaier F, Sun X, Hafner R, Rosenkranz SL, et al. Early Bactericidal Activity of Different Isoniazid Doses for Drug-Resistant Tuberculosis (INHindsight): A Randomized, Open-Label Clinical Trial. Am J Respir Crit Care Med. 2020; 201(11). https://doi.org/10.1164/rccm.201910-1960OC PMID: 31945300

71. Decroo T, de Jong BC, Piubello A, Souleymane MB, Lynen L, Van Deun A. High-Dose First-Line Treatment Regimen for Recurrent Rifampicin-Susceptible Tuberculosis. Am J Respir Crit Care Med. 2020; 201(12).

72. van Ingen J, Aarnoutse R, de Vries G, Boeree M, van Soolingen D. Low-level rifampicin-resistant Mycobacterium tuberculosis strains raise a new therapeutic challenge. Int J Tuberc Lung Dis. 2011; 15 (7):990–992. https://doi.org/10.5588/ijtld.10.0127 PMID: 21682979

73. Sirgel FA, Warren RM, Böttger EC, Klopper M, Victor TC, van Helden PD. The Rationale for Using Rifabutin in the Treatment of MDR and XDR Tuberculosis Outbreaks. PLoS ONE. 2013; 8(3):e59414. https://doi.org/10.1371/journal.pone.0059414 PMID: 23527189

74. Farhat MR, Jacobson KR, Franke MF, Kaur D, Sloutsky A, Mitnick CD, et al. Gyrase Mutations Are Associated with Variable Levels of Fluoroquinolone Resistance in Mycobacterium tuberculosis. J Clin Microbiol. 2016; 54(3). https://doi.org/10.1128/JCM.02775-15 PMID: 26763957

75. Disratthakit A, Prammananan T, Tribuddharat C, Thaipisuttikul I, Doi N, Leechawengwongs M, et al. Role of gyrB Mutations in Pre-extensively and Extensively Drug-Resistant Tuberculosis in Thai Clinical Isolates. Antimicrob Agents Chemother. 2016; 60(9):5189–5197. https://doi.org/10.1128/AAC.00539-16 PMID: 27297489

76. Malik S, Willby M, Sikes D, Tsodikov OV, Posey JE. New insights into fluoroquinolone resistance in Mycobacterium tuberculosis: functional genetic analysis of gyrA and gyrB mutations. PLoS ONE. 2012; 7(6):e39754. https://doi.org/10.1371/journal.pone.0039754 PMID: 22761889

77. World Health Organization. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. 2021. Report No.: ISBN: 9789240028173.

78. The CRyPTIC Consortium. A data compendium of M. tuberculosis antibiotic resistance. bioRxiv https://doi.org/10.1101/2021.09.14.460274 2021.

79. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat Commun. 2015; 6:10063. https://doi.org/10.1038/ncomms10063 PMID: 26686880

80. Hunt MH, Letcher B, Malone K, Nguyen G, Hall MB, Colquhoun RM, et al. Minos: graph adjudication and joint genotyping of cohorts of bacterial genomes. bioRxiv. 2021. https://doi.org/10.1101/2021.09.15.460475

81. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18(5):821–829. https://doi.org/10.1101/gr.074492.107 PMID: 18349386

82. Rizk G, Lavenier D, Chikhi R. DSK: k-mer counting with very low memory usage. Bioinformatics. 2013; 29(5):652–653. https://doi.org/10.1093/bioinformatics/btt020 PMID: 23325618

83. Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998; 393(6685):537–544. https://doi.org/10.1038/31159 PMID: 9634230

84. Kurtz S, Phillippy A, Delcher A, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004; 5(2):R12. https://doi.org/10.1186/gb-2004-5-2-r12 PMID: 14759262

85. Jamieson FB, Guthrie JL, Neemuchwala A, Lastovetska O, Melano RG, Mehaffy C. Profiling of rpoB Mutations and MICs for Rifampin and Rifabutin in Mycobacterium tuberculosis. J Clin Microbiol. 2014; 52(6):2157–2162. https://doi.org/10.1128/JCM.00691-14 PMID: 24740074

86. Kadura S, King N, Nakhoul M, Zhu H, Theron G, Köser CU, et al. Systematic review of mutations associated with resistance to the new and repurposed Mycobacterium tuberculosis drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid. J Antimicrob Chemother. 2020; 75:2031–2043. https://doi.org/10.1093/jac/dkaa136 PMID: 32361756