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

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...
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 catalogue variants of genetic origins [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
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 \textit{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 $\times$ 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)ton 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 board member and shareholder of Verus profit organization FIND. T.R. is a co-founder, 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. UCSD Conflict of Interest office has not received any financial support from the non-profit organization FIND. T.R. has not received any financial support from the non-profit organization FIND. T.R. has agreed to “donate all present and future financial benefits 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-wise 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 thecatalogues 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 literature (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., NSAGGSAKSGR, \( -\log_{10}\beta = 15.65 \), 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: DSAGGSAKSGR, \( -\log_{10}\beta = 13.47 \), \( \beta = 2.14 \), present in 7,278/7,300 genomes; 457V/461D: SELYVVEGDSA, \( -\log_{10}\beta = 12.51 \), \( \beta = 1.96 \), 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., NTDVQAIITAL, \( -\log_{10}\beta = 10.64 \), \( \beta = 1.86 \), present in 23/6,388 genomes). Oligopeptides capturing the reference allele at codons 499 and 501 were associated with lower MIC (e.g., NTEVQAIITAL, \( -\log_{10}\beta = 11.63 \), \( \beta = 1.33 \), present in 6,332/6,388 genomes). Amino acids 461 and 501 are at the interface between *gyrB* and the bound fluoroquinolone [34]. *gyrB* is included in the reference catalogues for predicting levofloxacin (including D461N) but not moxifloxacin resistance;
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., CAAACCAGCGGTATGGGCACAAAC- GAAGCTC, RIF: $-\log_{10}p = 12.82, \beta = 3.19$, present in 159/8,394 genomes; EMB:
log_{10} p = 10.86, β = 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 (β = 3.19) were larger than for ethambutol (β = 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., EVYTEGLLADR, log_{10} p = 8.46, β = −3.63, present in 8,919/8,929 genomes). Substitutions at these positions may
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;
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\textsuperscript{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 $\beta$ 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.
**Ethionamide.** Oligopeptides and oligonucleotides upstream and within the transcriptional regulator \(\text{whiB7} (\text{Rv3197A})\) 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 \(\text{whiB7}\), within the 5' untranslated region \([43]\) (e.g., AACCGTGCCGCGACTGAGTCCCT, \(-\log_{10}p = 18.18, \beta = 2.16,\) 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 \(\text{whiB7}\) promoter sequence, while deletions upstream of \(\text{whiB7}\) 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, \(\text{tap}\) and \(\text{erm}\) \([45]\). Variants in and upstream of another \(\text{whiB}\)-like transcriptional regulator, \(\text{whiB6}\), were previously found to be associated with resistance to ethionamide \([19,47]\), capreomycin, amikacin, kanamycin, and ethambutol \([22,23]\). \(\text{WhiB7}\) 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 \(\text{whiB7}\) were not found to be significantly associated with any of the other 12 antimicrobials. This could indicate yet another mechanism by which \(\text{whiB7}\) is involved in resistance to anti-TB drugs.

**Levofloxacin.** Oligopeptides in \(\text{tlyA} (\text{Rv1694})\) were significantly associated with MIC of the fluoroquinolone levofloxacin (minimum \(p\)-value \(p < 10^{-7.8}\), S11 Fig). \(\text{tlyA}\) encodes a methyltransferase that methylates ribosomal RNA. Variants in \(\text{tlyA}\), 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., GKGQVPGGKV, \(-\log_{10}p = 7.83, \beta = -1.86,\) present in 7,281/7,300 genomes). The resulting frameshift likely mimics the knockout effect of deleting the 27 C-terminal residues of \(\text{TlyA}\), 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 \(\text{tlyA}\) may therefore reveal genuine, previously unidentified cross-resistance.

**Rifabutin.** Oligonucleotides in \(\text{cysA2} (\text{Rv0815c})\) and \(\text{cysA3} (\text{Rv3117})\) 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 \(\text{cysA2}\) and \(\text{cysA3}\) 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., CATATGCCGTGACCA-TATTGCCGGCGCGAT, \(-\log_{10}p = 7.74, \beta = -2.65,\) 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 \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the first captures 2 substitutions that are synonymous in \( TTT \ TTT \) \( TAG \), \( CAGCT \ TGGCCGACTGGG \) \( G \) opposite strands. Of the 2 top oligonucleotides associated with higher MIC, the f...
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., GAATCGGA$^A$TGCTT$^G$GCCGCTGGTGCGAGTT, $-$$\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. Exhaustively 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 disequilibrium 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 \textit{Rv1218c} promoter binding site of the transcriptional repressor \textit{Rv1219c} (associated with MIC for isoniazid) upstream of the \textit{vapBC20} operon that cleaves 23S rRNA (linezolid) and in the region encoding a helix lining the active site of \textit{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 \cite{11,17}. We chose \cite{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 \cite{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., \cite{85,86}). To take another example, \textit{gyrB} did not appear in the catalogues we used for moxifloxacin \cite{11,17}. Yet our rediscovery of \textit{gyrB} 501D complements published reports associating the substitution with moxifloxacin resistance \cite{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 \cite{77}. Moreover, of the 5 new genes added to the forthcoming WHO catalogue \cite{77} but not featuring in the catalogues \cite{11,17} used here—\textit{eis} (amikacin), \textit{ethA} (ethionamide), \textit{inhA} (ethionamide), \textit{rplC} (linezolid), \textit{gyrB} (moxifloxacin)—we identify all as containing significant variants by GWAS except one, \textit{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 \textit{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 delamanid 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 \cite{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 $1 \times 10^9$, then the reads were randomly subsampled prior to assembly down to a sequence length of $1 \times 10^9$, 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 $\alpha/n_p$, 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. Petö, 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. Petö, 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. Petö, Guy Thwaites, A. Sarah Walker, Timothy M Walker, Daniel J. Wilson

**Supporting information**

S1 Data. The interpretation of oligopeptides and oligonucleotides required manual cura-
tion 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 analy-
\(s, \) 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 consid-
ered sensitive, and those to the right are considered resistant. AMI, amikacin; BDQ, bedaqui-
line; CFZ, clofazimine; DLM, delamanid; ECOFF, epidemiological cutoff; EMB, ethambutol;
ETH, ethionamide; GWAS, genome-wide association studies; INH, isoniazid; KAN, kanamy-
cin; LEV, levofloxacin; LZD, linezolid; MIC, minimum inhibitory concentration; MXF, moxi-
floxacinar; RFB, rifabutin; RIF, rifampicin.

(PDF)
S2 Fig. Oligopeptide and oligonucleotide sample heritability estimates for binary resistant vs. sensitive phenotypes compared to semiquantitative 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; MAF, minor allele frequency; 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 (oligonucleotides). 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, amikacin; BDQ, bedaquiline; CFZ, clofazimine; DLM, delamanid; EMB, ethambutol; ETH, ethionamide; INH, isoniazid; KAN, kanamycin; LEV, levofloxacin; LZD, linezolid; MAF, minor allele frequency; 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.

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.

S9 Fig. Variants in PPE42 associated with AMI and KAN MIC. Manhattan plots showing the oligopeptide association results for the PPE42 coding region for the 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.

(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.

(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.

(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.

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.

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.

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.

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 ammA2. 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.

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

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 Íñaki 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


25. Pantel A, Petrella S, Veziris N, Brossier F, Bastian S, Jarlier V, et al. Extending the definition of the classical PPE major Polymorphic Tandem Repeat open Reading Frame, Rv2608, shows a high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608. show a


33. Pantel A, Petrella S, Veziris N, Brossier F, Bastian S, Jarlier V, et al. Extending the definition of the classical PPE major Polymorphic Tandem Repeat open Reading Frame, Rv2608, shows a high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608. show a


