Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*

Francesc Coll[1,*], Jody Phelan[1*], Grant A. Hill-Cawthorne[2,3**], Mridul B Nair[2,**], Kim Mallard[1], Shahjahan Ali[2], Abdallah M Abdallah[2], Saad Alghamdi[4], Mona Alsomali[2], Abdallah O Ahmed[5], Stephanie Portelli [1,+], Yaa Oppong[1], Adriana Alves[6], Theolis Barbosa Bessa[7], Susana Campino[1], Maxine Caws[8,9], Anirvan Chatterjee[10], Amelia C Crampin[11,12], Keertan Dheda[13], Nicholas Furnham [1], Judith R Glynn[11,12], Louis Grandjean[14], Dang Thi Minh Ha[9], Rumina Hasan[15], Zahra Hasan[15], Martin L Hibberd[1], Moses Joloba[16], Edward C. Jones-López[17], Tomoshige Matsumoto[18], Anabela Miranda[6], David J Moore[1,14], Nora Mocillo[19], Stefan Panaiotov[20], Julian Parkhill[21], Carlos Penha[22], João Perdigão[23], Isabel Portugal[23], Zineb Rchiad[2], Jaime Robledo[24], Patricia Sheen[13], Nashwa Talaat Shesha[25], Frik A Sirgel[26], Christophe Sola[27], Erivelton de Oliveira Sousa[28], Elizabeth M Streicher[26], Paul Van Helden [26], Miguel Viveiros[29], Robert M Warren[26], Ruth McNerney [1,13,***], Arnab Pain[2,30,***], Taane G Clark[1,11,***]

1 Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom
2 Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia
3 Sydney Emerging Infections and Biosecurity Institute and School of Public Health, Sydney Medical School, University of Sydney, NSW 2006, Australia
4 Laboratory Medicine Department, Faculty of Applied Medical Sciences, Umm Al-Qura University, Kingdom of Saudi Arabia
5 Department of Microbiology, Faculty of Medicine, Umm Al-Qura University, Makkah, Saudi Arabia
6 National Mycobacterium Reference Laboratory, Porto, Portugal
7 Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz Bahia R. Waldemar Falcao 121 Candeal 40296-710 Salvador Bahia Brazil
8 Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, United Kingdom
9 Pham Ngoc Thach Hospital for TB and Lung Diseases, Hung Vuong, Ho Chi Minh City, Vietnam
10 The Foundation for Medical Research, 84-A, R. G. Thadani Marg, Worli, Mumbai 400018, India
11 Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

12 Karonga Prevention Study, Chilumba, Karonga, Malawi
13 Lung Infection and Immunity Unit, UCT Lung Institute, University of Cape Town, Groote Schuur Hospital, Observatory, 7925, Cape Town, South Africa.
14 Laboratorio de Enfermedades Infecciosas, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru
15 Department of Pathology and Laboratory Medicine, The Aga Khan University, Stadium Road, P.O. Box 3500, Karachi 74800, Pakistan
16 Department of Medical Microbiology, Makerere University College of Health Sciences, Kampala, Uganda
17 Section of Infectious Diseases, Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, Massachusetts, USA
18 Osaka Anti-Tuberculosis Association Osaka Hospital, Osaka, Japan
19 Reference Laboratory of Tuberculosis Control, Buenos Aires, Argentina
20 National Center of Infectious and Parasitic Diseases, 1504 Sofia, Bulgaria
21 Wellcome Trust Sanger Institute, Hinxton, United Kingdom
22 Instituto Gulbenkian de Ciência, Lisbon, Portugal
23 iMed.ULisboa - Research Institute for Medicines, Faculdade de Farmácia, Universidade de Lisboa, Portugal
24 Corporación para Investigaciones Biológicas, Universidad Pontificia Bolivariana, Medellín, Colombia
25 Regional Laboratory Directorate of Health Affairs, Makkah, Kingdom of Saudi Arabia.
26 Division of Molecular Biology and Human Genetics, SAMRC Centre for Tuberculosis Research, DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa
27 Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France
28 Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz Bahia R. Waldemar Falcao 121 Candeal 40296-710 Salvador Bahia Brazl and Laboratorio Central de Saude Publica Prof. Goncalo Moniz Rua Waldemar Falcao, 123 Horto Florestal 40295-010 Salvador Bahia Brazil
29 Unidade de Microbiologia Médica, Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, UNL, Lisboa, Portugal
30 Global Station for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, N20 W10 Kita-ku, Sapporo, 001-0020 Japan

[+] Present address: Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, VIC 3052, Australia

[*] Joint first authors, contributed equally.

[**] Contributed equally.

[***] Corresponding authors: Taane Clark (e-mail: taane.clark@lshtm.ac.uk) or Arnab Pain (e-mail: arnab.pain@kaust.edu.sa), or Ruth McNerney (ruth.mcnerney@uct.ac.za)

ABSTRACT

To further characterize the genetic determinants of resistance to anti-tuberculosis drugs we performed a genome-wide association study (GWAS) of 6,465 *Mycobacterium tuberculosis* clinical isolates from more than 30 countries. A GWAS approach within a mixed regression framework was followed by a phylogenetic-based test for independent mutations. In addition to mutations in established and recently described resistance-associated genes, novel mutations were discovered for resistance to cycloserine, ethionamide and para-aminosalicylic acid. The capacity to detect mutations associated with resistance to ethionamide, pyrazinamide, capreomycin, cycloserine and para-aminosalicylic acid was enhanced by inclusion of insertions and deletions. Odds ratios for mutations within candidate genes were found to reflect levels of resistance. Novel epistatic relationships between candidate drug resistance genes were identified. Findings also suggest the involvement of efflux pumps (*drrA a*nd *Rv2688c*) in the emergence of resistance. This study will inform the design of new diagnostic tests and expedite the investigation of resistance and compensatory epistatic mechanisms.

KEY WORDS: *Mycobacterium tuberculosis***,** tuberculosis, GWAS, drug resistance, MDR-TB, XDR-TB

Word count: 4,200

Introduction

The emergence and spread of *Mycobacterium tuberculosis* (Mtb) resistant to multiple anti-tuberculous drugs is of global concern. Programmatically incurable tuberculosis (TB), where effective treatment regimens cannot be provided due to resistance to the available drugs is a growing problem[1]**.** Resistance to rifampicin and isoniazid is classed as multidrug-resistant tuberculosis (MDR-TB), further resistance to the fluoroquinolones and any of the injectable drugs (amikacin, kanamycin or capreomycin) used to treat MDR-TB is termed extensively drug-resistant (XDR-TB). Treatment for patients with drug resistant tuberculosis is prolonged, expensive and outcomes are poor[2]. The drugs used are toxic and poorly tolerated, adverse events are common and may be severe and irreversible[3]. Inadequate treatment also risks amplification of resistance to further drugs and may prolong opportunities for transmission[4]**.**

Mtb has a clonal genome (size 4.4Mb) with a low mutation rate and no evidence of between-strain recombination or horizontal gene transfer[5]. The Mtb complex comprises seven lineages, of which four are predominant in humans: Lineage 1, Indo-Oceanic (e.g. East-African-Indian (EAI) spoligotype families); Lineage 2, East-Asian (e.g. W/Beijing spoligotype families); Lineage 3, East-African-Indian (e.g. Central-Asian-Strain (e.g. CAS-DELHI) spoligotype families) and Lineage 4, Euro-American (e.g. Latin American-Mediterranean (LAM), Haarlem and the "ill-defined" T spoligotype families)[5].

Resistance in Mtb is mainly conferred by nucleotide variations (single nucleotide polymorphisms, insertions and deletions (indels)) in genes coding for drug-targets or -converting enzymes. Changes in efflux pump regulation may have an impact on the

emergence of resistance[6] and putative compensatory mechanisms to overcome fitness impairment coincidental with the acquisition of resistance have been described for some drugs[7]. Detection of resistance conferring mutations offers a means of rapidly identifying resistance to anti-tuberculosis drugs[8] but, with the exception of rifampicin, current molecular tests for resistance lack high levels of sensitivity[8]. To improve knowledge of genetic determinants of drug resistance we undertook whole genome analysis of a large collection (n=6,465) of clinical isolates from more than 30 geographic locations, representing the four major Mtb lineages **(Figure 1, Supplementary table 1)**. We adopted a GWAS approach to identify nucleotide variation and loci underlying drug resistance as successfully applied in Mtb[9–11] and other bacteria[12,13]. A total of 14 drugs with available phenotypic data on drug susceptibility testing were investigated **(Supplementary table 2)**. Phenotypic drug susceptibility data was not available for each of the 14 drugs for every isolate and sample sizes ranged from over 6,000 for the most commonly tested first line drugs (isoniazid and rifampicin) to 255 and 248 for p-aminosalicylic acid and cycloserine, respectively, which are used to treat patients with XDR-TB. Here, we present findings from the most comprehensive study yet undertaken of the genetic determinants of resistance to anti-tuberculosis drugs or the Mtb resistome.

RESULTS

Genetic diversity and drug resistance

High quality genome-wide SNPs (102,160), indels (11,122), and large deletions (284) were identified across all samples (n=6,465). Most SNPs (93.1%) had rare minor alleles (allele frequency <1%) (**Supplementary Figure 1**). Similarly, small indels were rare (96.6% had frequency <1%), and ranged in size from 1 to 45bp. A phylogenetic tree and principal

component analysis constructed using all genome-wide SNPs revealed the expected clustering by lineage (**Figure 2, Supplementary Figure 2**).

Phenotypic analysis of susceptibility to anti-tuberculosis drugs found 31.2% of isolates were resistant to at least one drug, 15.1% were categorized as MDR-TB and 4.3% as XDR-TB (**Supplementary table 2, Figure 2**). Fourteen drugs were included in the genome-wide analysis: isoniazid (INH), rifampicin (RIF), ethionamide (ETH), pyrazinamide (PZA), ethambutol (EMB), streptomycin (STM), amikacin (AMK), capreomycin (CAP), kanamycin (KAN), ciprofloxacin (CIP), ofloxacin (OFL), moxifloxacin (MOX), cycloserine (CYS) and para-aminosalicylic acid (PAS). Drug family groups including the second-line injectable drugs (SLID: AMK, KAN, CAP) and fluoroquinolones (FLQ: CIP, OFL, MOX) were also analysed. Insufficient phenotypic data was available for the inclusion of the new and repurposed drugs, bedaquiline, delamanid and linezolid. To reveal loci associated with drug resistance complementary methods were applied to mutations and aggregated non-synonymous mutations: a tree-based "PhyC" test for convergent evolution to detect homoplastic variants[9] and a GWAS approach within a mixed regression framework (See **Online methods**). Unless stated otherwise, all analysis used the complete dataset. First, we consider MDR-TB and XDR-TB phenotypes **(Table 1)** and then individual drug GWAS and evolutionary results (**Table 2**).

GWAS and phyC tests for MDR-TB and XDR-TB

The gene-based GWAS of MDR-TB versus susceptible identified *rpoB* (RIF), *Rv1482c-fabG1* operon (INH, ETH), *inhA* (INH, ETH), *katG* (INH), and *oxyR'-ahpC* (compensatory mechanism for INH). The *katG* mutations at codon 315 (S315T, S315N, S315R) were all statistically

significant, and collectively were the most frequent mutations (75.2%) across all resistance loci identified, consistent with a recent study[14] and highlighting their pivotal role in the emergence of INH resistance and MDR-TB. The *katG* S315T mutation is thought to emerge before RIF resistance associated mutations and therefore, from an evolutionary standpoint, preclude the emergence of MDR-TB[14,15]. However, our analysis highlighted that *Rv1482c-fabG1* and *inhA* mutations, in the absence of *katG* S315T, can emerge prior to MDR-TB, as previously shown in two phylogenetically-independent clades in Lisbon[16,17]. The other frequent MDR-TB mutations in our study included *rpoB*-S450L (RIF, 64.2%), *embB*-M306L/V/I (EMB, 49.1%), and *rpsL*-K43R (STM, 42.2%) (**Supplementary table 3**), and the prevalence correlates with historical treatment practice and emergence of resistance. There are corresponding signals of INH/RIF co-resistance with other first-line drugs, with the detection of gene-based association signals for *gid* (STM) and *rpsL* (STM), and a SNP-based association signal for the *embC-embA* intergenic region (EMB). SNP-based PhyC analysis detected the above loci, but in addition *folC* (PAS), *pncA-Rv2044c* intergenic region (PZA), and *whiB6-Rv3863* intergenic (putative STM or ETH) regions.

The gene-based GWAS of XDR-TB versus MDR-TB identified mutations in *gyrA* (FLQ), *rrs* (aminoglycosides), the *embC-embA* intergenic region and *ubiA* (EMB). The PhyC test additionally revealed *eis-Rv2417c* (KAN), *gyrB (FLQ), rrs (*aminoglycosides), *folC* (PAS), *alr* (CYS), *gid* (STM) SNPs, and a novel mutation in the *thyX-hsdS.1* intergenic region (A-9T, PAS)[18,19]. In addition to loci identified above, the gene-based GWAS comparing XDR-TB to susceptible groups identified *rpoC* (a compensatory mechanism for RIF resistance), *ethA* (ETH), *eis-Rv2417c* (KAN) and *PPE52-nuoA* (novel intergenic region, G-314T). The PhyC test additionally detected SNPs in *gyrB* (FLQ, D461N, D641H, T500N, T500I and A504V),

supported the *thyX-hsdS.1* intergenic region SNP finding (PAS, A-9T), as well as identified a previously unreported *ubiA* SNP association (EMB, M180V).

The *drrA* Arg262Gly mutation was significantly associated with XDR-TB compared to susceptible (mutation frequency 18% vs. 0%, respectively, $P=1.5x10^{-8}$). We hypothesize that *drrA* may be involved in export of drugs across the membrane based on its strong association with XDR-TB in our study and its functional annotation as a probable transporter of antibiotics across the membrane (TubercuList, see URLs). This hypothesis is in accordance with the finding that *rpoB* mutations in Mtb may trigger compensatory transcriptional changes in genes involved in secondary metabolism, in particular, in the biosynthesis and export of phthiocerol dimycocerosate (PDIM), increasing expression and activity. As a consequence these strains became more virulent and multidrug resistant, increasing their fitness by increased efflux activity and lipid metabolism[20,21]. Similarly, a mutation in the *Rv1144-mmpL13a* intergenic region (C-102A) was highly associated with XDR-TB versus susceptible (mutation frequency 17% vs. 0%, respectively, $P=1.5x10^{-7}$). This mutation sits in the promoter to the operon containing *mmpL13a* and *mmpL13b,* which code for transmembrane transport proteins and could influence expression of these proteins[6].

Lineage-specific and compensatory mechanisms

We conducted a stratified GWAS per lineage to identify lineage-specific loci associated with drug resistance. Most associations were present in more than one lineage. The largest number of lineage-specific drug resistance mutations were found in lineage 4, which was the largest collection investigated and contained more genetically diverse clones[5], implying that geographically restricted mutations are being captured (**Supplementary table 4**). A

previously unreported putative compensatory locus was identified for pyrazinamide (*pncB1*) through analysis of lineage 1 which reached borderline significance for lineage 3.

We applied a systematic approach to reveal epistatic interactions between GWAS loci (from **Table 2**) or explore known compensatory effects using a test of non-random association to detect the frequent co-occurrence of mutations in pairs of loci (Fisher exact test, P-value cut-off $<1x10^{-8}$) (**Supplementary table 5**). Deep phylogenetic mutations were removed to increase robustness. This approach proved to be successful at identifying well-known compensatory relationships between *rpoB* and *rpoC* loci (RIF)[7], *rpoB* and *rpoA* (RIF)[22] and *katG* and *oxyR'-ahpC* (INH)[23]. We captured the frequent co-occurrence of *embB* and *ubiA* mutations which together are known to lead to high levels of EMB resistance[24], and they are therefore unlikely to represent a compensatory mechanism. Novel epistatic relationships included *pncA* with *pncB2* (PZA) and *thyA* with *thyX-hsdS.1* (PAS). The *pncB2* epistatic effect with *pncA* appears to be specific to lineage 4 **(Supplementary table 6)**. The other nicotinamide co-factor, *pncB1,* had weaker evidence of an epistatic relationship with *pncA* in lineage 1 (P=0.0016) **(Supplementary table 6).** Similarly, there was marginal evidence for *pyrG* (lineage 4, P=0.00016)[25] and *Rv0565c* (lineage 2, P=0.00027) with *ethA* (ETH)[26] **(Supplementary table 6)**. Follow-up investigations will need to determine whether mutations in these loci have an impact on the minimal inhibitory concentration (MIC) values or function as compensatory mechanisms.

Overall, the GWAS approach was effective at detecting known drug resistance determinants and epistatic (gene-gene) relationships and identified novel ones that warrant functional validation in future studies. As resistance loci for individual drugs, especially second-line

treatments, may be masked by an analysis of the composite MDR-TB and XDR-TB outcomes, we repeated the GWAS, PhyC test and epistatic analysis for the 14 individual drugs considered.

GWAS and phyC tests for individual drugs

Rifampicin, isoniazid and ethionamide

The *rpoB* locus showed the strongest association with RIF resistance, but the compensatory effects of *rpoC* and *rpoA* were also evident through homoplasy SNP analysis. As previously reported non-synonymous SNPs in *rpoC* (272 identified) were spread across the whole gene[27]. Altered or diminished activity of the catalase-peroxidase enzyme *KatG* is the most frequent mechanism of isoniazid resistance[28]**,** and as expected, the *katG* gene ranked first in the GWAS for this drug. Mutations in proposed INH drug targets, *kasA* and *kasB* previously included in some drug resistance databases, did not reach statistical significance in our study[29], suggesting an odds ratio below our detection level of 1.4 (with 99% confidence of detection, 90% statistical power)**.** Both *inhA,* encoding the molecular target of isoniazid[30] and the *Rv1482c-fabG1* intergenic region harbouring its promoter, showed strong associations with INH and ETH, with greater effects in the former. In addition, mutations associated with the *oxyR'–ahpC* intergenic region (20 detected) were found in the presence of *katG* polymorphisms (28), supporting its role as a compensatory mechanism in isoniazid resistant strains. For ethionamide, the *ethA* locus, encoding the drug-metabolising enzyme was found to be associated with resistance as described previously[31]. A total of 153 non-synonymous mutations were identified in *ethA,* scattered throughout the gene and mostly affecting codons different from those already described[8].

Ethambutol

Mutations in the *embCAB* operon, which encodes for enzymes involved in the biosynthesis of arabinan components of the mycobacterial cell wall, are mostly responsible for EMB resistance but are not fully penetrant for resistance[32]**.** The *embB* and the *embC-embA* intergenic region had the strongest associations. *Rv3806c* (*ubiA*), described to contribute to high levels of EMB resistance *in vitro*[17] was also significantly associated in our analysis demonstrating a role in clinical samples too across all four lineages. Two novel loci were identified: *Rv2820c* thought to enhance mycobacterial virulence *ex vivo* and *in vivo,* and *Rv3300c* a conserved protein with unknown function (TubercuList, see URLs).

Pyrazinamide

The *pncA* locus was the highest ranked association with PZA resistance in the GWAS and was a target of independent mutation, consistent with its established role[33]**.** Additionally, many low frequency SNPs were reported which were not used in the association analysis and could potentially confer resistance (**Supplementary data 1**). Other proposed PZA targets, namely *rpsA*[34] and *panD*[35], did not reach statistical significance in the GWAS and were not targets of independent mutation among PZA resistant strains in our collection.

Streptomycin

The rpsL, *rrs* and *gid* loci, all known to be involved in STM resistance[18] were identified by GWAS. Mutations in *rpsL* are known to lead to intermediate to high levels of STM resistance[36], and accordingly we observed high odds ratios indicative of high penetrance in association signals in this locus (**Figure 3A**). In contrast, candidate *rrs* and *gid* gene polymorphisms showed weaker overall signals (lower odds ratio) in the GWAS, which concurs with existing evidence that *gid* and *rrs* mutations confer lower levels of resistance[36] (differences in odds ratios: *rpsl* vs. *rrs/gid* Wilcoxon P = 0.03; *rpsl* vs. *gid* Wilcoxon P = 0.04).

Fluoroquinolones and Second-line injectables

The gene- and SNP-based GWAS analysis revealed the *gyrA* locus, which encodes for the molecular target of FLQ[37], as the strongest association signal. In addition to homoplastic mutations in *gyrA*, evidence of independent mutation was detected in *gyrB*[38]. The *Rv2688c* C213R mutation was associated with MOX and FLQ resistance but did not reach statistical significance in OFL. The antibiotic transport ATP-binding protein encoded by *Rv2688c* is a known FLQ efflux gene[39]**.** As expected the strongest resistance gene and SNP-based association signals across AMK, KAN, and CAP was with the aminoglycoside (SLID) target gene *rrs*[18]**.** Association was observed with mutations in the *eis* promoter known to result in low levels of KAN resistance but not in co-resistance with other aminoglycosides[40]**.** Although the *eis* promoter mutations had a lower median odds ratio than that of *rrs* mutations, potentially supporting evidence that *rrs* mutations confer higher levels of KAN resistance[40], this was not statistically significant due to small sample size (differences in odds ratios Wilcoxon P=0.24) (see **Figure 3A**).

D-Cycloserine

CYS inhibits the Alr enzyme, responsible for the conversion of L-Alanine into D-Alanine, by competing with L-Alanine for the active site. Resistance to CYS results from mutations in the *alr* coding region[41]**.** In our study *alr* was significantly associated with CYS resistance (**Table 2**) in line with recent evidence showing that clinical strains with *alr* mutations exhibit increased resistance to CYS[11] and harboured multiple homoplastic mutations including Phe4Leu, Lys113Arg and Met343Thr. In a previous study, the Met343Thr mutation was detected in an XDR-TB strain that had been exposed to CYS treatment, predicted to alter the protein structure of Alr, and therefore it was hypothesised to be involved in CYS resistance[42]. To further understand the functional impact of the mutations found in *alr* we modelled the effect of these variants using the available crystal protein structure (PDB 1XFC,

Supplementary figure 3). Mutations in *alr* were found to differ in their proximity to the CYS binding site and their effect on protein stability and ligand binding (**Supplementary table 7**). The Met343Thr mutation (found in 12 susceptible and 2 resistant isolates) was predicted to have more drastic effect on protein structure compared to Lys113Arg, the most frequent mutation among CYS resistant isolates (in 7 susceptible and 23 resistant isolates). There appears to be a balance between the fitness cost associated with mutations and their frequency (**Supplementary table 7)**. The Met343Thr mutation appears independently throughout the phylogenetic tree, but did not reach statistical significance for association to drug resistance (XDR-TB or CYS), implying that selection may be acting on this mutation but drug resistance may not be the driving factor.

Para-aminosalicylic acid

PAS is a pro-drug that is converted into its active form by *thyA* - a thymidylate synthase, which is an essential gene for Mtb survival. The candidate drug resistance loci are those involved in folate metabolism and biosynthesis of thymidine nucleotides (*thyA*, *dfrA*, *folC*, *folP1*, *folP2* and *thyX*)[19]**.** Of these, *thyA* and *thyX-hsdS.1* (directly upstream of *thyX*) and were found to be associated with PAS drug resistance in both gene- and SNP-based GWAS. Importantly, it has been shown that G-16A SNP found in our study increased *thyX* expression by 18-fold relative to wild-type promoter although no link with PAS resistance was made[18]**.** Of 3 PAS resistance strains with the G-16A *thyX* promoter mutation, 2 also had a *thyA* mutation (P145L, H207R), further supporting that up-regulation of *thyX* is involved in resistance to PAS[26], or has a compensatory role. The G-16A *thyX* is a homoplastic mutation, and therefore more likely to be compensatory.

Overall, the log-transformed odds ratios for the association of mutations with known levels of resistance followed an increasing trend from low to intermediate to high (**Figure 3B**; log odds ratios: linear regression trend $P = 1.5 \times 10^{-9}$, high versus intermediate $P = 5.2 \times 10^{-5}$; intermediate versus low $P = 5.8 \times 10^{-10}$). This analysis demonstrates a potential utility of using odds ratios and their statistical significance to indicate the impact of a mutation and its propensity to cause low, intermediate or high-level resistance. Further, the odds ratios for the novel findings were marginally lower than those for known ones (Wilcoxon test $P = 8.3 \times 10^{-5}$), reflecting the ability of the GWAS to discover effect sizes of lower magnitude **(Figure 3C)**. A pathway analysis comparing MDR-TB/XDR-TB to susceptible strains revealed only one significant annotation cluster with 17.7-fold enrichment for antibiotic resistance and response to antibiotics ($P=1.6 \times 10^{-7}$), further confirming the robustness of the GWAS approach.

Association tests using small indels and large deletions

An analysis of genome-wide small indels revealed associations in candidate resistance genes and operons (**Supplementary table 8, Supplementary data 1**). The candidate genes differed in their abundance of small indels, reflecting their essentiality for survival: drug targets had less density of indels whereas drug-metabolising enzymes had a greater density. For example, the *pncA* gene was the most polymorphic coding region (PZA, 44.72 indels /kb) while the least polymorphic was *rpoB* (RIF, 2.3 indels /kb). Although, most small indels (83%) in the candidate regions were 1bp in length and caused frame-shifts, the indels in *rpoB* inserted or deleted whole codons, i.e. they did not cause a shift in the codon reading frame. Indels in *rpoB, pncA* and the *embAB* promoter region were associated with MDR-TB, XDR-TB

and their respective targets/activators. Indels in *ethA* were associated with ETH and XDR-TB resistance. Similarly, *gid* indels were associated with STM as expected.

The analysis of CYS revealed indel associations with the *ald* gene, supporting recent reports that loss of function in *ald* confers resistance[11]. Thus resistance to CYS appears to be conferred by both SNPs in *alr* and indels in *ald*. Indels found in *rrs* were associated with KAN and CAP resistance, however they did not reach statistical significance for STM, which has a different drug binding site. CAP resistance was also found to be associated with three indels in *tlyA,* two of which are located at the 3' end of the gene. In general, indels were distributed throughout the gene lengths however there was some evidence of areas of higher density such as the *pncA* region between codons 130 and 132 (close to the catalytic centre) and the *rpoB* 427-434 codon region.

The only large deletion association identified by GWAS was a region encompassing the *thyA* and *dfrA* genes and PAS resistance. Five samples across 4 countries contained large *thyA-dfrA* deletions of varying length **(Supplementary table 9, Supplementary figure 4).** Associations in partial or whole gene deletions in *katG, ethA* and *pncA*, were close to statistical significance ($P<0.05$). These genes activate pro-drugs, and none are considered to be essential to Mtb survival. The large deletions detected occur independently in different branches of the phylogenetic tree and are likely to offer an alternative route to resistance compared to small genomic variants, across lineages and populations.

Effects on resistance prediction using GWAS variants

We sought to establish if any of the mutations found in association and homoplastic analysis increased the predictability of individual drug resistance phenotypes (**Table 3**). We used the reported phenotypic drug susceptibility test result as the reference standard to calculate the sensitivity and specificity for mutation-resistance predictions. Using a previously established library of mutations[8,17] (TBDR library), we found that although the sensitivity was greater than 80% in 8/14 drugs, a substantial proportion of resistance phenotypes were not explained by known mutations, particularly in second-line drugs. Using the novel SNPs identified in this study we gained sensitivity for PAS (+10%), ETH (+14%) and CYS (+50%, not included in the TBDR library) (**Table 3**). The additional inclusion of small indels and large deletions further improved the predictive ability for 9 drugs while maintaining specificities of at least 90%, except for ETH which is 72% (**Table 3**).

DISCUSSION

To provide genomic insights into Mtb drug resistance we have combined the power of whole genome sequencing with a genome-wide association analytical approach in the largest and most geographically widespread study to date, encompassing a total of 6,465 clinical isolates of Mtb from more than 30 countries. Large sample sizes are required to identify complex or infrequent genetic effects, but also to negate effects due to possible errors in phenotypic drug susceptibility testing and misclassification[43]**.** The lack of standardization of phenotypic testing methodologies for Mtb is also a potential source of bias which was reduced by the inclusion of samples from different countries and laboratories using a variety of quality assured testing methodologies. Whilst resistant phenotypes may be imputed from established resistance causing mutations, inferring susceptibility to a drug cannot be assumed in the absence of corroborating evidence[17]. The

completeness of our susceptibility test data meant that both GWAS and homoplasy-based methods could be applied across 14 drugs.

The GWAS identified well-established resistance loci and compensatory relationships, thereby confirming the authenticity and robustness of the approach. It also revealed several recently discovered loci (*folC*, *ubiA*, *thyX-hsdS.1*, *thyA*, *alr*, *ald*, *dfrA-thyA*), new epistatic relationships (*pncA* with *pncB2*, and *thyA* with *thyX-hsdS.1*) and efflux pumps represented by the ABC transporters *drrA* and *Rv2688c* associated with drug resistance. The novel genetic markers associated with resistance identified in this GWAS included SNPs in the *ethA* and *thyX* promoters, small indels in *pncA* and *ald,* and large deletions in pro-drug activators such as *ethA* and *katG*. These loci warrant functional follow-up and characterization studies to fully elucidate their role in treatment failure. The associations identified may shed light on the molecular mechanisms underlying drug resistance and assist in the design of novel antibiotics.

In our study, sample sizes for second-line drugs were reduced compared to the first-line drugs. This was due to the lower prevalence of resistance to second-line drugs and the fact that isolates susceptible to first-line drugs are not routinely tested for second-line drugs. However, due to the large effect that causal mutations have on drug resistance phenotypes, although not ideal, relatively small samples of bacterial genomes can be sufficient to identify causal mutations [43] as has been demonstrated in previous studies on Mtb [10-12]. It should be noted that bedaquiline, delamanid and linezolid were excluded from our analysis due to the paucity of phenotypic susceptibility data.

The analysis also highlighted the importance of indels on drug resistance, particularly their high density in drug-metabolizing genes, in contrast to highly essential drug-target genes where their density was low. The inclusion of small indels and large deletions improved the predictability of resistance phenotypes. However, for drugs like CYS and PAS mechanisms of drug resistance remain unknown and larger numbers of resistant cases will be required to elucidate them. It is also possible that unknown mechanisms may be explained by the role of epigenetics and gene expression[44].

Mtb strains are usually classified as drug resistant or susceptible based on their capacity to grow *in vitro* when exposed to a critical concentration of the drug. Phenotypic testing methods have a degree of uncertainty, especially close to the threshold[43]**.** Testing against a range of drug concentrations to establish the minimum inhibitory concentration (MIC) is a preferred approach to determine the level of resistance but is not routinely undertaken[40]. MIC values were not available for every isolate presented here, but despite this limitation, loci known to be involved in low-levels of resistance (**Table 3**), were identified by our analysis. Indeed, our analysis revealed a relationship between known levels of resistance and the odds ratios from the GWAS, which could aid the clinical interpretation of molecular diagnostic data including measuring the sensitivity and specificity of individual mutations when diagnosing drug resistance.

Emergence of resistance is driven by drug exposure and local TB treatment practices are a major influence on the prevalence and pattern of resistance. A limitation of this study was the sampling methodology since collection of the isolates was not controlled or systematic and resistant isolates were not evenly distributed across collection sites. However, within

our study population we covered the four major Mtb lineages across 5 continents and sampled multiple geographical regions, allowing us to observe differences in the prevalence of drug resistance mutations and mechanisms. Some of drug resistance and compensatory/epistatic relationships were found to vary across geographical populations and bacterial lineage, implying that regional variation should be considered to fully characterise genotype-phenotype relationships. The differential lineage effects could impact on relative virulence between strain-types. Enhanced understanding of the genetic basis of anti-tuberculous phenotypic drug resistance will also aid in the development of more accurate molecular diagnostics for drug-resistant TB. An important finding of this study is the significance of genomic variation other than SNPs which has implications for the design of molecular tests for resistance. Improved tools are needed to guide treatment of patients with multidrug-resistant disease where personalized treatment offers improved rates of cure[45]**.** Next generation sequencing offers a comprehensive assessment and may be used to guide treatment[45]. Although such technology is currently being implemented in some low burden countries such as the United Kingdom, it remains to be trialled in resource-poor settings that are representative of most TB patients worldwide.

ACKNOWLEDGMENTS

The project was supported by the KAUST faculty baseline research fund (BAS/1/1020-01-01) to A.P. The authors wish to thank members of KAUST Bioscience Core laboratory who sequenced samples. We thank the Wellcome Trust Sanger Institute core and pathogen sequencing and informatics teams who were involved in the Malawi and Uganda studies. The work was funded in part by the Wellcome Trust (Grant numbers WT096249/Z/11/B, WT088559MA, WT081814/Z/06/Z, and WT098051), and the Wellcome Trust-Burroughs

Wellcome Fund Infectious Diseases Initiative grant (number 063410/ABC/00/Z). F.C. was the recipient of a Bloomsbury College PhD Studentship and was supported by the Wellcome Trust (201344/Z/16/Z); J. Perdigao received a *Fundação* para a *Ciência* e a *Tecnologia* (Portugal) Post-doctoral fellowship fund (SFRH/BPD/95406/2013). The Calouste Gulbenkian Foundation, the Institute Gulbenkian in Lisbon and European Society of Clinical Microbiology and Infectious Diseases supported the research of C.P., J. Perdigao, I.P. and M.V. J. Phelan is funded by a BBSRC PhD studentship. N.F. is funded by the Medical Research Council (grant reference no. MR/K020420/1). T.G.C is funded by the Medical Research Council UK (Grant no. MR/K000551/1 and MR/M01360X/1, MR/N010469/1, MC_PC_15103). T.M. is supported by the Ministry of Health, Labor and Welfare of Japan (H21- Shinkou-Ippan-008 and H24-Shinkou-Ippan-010). We thank Nerges Mistry (Foundation for Medical Research, Mumbai) for contributing *M. tuberculosis* archived strains and drug-sensitivity testing data. We wish to thank Prof. Goncalo Moniz at the Laboratorio Central de Saude Publica for supporting the collection of samples in Brazil, and the South African National Health Laboratory Service for their contribution providing access clinical Mtb isolates. The MRC eMedLab computing resource was used for bioinformatics and statistical analysis. The authors declare no conflicts of interest. The work has been performed as part of the TB Global Drug Resistance Collaboration (see URLs).

AUTHOR CONTRIBUTIONS

R.M., A.P. and T.G.C. conceived and directed the project. G.A.H.-C., K.M. and R.M. coordinated sample collection and undertook DNA extraction. S. Alghamdi, A.M.A, A.O.A., A.A., T.B.B., M.C., A.C., A.C.C., K.D., L.G., J.R.G., D.T.M.H., R.H., Z.H., P.V.H., M.J., E.C.J.-L., T.M., A.M., N.M., D.J.M., S. Panaiotov, I.P., C.P., J. Perdigão, J.R., P.S., N.T.S., F.A.S., C.S.,

E.d.O.S., E.M.S., P.V.H., M.V. and R.M.W. undertook sample collection, DNA extraction, genotyping and phenotypic drug resistance testing. G.A.H.-C., M.B.N., M.A., Z.R. and S.Ali. prepared libraries for Illumina sequencing. J. Parkhill led the generation of Malawian and Ugandan sequencing data. F.C. and J. Phelan performed bioinformatic and statistical analyses under the supervision of T.G.C. S. Portelli and Y.O. performed additional confirmatory analysis under the supervision of M.L.H., N.F. and T.G.C. F.C., J. Phelan, S. Portelli, S.C., N.F., M.L.H., R.M., A.P. and T.G.C. interpreted results. F.C., J. Phelan, R.M. and T.G.C. wrote the first draft of the manuscript. All authors commented to and edited various versions of the draft manuscript. F.C., J. Phelan, R.M. and T.G.C. compiled the final manuscript. All authors approved the final manuscript.

COMPETING FINANCIAL INTERESTS STATEMENT

There are no conflicts of interest.

REFERENCES

1. Dheda, K. *et al.* Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *The Lancet Respiratory Medicine* **2,** 321–338 (2014).
2. Bastos, M. L. *et al.* Treatment Outcomes of Patients With Multidrug-Resistant and Extensively Drug-Resistant Tuberculosis According to Drug Susceptibility Testing to First- and Second-line Drugs: An Individual Patient Data Meta-analysis. *Clinical Infectious Diseases* **59,** 1364–1374 (2014).
3. Shean, K. *et al.* Drug-Associated Adverse Events and Their Relationship with Outcomes in Patients Receiving Treatment for Extensively Drug-Resistant Tuberculosis in South Africa. *PLoS ONE* **8,** e63057 (2013).
4. Clark, T. G. *et al.* Elucidating Emergence and Transmission of Multidrug-Resistant Tuberculosis in Treatment Experienced Patients by Whole Genome Sequencing. *PLoS ONE* **8,** e83012 (2013).
5. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature Communications* **5,** 4812 (2014).
6. Black, P. a *et al.* Energy Metabolism and Drug Efflux in Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy* **58,** 2491–2503 (2014).
7. de Vos, M. *et al.* Putative Compensatory Mutations in the rpoC Gene of Rifampin-

Resistant Mycobacterium tuberculosis Are Associated with Ongoing Transmission. *Antimicrobial Agents and Chemotherapy* **57,** 827–832 (2013).
8. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Medicine* **7,** 51 (2015).
9. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nature Genetics* **45,** 1183–1189 (2013).
10. Zhang, H. *et al.* Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature Genetics* **45,** 1255–1260 (2013).
11. Desjardins, C. A. *et al.* Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. *Nature Genetics* **48,** 544–551 (2016).
12. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology* **1,** 16041 (2016).
13. Chewapreecha, C. *et al.* Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS genetics* **10,** e1004547 (2014).
14. Manson, A. L. *et al.* Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nature Genetics* **49,** 395–402 (2017).
15. Cohen, K. a. *et al.* Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLOS Medicine* **12,** e1001880 (2015).
16. Perdigão, J. *et al.* Unraveling Mycobacterium tuberculosis genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. *BMC genomics* **15,** 991 (2014).
17. Phelan, J. *et al.* The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Medicine* **8,** 132 (2016).
18. Meier, A., Sander, P., Schaper, K. J., Scholz, M. & Böttger, E. C. Correlation of molecular resistance mechanisms and phenotypic resistance levels in streptomycin-resistant Mycobacterium tuberculosis. *Antimicrobial agents and chemotherapy* **40,** 2452–4 (1996).
19. Zhang, X. *et al.* Genetic Determinants Involved in *p* -Aminosalicylic Acid Resistance in Clinical Isolates from Tuberculosis Patients in Northern China from 2006 to 2012. *Antimicrobial Agents and Chemotherapy* **59,** 1320–1324 (2015).
20. Bisson, G. P. *et al.* Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, rpoB mutant Mycobacterium tuberculosis. *Journal of bacteriology* **194,** 6441–52 (2012).
21. Chatterjee, A., Saranath, D., Bhatter, P. & Mistry, N. Global transcriptional profiling of longitudinal clinical isolates of Mycobacterium tuberculosis exhibiting rapid accumulation of drug resistance. *PloS one* **8,** e54717 (2013).
22. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics* **44,** 106–110 (2011).

23. Sherman, D. R. *et al.* Compensatory ahpC Gene Expression in Isoniazid-Resistant Mycobacterium tuberculosis. *Science* **272,** 1641–1643 (1996).
24. Safi, H. *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes. *Nature genetics* **45,** 1190–1197 (2013).
25. Mori, G. *et al.* Thiophenecarboxamide Derivatives Activated by EthA Kill Mycobacterium tuberculosis by Inhibiting the CTP Synthetase PyrG. *Chemistry & Biology* **22,** 917–927 (2015).
26. Merker, M. *et al.* Whole genome sequencing reveals complex evolution patterns of multidrug-resistant Mycobacterium tuberculosis Beijing strains in patients. *PloS one* **8,** e82551 (2013).
27. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature genetics* **46,** 279–86 (2014).
28. Zhang, Y., Heym, B., Allen, B., Young, D. & Cole, S. The catalase—peroxidase gene and isoniazid resistance of Mycobacterium tuberculosis. *Nature* **358,** 591–593 (1992).
29. Larsen, M. H. *et al.* Overexpression of inhA, but not kasA, confers resistance to isoniazid and ethionamide in Mycobacterium smegmatis, M. bovis BCG and M. tuberculosis. *Molecular microbiology* **46,** 453–66 (2002).
30. Banerjee, A. *et al.* inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. *Science* **263,** 227–230 (1994).
31. DeBarber, a E., Mdluli, K., Bosman, M., Bekker, L. G. & Barry, C. E. Ethionamide activation and sensitivity in multidrug-resistant Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 9677–82 (2000).
32. Telenti, A. *et al.* The emb operon, a gene cluster of Mycobacterium tuberculosis involved in resistance to ethambutol. *Nature Medicine* **3,** 567–570 (1997).
33. Scorpio, A. & Zhang, Y. Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. *Nature Medicine* **2,** 662–667 (1996).
34. Shi, W. *et al.* Pyrazinamide Inhibits Trans-Translation in Mycobacterium tuberculosis. *Science* **333,** 1630–1632 (2011).
35. Shi, W. *et al.* Aspartate decarboxylase (PanD) as a new target of pyrazinamide in Mycobacterium tuberculosis. *Emerging Microbes & Infections* **3,** e58 (2014).
36. Perdigão, J. *et al.* GidB mutation as a phylogenetic marker for Q1 cluster Mycobacterium tuberculosis isolates and intermediate-level streptomycin resistance determinant in Lisbon, Portugal. *Clinical Microbiology and Infection* **20,** O278–O284 (2014).
37. Takiff, H. E. *et al.* Cloning and nucleotide sequence of Mycobacterium tuberculosis gyrA and gyrB genes and detection of quinolone resistance mutations. *Antimicrobial Agents and Chemotherapy* **38,** 773–80 (1994).
38. Kocagöz, T. *et al.* Gyrase mutations in laboratory-selected, fluoroquinolone-resistant mutants of Mycobacterium tuberculosis H37Ra. *Antimicrobial Agents and Chemotherapy* **40,** 1768–74 (1996).
39. Pasca, M. R. *et al.* Rv2686c-Rv2687c-Rv2688c, an ABC Fluoroquinolone Efflux Pump in Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy* **48,** 3175–3178 (2004).
40. Zaunbrecher, M. A., Sikes, R. D., Metchock, B., Shinnick, T. M. & Posey, J. E.

Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences* **106,** 20004–20009 (2009).
41. Awasthy, D., Bharath, S., Subbulakshmi, V. & Sharma, U. Alanine racemase mutants of Mycobacterium tuberculosis require D-alanine for growth and are defective for survival in macrophages and mice. *Microbiology* **158,** 319–327 (2012).
42. Köser, C. U. *et al.* Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis. *The New England journal of medicine* **369,** 290–2 (2013).
43. Schön, T. *et al.* Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. *Clinical Microbiology and Infection* **23,** 154–160 (2017).
44. Smith, T., Wolff, K. A. & Nguyen, L. Molecular biology of drug resistance in Mycobacterium tuberculosis. *Curr. Top. Microbiol. Immunol.* **374**, 53–80 (2013).
45. McNerney, R. *et al.* Removing the bottleneck in whole genome sequencing of Mycobacterium tuberculosis for rapid drug resistnce analysis: a call to action. *International Journal of Infectious Diseases* **56,** 130–135 (2017).

FIGURE LEGENDS

Figure 1. Geographical distribution of the 6,465 *Mycobacterium tuberculosis* isolates analysed in the study

This world map shows the main geographical origins of the *M. tuberculosis* isolates included in this study. The study comprises strains from more than 30 countries, of which the 18 major contributors are shown on this map. See **Supplementary table 1** for a detailed description of each dataset. Inner pie charts show the proportion of each of the main four lineages, and the outer charts summarise the drug resistance phenotypes. 'Drug-resistant' refers to non-MDR-TB/XDR-TB resistance.

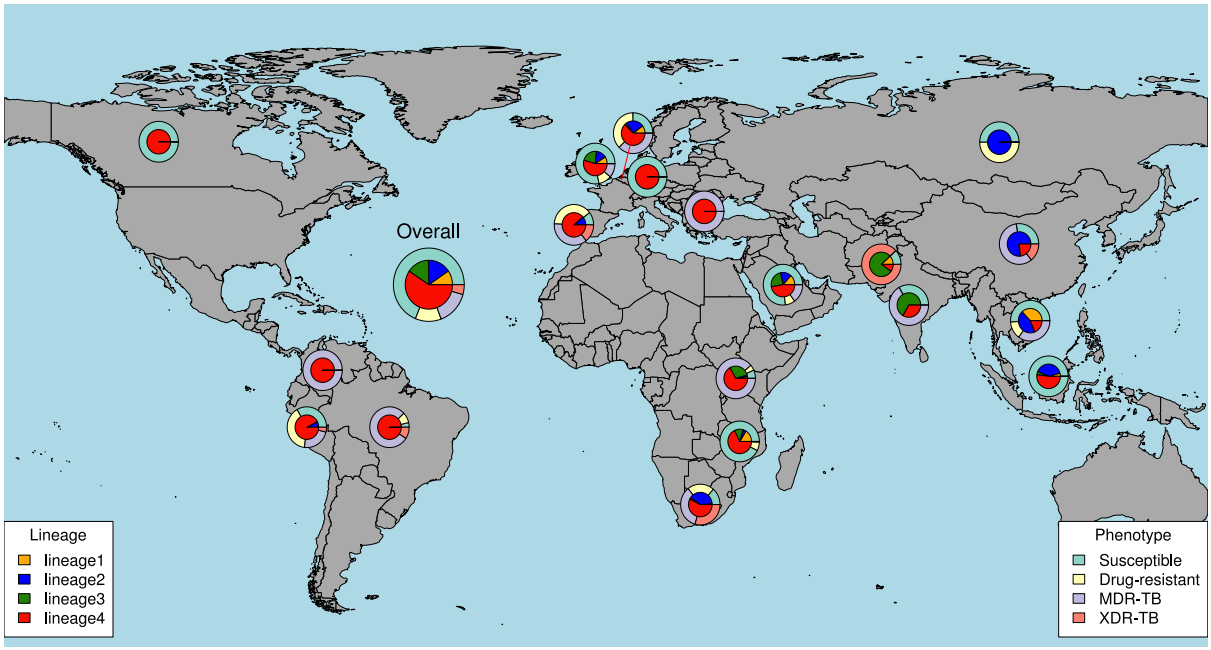

Figure 2. Whole genome phylogeny of the 6,465 *M. tuberculosis* isolates

Maximum likelihood phylogenetic tree constructed using 102,160 SNPs and 11,122 insertions and deletions spanning the whole genome and rooted on *M. canetti* (not shown), colour-coded by lineage (inner circle) and drug resistance status (outer circle). 'Susceptible' refers to isolates being susceptible to all drugs tested. 'Drug-resistant' refers to strains being resistant to multiple drugs but not classified as multidrug-resistant (MDR-TB) or extensively drug-resistant XDR-TB.

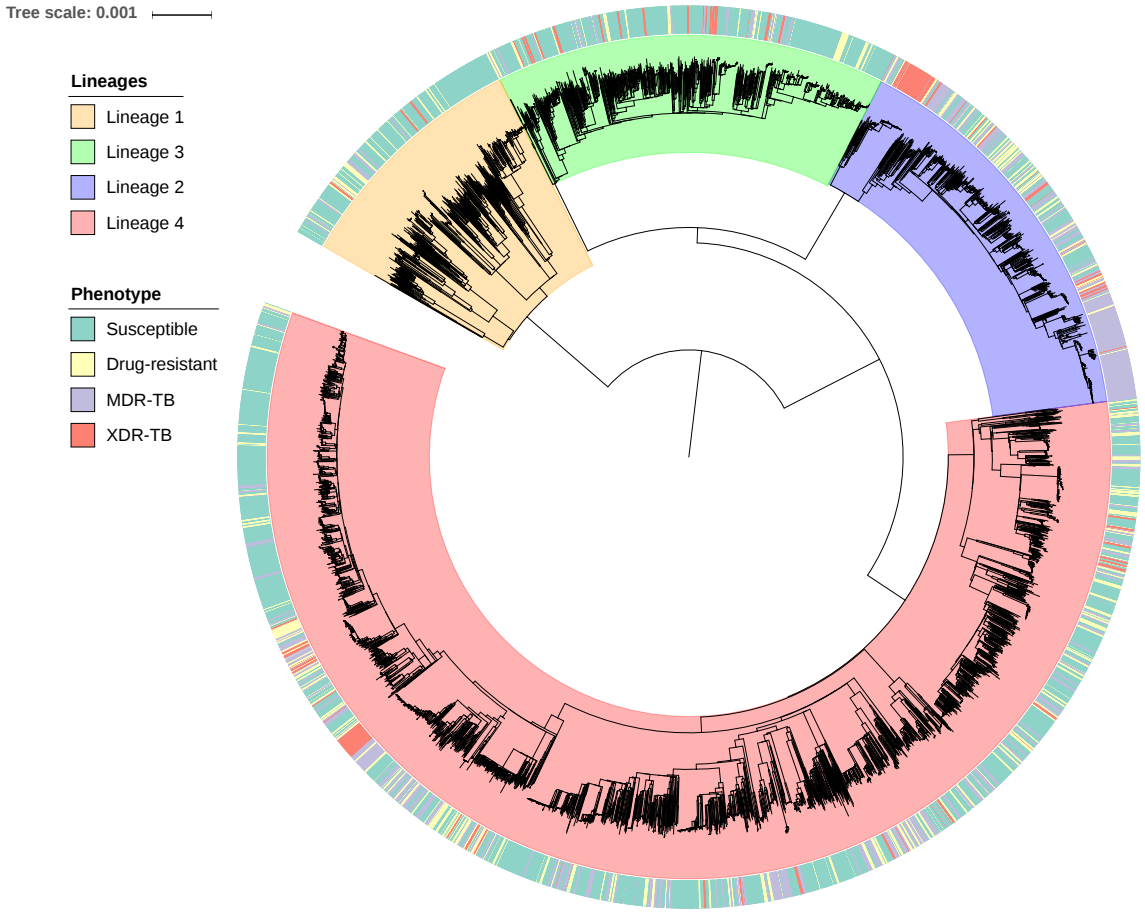

Figure 3

(Log) Odds ratios from SNP-drug resistance associations are a potential surrogate for resistance level. (A) Within each drug, boxplots for the log odds ratios ($P<1x10^{-5}$) for each gene are arranged by increasing median values (as indicated by the horizontal line in the boxes) to show their relative effect on resistance. Mutations known to confer low, intermediate or high levels of resistance (**See Online Methods**) are represented by points coloured blue, yellow or red, respectively, and their size is proportional to their frequency; overall, higher levels of resistance are reflected by higher odds ratios; one exception is for *rrs* and CAP, where the G1484C/T (high level resistance) mutation has a lower odds ratio than A1401G (intermediate level) due to its low frequency; a similar effect is seen for the same G1484C/T mutation in KAN resistance; **(B)** The distribution of (log) odds ratios ($P<1x10^{-5}$) for the mutations within unknown (n=167), or known low (n=17) (blue), intermediate (n=16)(yellow) or high (n=11)(red) levels of resistance; **(C)** The distribution of (log) odds ratios for known (n=171) and novel (n=40) drug resistance mutations ($P<1x10^{-5}$). All boxplots consist of boxes (median and interquartile range) and whiskers that extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box.

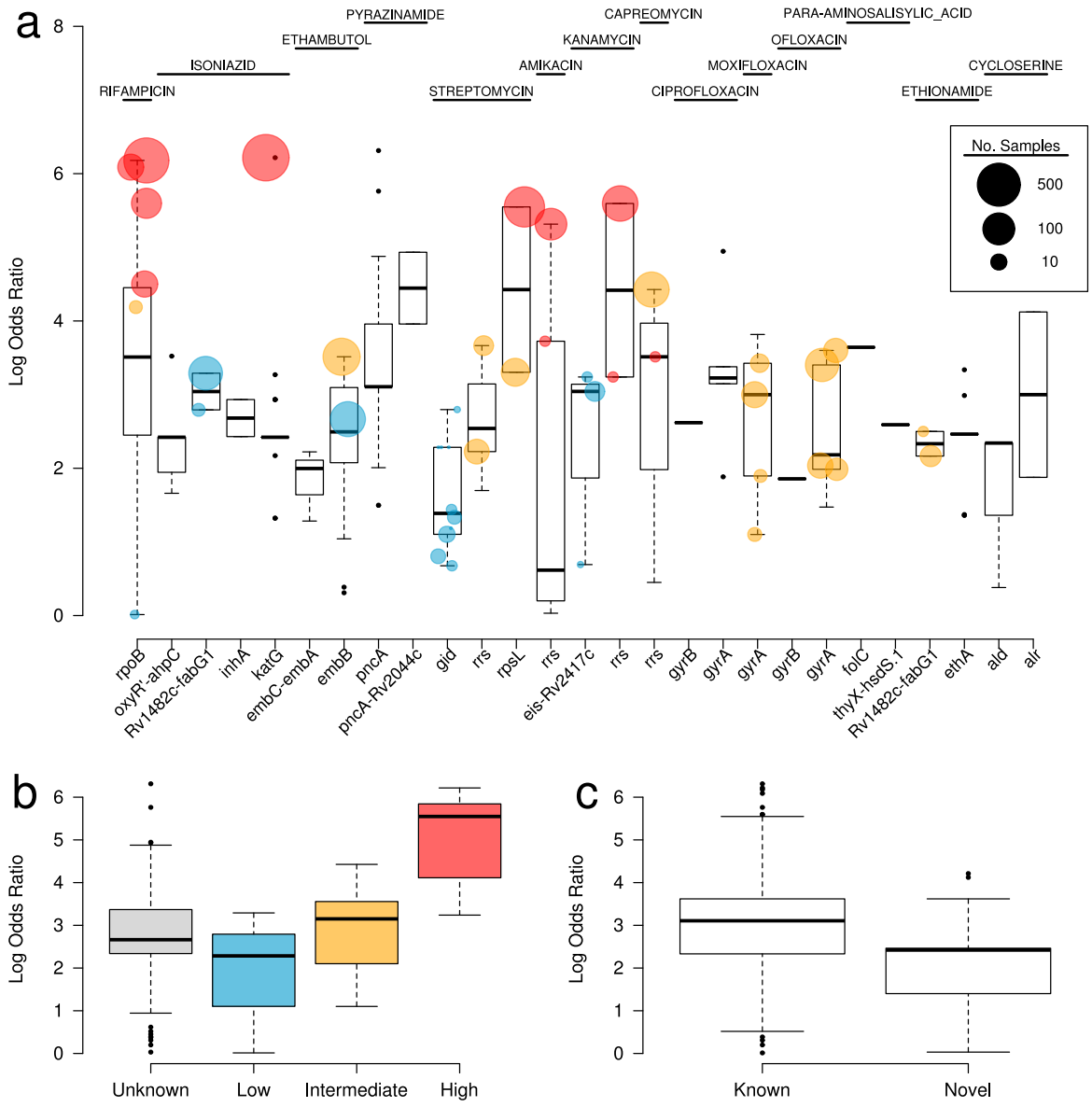
a
Log Odds Ratio
RIFAMPICIN
ISONIAZID
ETHAMBUTOL
PYRAZINAMIDE
STREPTOMYCIN
AMIKACIN
KANAMYCIN
CAPREOMYCIN
CIPROFLOXACIN
MOXIFLOXACIN
OFLOXACIN
PARA-AMINOSALISYLIC_ACID
ETHIONAMIDE
CYCLOSERINE
No. Samples
500
100
10
rpoB
oxyR'-ahpC
Rv1482c-fabG1
inhA
katG
embC-embA
embB
pncA
pncA-Rv2044c
gid
rrs
rpsL
rrs
eis-Rv2417c
rrs
rrs
gyrB
gyrA
gyrA
gyrB
gyrA
folC
thyX-hsdS.1
Rv1482c-fabG1
ethA
ald
alr
b
Log Odds Ratio
Unknown
Low
Intermediate
High
c
Log Odds Ratio
Known
Novel

Table 1
MDR-TB and XDR-TB gene-based associations

Comparison	Rv number	Gene name	P-value	NS SNPs	Indels (frame.)	Assoc. SNPs	PhyC SNPs
MDR-TB vs. Susc.	*Rv0667*	*rpoB*	2.99E-103	159	7 (0)	6	33
MDR-TB vs. Susc.	*Rv1908c*	*katG*	2.44E-65	177	12 (9)	2	8
MDR-TB vs. Susc.	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	1.28E-17	8	0	1	4
MDR-TB vs. Susc.	*Rv2427A-Rv2428*	*oxyR'-ahpC*	5.26E-15	17	3	0	7
MDR-TB vs. Susc.	*Rv3919c*	*gid*	1.09E-08	137	26 (26)	0	15
MDR-TB vs. Susc.	*Rv1484*	*inhA*	8.55E-07	9	0	0	3
MDR-TB vs. Susc.	*Rv0682*	*rpsL*	7.31E-06	6	0	0	2
XDR- vs. MDR-TB	*Rv0006*	*gyrA*	2.46E-37	147	0	4	5
XDR- vs. MDR-TB	*rrs*	*rrs*	4.33E-17	91	4	1	5
XDR- vs. MDR-TB	*Rv3806c*	*ubiA*	4.22E-07	47	0	1	1
XDR- vs. MDR-TB	*Rv3793-Rv3794*	*embC-embA*	8.73E-06	6	6	0	6
XDR-TB vs. Susc.	*Rv0667*	*rpoB*	4.13E-183	159	7 (0)	5	3
XDR-TB vs. Susc.	*Rv3795*	*embB*	1.54E-75	168	2 (0)	4	2
XDR-TB vs. Susc.	*Rv2043c*	*pncA*	4.33E-65	117	25 (22)	1	9
XDR-TB vs. Susc.	*Rv1908c*	*katG*	9.52E-60	177	12 (9)	1	1
XDR-TB vs. Susc.	*Rv3793-Rv3794*	*embC-embA*	1.07E-31	6	6	2	4
XDR-TB vs. Susc.	*rrs*	*rrs*	5.14E-28	91	4	2	3
XDR-TB vs. Susc.	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	1.98E-27	8	0	2	1
XDR-TB vs. Susc.	*Rv1484*	*inhA*	3.09E-26	9	0	1	1
XDR-TB vs. Susc.	*Rv0006*	*gyrA*	8.62E-26	147	0	4	5
XDR-TB vs. Susc.	*Rv0668*	*rpoC*	2.62E-21	153	1 (0)	1	9
XDR-TB vs. Susc.	*Rv0682*	*rpsL*	2.02E-18	6	0	1	3
XDR-TB vs. Susc.	*Rv3144c-Rv3145*	*PPE52-nuoA*	3.65E-11	24	1	1	2
XDR-TB vs. Susc.	*Rv3854c*	*ethA*	1.80E-10	163	38 (35)	0	1
XDR-TB vs. Susc.	*Rv2936*	*drrA*	1.46E-08	19	0	1	9
XDR-TB vs. Susc.	*Rv2416c-Rv2417c*	*eis-Rv2417c*	2.53E-07	12	1	0	3
XDR-TB vs. Susc.	*Rv1144-Rv1145*	*Rv1144-mmpL13a*	1.48E-07	33	4	1	2
XDR-TB vs. Susc.	*Rv3854c-Rv3855*	*ethA-ethR*	9.87E-06	12	0	1	0

This table shows loci (protein and RNA coding regions, intergenic regions) associated with MDR- and XDR-TB resistance ($P<1x10^{-5}$). The column labelled as 'NS SNPs' shows the number of non-synonymous SNPs in the genes; the column 'Indels (frame.)' refers to the number of small indels resulting in frameshifts in the genes; 'Assoc. SNPs' refers to the number of SNPs identified by GWAS and 'PhyC SNPs' is the number of homoplastic SNPs identified using the PhyC test. The PhyC test additionally detected *folC*, *pncA-Rv2044c* and *whiB6-Rv3863* loci when comparing MDR-TB against the susceptible group; and *eis-Rv2417c, gyrB, rrs, folC, alr, gid,* and the *thyX-hsdS.1* intergenic region when comparing XDR-TB against MDR-TB; and *alr, gyrB, pyrG, rpoA, and thyX-hsdS.1* loci when comparing XDR-TB against susceptible. Similarly, GWAS using SNPs additionally identified *embC-embA* for MDR-TB vs susceptible (1 SNP), *rrs* and *ubiA* genes for XDR-TB vs MDR-TB (each 1 SNP), and the *ubiA* gene for XDR-TB vs. susceptible (2 SNPs).

Table 2

Individual drug gene-based associations in the complete dataset

Drug*	Rv number	Gene name	P-value	NS SNPs	Indels (frame.)	Assoc. SNPs	PhyC SNPs
Isoniazid	*Rv1908c*	*katG*	1.02E-112	177	12 (9)	1	3
Isoniazid	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	5.41E-54	8	0	2	2
Isoniazid	*Rv2427A-Rv2428*	*oxyR'-ahpC*	8.51E-27	17	3	0	3
Isoniazid	*Rv1484*	*inhA*	3.29E-07	9	0	1	1
Rifampicin	*Rv0667*	*rpoB*	8.47E-226	159	7 (0)	7	9
Rifampicin	*Rv0668*	*rpoC*	2.57E-08	153	1 (0)	0	9
Ethambutol	*Rv3795*	*embB*	2.48E-129	168	2 (0)	4	10
Ethambutol	*Rv3793-Rv3794*	*embC-embA*	8.49E-42	6	6	2	5
Ethambutol	*Rv3806c*	*ubiA*	3.93E-13	47	0	1	2
Ethambutol	*Rv2820c*	*.*	2.55E-08	16	0	1	0
Ethambutol	*Rv3300c*	*.*	1.33E-07	39	5 (3)	0	0
Ethionamide	*Rv1482c-Rv1483*	*Rv1482c-fabG1*	6.01E-16	8	0	2	2
Ethionamide	*Rv1484*	*inhA*	6.72E-07	9	0	1	0
Pyrazinamide	*Rv2043c*	*pncA*	3.62E-99	117	25 (22)	2	1
Pyrazinamide	*Rv2043c-Rv2044c*	*pncA-Rv2044c*	6.64E-30	4	1	1	1
Streptomycin	*Rv0682*	*rpsL*	2.67E-85	6	0	2	2
Streptomycin	*Rv3919c*	*gid*	3.54E-26	137	26 (26)	0	1
Streptomycin	*rrs*	*rrs*	3.95E-13	91	4	1	3
Amikacin	*rrs*	*rrs*	5.28E-48	91	4	1	1
Kanamycin	*rrs*	*rrs*	1.76E-48	91	4	2	2
Kanamycin	*Rv2416c-Rv2417c*	*eis-Rv2417c*	9.84E-21	12	1	1	1
Capreomycin	*rrs*	*rrs*	1.68E-39	91	4	1	1
Capreomycin	*Rv2172c-Rv2173*	*Rv2172c-idsA2*	7.18E-06	18	0	0	0
Ciprofloxacin	*Rv0006*	*gyrA*	4.48E-45	147	0	2	2
Moxifloxacin	*Rv0006*	*gyrA*	2.98E-23	147	0	3	5
Ofloxacin	*Rv0006*	*gyrA*	4.87E-115	147	0	4	6
D-Cycloserine	*Rv3423c*	*alr*	1.23E-13	57	0	1	0
D-Cycloserine	*Rv0342*	*iniA*	3.36E-08	76	13 (12)	1	0
PAS	*Rv2764c*	*thyA*	3.74E-10	36	4 (4)	0	0
PAS	*Rv2754c-Rv2755c*	*thyX-hsdS.1*	4.27E-07	21	0	1	1

This table shows loci (protein and RNA coding and intergenic regions) associated with resistance to individual drugs ($P<1x10^{-5}$). The column labelled as 'NS SNPs' shows the number of non-synonymous SNPs in the genes; the column 'Indels (frame.)' refers to the number of small indels resulting in frameshifts in the genes; 'Assoc. SNPs' is the number of SNPs identified by GWAS, and 'PhyC SNPs' refers to the number of homoplastic SNPs identified using the PhyC test. * The GWAS additionally detected a significant association of a SNP (C213R) in the *Rv2688c l*ocus (known efflux gene) with Moxifloxacin and Fluoroquinolones; the PhyC test additionally detected other associated loci for Amikacin (*eis-Rv2417c*), Capreomycin and D-Cycloserine (*lhr),* Kanamycin (*thyX-hsdS.1*), Rifampicin (*rpoA*). Abbreviations: PAS, para-aminosalicylic acid.

Table 3

Impact on drug resistance prediction (%) from GWAS findings

Drug	TBDR panel		+ SNPs		+ small indels + SNPs		+ big deletions + small indels + SNPs	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Isoniazid	89	97	89	97	**90**	97	90	97
Rifampicin	92	98	92	98	**93**	98	93	98
Ethambutol	90	92	90	92	90	92	90	92
Ethionamide	64	78	**78**	74	**84**	72	**88**	72
Pyrazinamide	52	98	52	98	**63**	97	**65**	97
Streptomycin	76	93	76	93	**80**	91	80	91
Amikacin	83	96	83	96	**85**	93	85	93
Kanamycin	84	98	84	98	84	98	84	98
Capreomycin	75	96	75	96	**81**	95	81	95
Ciprofloxacin	89	98	89	98	89	98	89	98
Moxifloxacin	85	90	85	90	85	90	85	90
Ofloxacin	86	96	86	96	86	96	86	96
D-Cycloserine	-	-	**55**	**92**	**61**	90	61	90
PAS	10	100	**20**	99	**40**	94	**65**	94
MDR-TB	87	100	87	100	**88**	100	**89**	100
XDR-TB	77	99	**78**	99	**79**	98	79	98

This table shows the sensitivity and specificity achieved by known drug resistance SNPs and indels (TBDR, tbdr.lshtm.ac.uk)[9, 31] when predicting phenotypic drug resistance ("TBDR panel" columns). The SNPs in the TBDR contribute 100% to the stated sensitivity, except rifampicin (99.8%) and ethionamide (99.3%). The other columns show the improvements achieved when including the SNPs, small indels and large deletions found associated with drug resistance in this study. The improvements in sensitivity are highlighted in grey. Abbreviations: MDR-TB, multidrug-resistant; PAS Para-aminosalicylic acid; Sens., sensitivity; Spec., specificity; SNPs, single nucleotide polymorphisms; XDR-TB, extensively drug-resistant.

ONLINE METHODS

Sequence data and variant calling

Sequence data for 6,465 *Mycobacterium tuberculosis complex* clinical isolates were generated as part of a collaborative global drug resistance project (n=2,637, pathogenseq.lshtm.ac.uk) or downloaded from the public domain (n=3,828) (**Supplementary table 1**). All isolates had undergone drug susceptibility testing by phenotypic methods. These isolates represented multiple populations from different geographic areas, and all four main lineages (1 to 4) (**Supplementary table 1**). The 2,637 samples not previously sequenced were Illumina sequenced generating paired-end reads of at least 50 bp with at least 50-fold genome coverage. The analytical workflow for the raw sequence data is summarised in **Supplementary figure 5**. The new and archived raw sequence data were aligned to the H37Rv reference genome (Genbank accession number: NC_000962.3) using the *BWA mem* algorithm[46] (settings: –c 100 –T 50)**.** The *SAMtools/BCFtools*[47] (default settings) and *GATK*[48] software were used to call SNPs and small indels. The GATK parameters used are "-T UnifiedGenotyper -ploidy 1 -glm BOTH -allowPotentiallyMisencodedQuals 2". The overlapping set of variants from the two algorithms was retained for further analysis. Alleles were additionally called across the whole genome (including SNP sites) using a coverage-based approach[5,49]**.** A missing call was assigned if the total depth of coverage at a site did not reach a minimum of 20 reads or none of the four nucleotides accounted for at least 75% of the total coverage. Samples or SNP sites having an excess of 10% missing genotype calls were removed. This quality control step was implemented to remove samples with bad quality genotype calls due to poor

depth of coverage or mixed infections. The final dataset included 6,465 isolates and 102,160 genome-wide SNPs. *Delly2* software[50] was used to identify large deletions. All large deletions were confirmed using localised *de novo* assembly, and those found in association analysis (*dfrA/thyA, pncA, ethA/ethR, katG*) confirmed using PCR.

Phenotypic drug susceptibility testing

Drug susceptibility data was obtained from World Health Organisation recognised testing protocols[51]**.** The *M. tuberculosis* (Mtb) isolates that provided sequence data included in this study are summarised in **Supplementary table 1.** Each sequence included in the study was derived from an isolate from an individual patient. Some DNA samples were from archived stocks (e.g. India, collected prior to 2009 and Malawi, collected between 1996 and 2010) and others were extracted specifically for this study. Information regarding isolates with previously reported sequence data was derived from published materials. Isolates were classed as resistant or susceptible to a drug on the basis of phenotypic testing using either the BACTEC 460 TB System (Becton Dickinson), the BACTEC Mycobacterial Growth Indicator Tube (MGIT) 960 system (Becton Dickinson)[52], solid agar or Lowenstein Jensen slopes[53,54]**.** Not all samples were tested for resistance to all drugs, most notably some isolates found susceptible to the first-line drugs were not subjected to testing for resistance to second-line drugs. Where isolates were not tested for resistance to a particular drug they were excluded from the analysis for that drug. Drug susceptibility testing was mainly undertaken in local laboratories participating in the WHO supranational laboratory network using the recognised testing protocols[51]**.** Isolates from Malawi were shipped to the United Kingdom's Mycobacterium Reference Laboratory for testing. Isolates from Uganda were tested at the Joint Clinical Research Centre (JCRC) in Kampala with quality control performed by the US

Centers for Disease Control and Prevention (CDC). The Peruvian isolates were initially tested for resistance to rifampicin and isoniazid using the Microscopic Observation Drug Susceptibility assay (MODS)[54] at the Universidad Peruana Cayetano Heredia (UPCH) prior to transfer to the national reference laboratory for further testing. In Peru susceptibility to pyrazinamide (PZA) was assessed by the Wayne assay; a colorimetric biochemical test during which PZA is hydrolysed to free pyrazinoic acid[55]. Testing using the BACTEC 960® MGIT® or BACTEC 460® (Becton-Dickinson®) was performed according to the manufacturer's indications[56]**.** Pyrazinamide sensitivity was determined by using BACTEC 7H12 liquid medium, pH 6.0, at 100 µg/mL (BACTEC PZA test medium, Becton Dickinson). When testing on agar critical drug concentrations used were rifampicin 1 µg/mL, isoniazid 0.2 µg/mL, streptomycin 2 µg/mL, and ethambutol 5 µg/mL, ciprofloxacin 2 µg/mL, amikacin 5µg/mL, capreomycin 10 µg/mL, kanamycin 5 µg/mL (Pakistan 6 µg/mL), ethionamide 5 µg/mL and para-aminosalicylic acid 2 µg/mL[53]**.** For Lowenstein-Jensen drug concentrations used were for streptomycin 4.0 µg/ml, isoniazid 0.2 µg/ml, rifampicin 40.0 µg/ml, ethambutol 2.0 µg/ml, capreomycin 40.0 µg/ml, kanamycin 30.0 µg/ml (China) or 20.0 µg/ml (Vietnam), ofloxacin 2.0 µg/ml, ethionamide 40 µg/ml, thioacetone (10 µg/ml), pyrazinamide 200 µg/ml, cycloserine 30 µg/ml and *para*-aminosalicylic acid (PAS) 0.5 µg/ml[55].

Phylogenetic tree and association analysis

The best-scoring maximum likelihood phylogenetic tree rooted on *Mycobacterium canettii* (Genbank accession number: HE572590) was constructed by *RAxML* software[57] (10,000 bootstrap samples) using the 102,160 high quality SNP sites. Spoligotypes were inferred *in silico* using *SpolPred*[58]**,** and strain-types determined using lineage-specific SNPs[5]**.** Further population structure assessment was performed using principal components analysis

(**Supplementary figure 2)**, which clustered samples by genotype congruent with the phylogenetic tree. The principal components were calculated from a SNP pair-wise distance matrix between each sample, and the first five components (summarising 82.7% of genetic variation) were used as covariates in the regression-based association models. Mixed regression models were employed to estimate the strength of association between the binary drug resistance outcome (resistance vs. susceptible) and the aggregate number of mutations (SNPs, indels or large deletions) by coding region, RNA loci and intergenic regions, as well as operons[49]**.** The low frequency of variants required the aggregation of mutations to increase the power of detecting associated loci**,** and a mixed model approach has been demonstrated to work well at adjusting for the confounding effects of Mtb lineage, sub-lineage and outbreak-based population structure.[49] The operons or functional units containing clusters of genes under the control of the same promoter were determined from TBDB (see URLs). Gene function was extracted from the Tuberculist webserver (see URLs). The mixed models also included the principal components to account for the main Mtb lineage and sub-lineage effects, and a SNP inferred kinship matrix as a random effect to account for highly related samples and fine-scale population structure due to potential outbreaks[49]. These models were implemented in GEMMA (v.1.1.2) software[59]**.** A SNP-based GWAS was used to identify individual variants associated with drug resistance expected to fall within the genes found associated in the 'main' analysis. To minimise any co-resistance between drugs, we adjusted for the presence of other resistance in the regression models. Co-resistance is expected to result from exposure to multiple anti-tuberculous drugs and the step-wise accumulation of mutations. Statistical significance thresholds to account for multiple testing were established using a permutation approach that sorted phenotypic test data without replacement and re-performed GWAS analysis (10,000 times). We report all

findings that are below a calculated permutation threshold of $P<1x10^{-5}$. All statistical analyses were performed using R software. To identify SNPs enriched by convergent evolution and provide further evolutionary evidence, the phylogenetic-based *phyC* approach was employed[9] using the implementation made available in a previous study[60]. Any potential co-resistance effects were dissected through consulting gene annotation and published literature to report the most plausible role in drug resistance. Additionally, long branches in the phylogenetic tree leading up to clades enriched with drug resistant isolates leads to spurious associations. Truly drug resistant mutations often originate multiple times independently in the phylogeny. Mutations which originated once in the tree (i.e. clade-specific mutations), which are likely to lead to spurious associations, were removed from the GWAS results.

Detection of putative compensatory mechanisms

Loci were identified as being putative compensatory if they: (i) were associated with drug resistance, (ii) harboured homoplastic mutations, (iii) shared a similar biological function with a known drug-target or drug-activating enzyme, and (iv) were significantly more mutated in the presence of mutations in the drug-target or drug-activating enzyme coding gene. For the fourth analysis, deep phylogenetic and synonymous SNPs were removed before calculating the number of samples with nonsynonymous SNPs at genes of interest (e.g. Ala1075Ala at *rpoB* or Glu1092Asp at *rpoC*). The significance of differences between studied genes was calculated using Fisher's exact test (cut-off of $P<1x10^{-8}$).

Protein mutation modelling

Apo crystal structures for *alr* were downloaded from the Protein Data Bank (PDBe1XFC[61]) and then subjected to modelling of missing residues, WinCOOT regularisation, and removal of pyridoxal 5´-phosphate from both chains. The mCSM and DUET web servers were used to

assess changes in protein stability, mCSM-PPI to quantify effects on protein-protein interactions and mCSM-Lig to quantify effects on drug binding[62–64]. For ligand binding, D-Cycloserine was modelled in the active site using UCSF Chimera v1.11[65] from the coordinates of the closest holo-homolog *Clostridium difficile* 630 (PDBe4LUT[66]).

Statistical analyses

The statistical mixed models used for association analysis are described above. The terms 'low', 'intermediate' or 'high' levels of resistance referred to in the text and **Figure 3** denote whether a mutation is known to confer low, intermediate or high MIC values, respectively, as reported in the literature[18,40,67–71]. Wilcoxon tests and linear regression models were used to compare differences in (log) odds ratios between resistance levels. Samples which had more than one known resistance causing variant were removed from these calculations. R statistical software (v3.4.1; see URLs) was used to perform this analysis. The R library "maps" was used to generate the world map with lineage and drug resistance frequencies.

DATA AVAILABILITY

All raw sequencing data are available, and the study accession numbers are listed in **Supplementary table 1.** For samples sequenced as part of our collaborative global drug resistance project, the ENA accession numbers for the isolates and their phenotypic drug susceptibility data are provided in **Supplementary data 2**.

URLs

The TubercuList knowledge base, http://tuberculist.epfl.ch; Tuberculosis Database, http://www.tbdb.org; R Statistical software, https://www.r-project.org; TB Global Drug

Resistance Collaboration, http://pathogenseq.lshtm.ac.uk/#tuberculosis); MycoBrowser, https://mycobrowser.epfl.ch/

METHODS-ONLY REFERENCES

46. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30,** 2843–2851 (2014).
47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
48. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43,** 491–8 (2011).
49. Phelan, J. *et al.* Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC medicine* **14,** 31 (2016).
50. Rausch, T. *et al.* DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).
51. World Health Organization. WHO | Guidelines for surveillance of drug resistance in tuberculosis. (2009).
52. Kubica, G. & Kent, K. Public health mycobacteriology: a guide for the level III laboratory. *Centers for Disease Control, U.S. Department of Health and Human Services, Atlanta, GA* 60–63 (1985).
53. Canetti, G. *et al.* Mycobacteria: Laboratory Methods For Testing Drug Sensitivity and Resistance. *Bulletin of the World Health Organization* **29,** 565–78 (1963).
54. Minion, J., Leung, E., Menzies, D. & Pai, M. Microscopic-observation drug susceptibility and thin layer agar assays for the detection of drug resistant tuberculosis: a systematic review and meta-analysis. *The Lancet Infectious Diseases* **10,** 688–698 (2010).
55. Wayne, L. G. Simple pyrazinamidase and urease tests for routine identification of mycobacteria. *The American review of respiratory disease* **109,** 147–51 (1974).
56. Palicova, F., Jahn, E. I. M. & Pfyffer, G. E. Susceptibility Testing of Mycobacterium tuberculosis to Anti-Tuberculosis Drugs: BACTEC ™ MGIT ™ 960 vs BACTEC ™ 460TB System.
57. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57,** 758–71 (2008).
58. Coll, F. *et al.* SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics* **28,** 2991–3 (2012).
59. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44,** 821–4 (2012).
60. Alam, M. T. *et al.* Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome biology and evolution* **6,** 1174–85 (2014).
61. Velankar, S. *et al.* PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Research* **44,** D385–D395 (2016).
62. Pires, D. E. V, Ascher, D. B. & Blundell, T. L. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30,** 335–342 (2014).

63. Pires, D. E. V., Blundell, T. L. & Ascher, D. B. mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Scientific Reports* **6,** 29575 (2016).
64. Pires, D. E. V, Ascher, D. B. & Blundell, T. L. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Research* **42,** W314–W319 (2014).
65. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*, **25** (13), 1605-12 (2004).
66. Asojo, O. A.; Nelson, S. K.; Mootien, S.; Lee, Y.; Rezende, W. C.; Hyman, D. A.; Matsumoto, M. M.; Reiling, S.; Kelleher, A.; Ledizet, M.; Koski, R. A.; Anthony, K. G., Structural and biochemical analyses of alanine racemase from the multidrug-resistant Clostridium difficile strain 630. *Acta crystallographica*. Section D, Biological crystallography, **70** (Pt 7), 1922-33 (2014).
67. Wong, S. Y. *et al.* Mutations in gidB Confer Low-Level Streptomycin Resistance in Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy* **55,** 2515–2522 (2011).
68. Rueda, J. *et al.* Genotypic Analysis of Genes Associated with Independent Resistance and Cross-Resistance to Isoniazid and Ethionamide in Mycobacterium tuberculosis Clinical Isolates. *Antimicrobial Agents and Chemotherapy* **59,** 7805–7810 (2015).
69. Kambli, P. *et al.* Correlating rrs and eis promoter mutations in clinical isolates of Mycobacterium tuberculosis with phenotypic susceptibility levels to the second-line injectables. *International Journal of Mycobacteriology* **5,** 1–6 (2016).
70. Domínguez, J. *et al.* Clinical implications of molecular drug resistance testing for Mycobacterium tuberculosis; a TBNET/RESIST-TB consensus statement. *Int. J. Tuberc. Lung Dis.* **20**, 24–42 (2016).
71. Cambau, E. *et al*. Revisiting susceptibility testing in MDR-TB by a standardized quantitative phenotypic assessment in a European multicentre study. *J. Antimicrob. Chemother.* **70**, 686–696 (2015).