nature genetics

ARTICLES

https://doi.org/10.1038/s41588-017-0029-0

Corrected: Author Correction

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[1,6], Yaa Oppong[1], Adriana Alves[7], Theolis Barbosa Bessa[8], Susana Campino[1], Maxine Caws[9,10], Anirvan Chatterjee[11], Amelia C. Crampin[12,13], Keertan Dheda[14], Nicholas Furnham[1], Judith R. Glynn[12,13], Louis Grandjean[15], Dang Minh Ha[10], Rumina Hasan[16], Zahra Hasan[16], Martin L. Hibberd[1], Moses Joloba[17], Edward C. Jones-López[18], Tomoshige Matsumoto[19], Anabela Miranda[7], David J. Moore[1,15], Nora Mocillo[20], Stefan Panaiotov[21], Julian Parkhill[22], Carlos Penha[23], João Perdigão[24], Isabel Portugal[24], Zineb Rchiad[2], Jaime Robledo[25], Patricia Sheen[14], Nashwa Talaat Shesha[26], Frik A. Sirgel[27], Christophe Sola[28], Erivelton Oliveira Sousa[8,29], Elizabeth M. Streicher[27], Paul Van Helden[27], Miguel Viveiros[30], Robert M. Warren[27], Ruth McNerney[1,14]*, Arnab Pain[2,31]* and Taane G. Clark[1,12]*

To characterize the genetic determinants of resistance to antituberculosis 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 phylogenetics-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. New epistatic relationships between candidate drug-resistance-associated genes were identified. Findings also suggest the involvement of efflux pumps (*drrA* and *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.

The emergence and spread of *Mycobacterium tuberculosis* (Mtb) resistant to multiple antituberculosis drugs is of global concern. Programmatically incurable tuberculosis (TB), where effective treatment regimens cannot be provided owing to resistance to the available drugs, is a growing problem[1]. Resistance to rifampicin and isoniazid is classed as multidrug-resistant tuberculosis

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [2]Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. [3]Sydney Emerging Infections and Biosecurity Institute and School of Public Health, Sydney Medical School, University of Sydney, Sydney, New South Wales, Australia. [4]Laboratory Medicine Department, Faculty of Applied Medical Sciences, Umm Al-Qura University, Makkah, Saudi Arabia. [5]Department of Microbiology, Faculty of Medicine, Umm Al-Qura University, Makkah, Saudi Arabia. [6]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Melbourne, Victoria, Australia. [7]National Mycobacterium Reference Laboratory, Porto, Portugal. [8]Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil. [9]Liverpool School of Tropical Medicine, Liverpool, UK. [10]Pham Ngoc Thach Hospital for TB and Lung Diseases, Ho Chi Minh City, Vietnam. [11]Foundation for Medical Research, Mumbai, India. [12]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [13]Karonga Prevention Study, Chilumba, Karonga, Malawi. [14]Lung Infection and Immunity Unit, UCT Lung Institute, University of Cape Town, Groote Schuur Hospital, Cape Town, South Africa. [15]Laboratorio de Enfermedades Infecciosas, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Peru. [16]Department of Pathology and Laboratory Medicine, Aga Khan University, Karachi, Pakistan. [17]Department of Medical Microbiology, Makerere University College of Health Sciences, Kampala, Uganda. [18]Section of Infectious Diseases, Department of Medicine, Boston Medical Center and Boston University School of Medicine, Boston, MA, USA. [19]Osaka Anti-Tuberculosis Association, Osaka Hospital, Osaka, Japan. [20]Reference Laboratory of Tuberculosis Control, Buenos Aires, Argentina. [21]National Center of Infectious and Parasitic Diseases, Sofia, Bulgaria. [22]Wellcome Trust Sanger Institute, Hinxton, UK. [23]Instituto Gulbenkian de Ciência, Lisbon, Portugal. [24]iMed.ULisboa–Research Institute for Medicines, Faculdade de Farmácia, Universidade de Lisboa, Lisbon, Portugal. [25]Corporación para Investigaciones Biológicas, Universidad Pontificia Bolivariana, Medellín, Colombia. [26]Regional Laboratory Directorate of Health Affairs, Makkah, Saudi Arabia. [27]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. [28]Institute for Integrative Cell Biology, CEA, CNRS, Université Paris-Saclay, Orsay, France. [29]Laboratorio Central de Saúde Pública Professor Gonçalo Moniz, Salvador, Brazil. [30]Unidade de Microbiologia Médica, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa (UNL), Lisbon, Portugal. [31]Global Station for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, Sapporo, Japan. Francesc Coll and Jody Phelan contributed equally to this work. Ruth McNerney, Arnab Pain and Taane G. Clark jointly directed this work. *e-mail: ruth.mcnerney@uct.ac.za; arnab.pain@kaust.edu.sa; taane.clark@lshtm.ac.uk

(MDR-TB), and 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 tuberculosis (XDR-TB). Treatment for patients with drug-resistant tuberculosis is prolonged and expensive, and outcomes are poor[2]. The drugs used are toxic and poorly tolerated, and 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 of 4.4 Mb) 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 (for example, East African–Indian (EAI) spoligotype families); lineage 2, East Asian (for example, W/Beijing spoligotype families); lineage 3, East African–Indian (for example, Central Asian strain (e.g., CAS-DELHI) spoligotype families); and lineage 4, Euro-American (for example, Latin American–Mediterranean (LAM), Haarlem and the 'ill-defined' T spoligotype families)[5].

Resistance in Mtb is mainly conferred by nucleotide variations (SNPs and indels) in genes encoding drug targets or drug-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 antituberculosis 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 of clinical isolates ($n=6,465$) from more than 30 geographic locations, representing the four major Mtb lineages (Fig. 1 and 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 were 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 *para*-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 antituberculosis 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 Fig. 1). Similarly, small indels were rare (96.6% had frequency < 1%) and ranged in size from 1 to 45 bp. A phylogenetic tree and principal-component analysis constructed using all genome-wide SNPs showed the expected clustering by lineage (Fig. 2 and Supplementary Fig. 2).

Phenotypic analysis of susceptibility to antituberculosis drugs found that 31.2% of isolates were resistant to at least one drug, with 15.1% categorized as MDR-TB and 4.3% categorized as XDR-TB (Fig. 2 and Supplementary Table 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 (SLIDs: AMK, KAN and CAP) and fluoroquinolones (FLQs: CIP, OFL and MOX) were also analyzed. Insufficient phenotypic data were available for

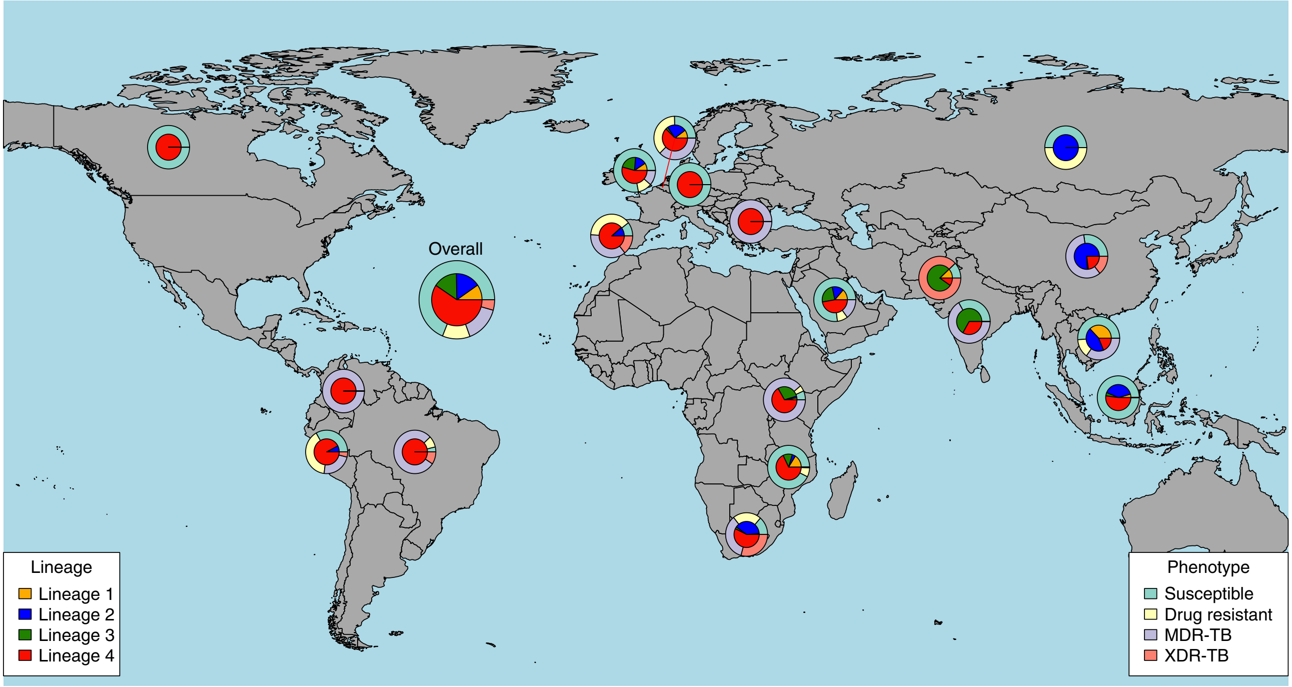

Fig. 1 | Geographic distribution of the 6,465 *Mycobacterium tuberculosis* isolates analyzed in the study. The world map shows the main geographic origins of the Mtb isolates included in this study. The study comprises strains from more than 30 countries, of which the 18 major contributors are shown. 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 summarize the drug resistance phenotypes. "Drug-resistant" refers to resistant strains not classified as MDR-TB or XDR-TB.

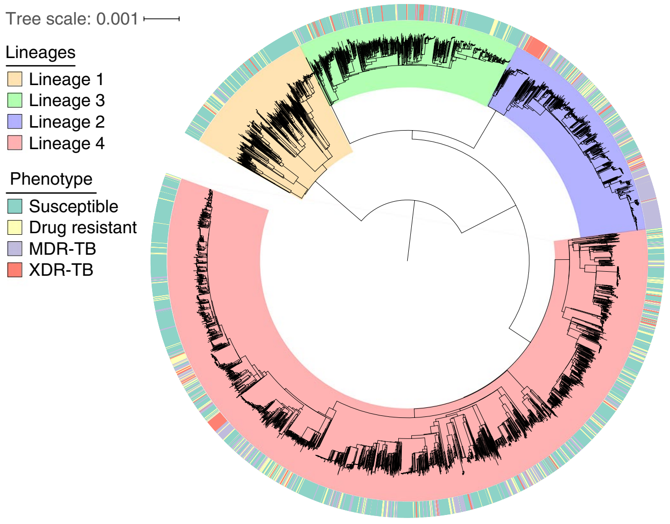

Fig. 2 | Whole-genome phylogeny of the 6,465 *Mycobacterium tuberculosis* isolates. Maximum-likelihood phylogenetic tree constructed using 102,160 SNPs and 11,122 indels spanning the whole genome and rooted on *Mycobacterium canetti* (not shown), with isolates color-coded by lineage (inner circle) and 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 MDR-TB or XDR-TB.

inclusion of the new and repurposed drugs bedaquiline, delamanid and linezolid. To identify loci associated with drug resistance, complementary methods were applied to mutations and aggregated nonsynonymous mutations: a tree-based 'PhyC' test for convergent evolution to detect homoplastic variants[9] and a GWAS approach within a mixed-regression framework (Methods). Unless stated otherwise, all analysis used the complete dataset. First, we considered 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 isolates identified *rpoB* (RIF), the *Rv1482c–fabG1* operon (INH, ETH), *inhA* (INH, ETH), *katG* (INH) and *oxyR´–ahpC* (compensatory mechanism for INH). The *katG* mutations at codon 315 (encoding p.Ser315Thr, p.Ser315Asn, p.Ser315Arg) were all statistically significant and collectively were the most frequent mutations (75.2%) across all the resistance-associated loci identified, consistent with a recent study[14] and highlighting their pivotal role in the emergence of INH resistance and MDR-TB. The *katG* mutation encoding p.Ser315Thr is thought to emerge before mutations associated with RIF resistance and, therefore, from an evolutionary standpoint, to precede the emergence of MDR-TB[14,15]. However, our analysis highlighted that *Rv1482c–fabG1* and *inhA* mutations, in the absence of *katG* p.Ser315Thr, can emerge before MDR-TB, as previously shown in two phylogenetically independent clades in Lisbon[16,17]. The other frequent MDR-TB-associated mutations in our study included *rpoB* p.Ser450Leu (RIF, 64.2%), *embB* p.Met306Leu, p.Met306Val and p.Met306Ile (EMB, 49.1%), and *rpsL* p.Lys43Arg (STM, 42.2%) (Supplementary Table 3), and the prevalence correlated with historical treatment practice and emergence of resistance. There were corresponding signals of INH–RIF co-resistance with resistance for other first-line drugs, and gene-based association signals were detected for *gid* (STM) and *rpsL* (STM) and a SNP-based association signal was detected for the *embC–embA* intergenic region (EMB). SNP-based PhyC analysis detected the above loci but in addition identified *folC* (PAS), the *pncA–Rv2044c* intergenic region (PZA) and the *whiB6–Rv3863* intergenic region (putative STM or ETH).

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) and *gid* (STM) SNPs and a new mutation in the *thyX–hsdS.1* intergenic region (c.–9A>T; PAS)[18,19]. In addition to the 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* (a new intergenic region; c.–314G>T). The PhyC test additionally detected SNPs in *gyrB* (FLQ; encoding p.Asp461Asn, p.Asp641His, p.Thr500Asn, p.Thr500Ile and p.Ala504Val) and supported the SNP finding in the *thyX–hsdS.1* intergenic region (PAS; c.–9A>T), as well as identified a previously unreported *ubiA* SNP association (EMB; p.Met180Val).

The *drrA* mutation encoding p.Arg262Gly was significantly associated with XDR-TB as compared to susceptible bacteria (mutation frequency = 18% versus 0%, $P = 1.5 \times 10^{-8}$). We hypothesize that *drrA* may be involved in export of drugs across the membrane on the basis of 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 through increased efflux activity and lipid metabolism[20,21]. Similarly, a mutation in the *Rv1144–mmpL13a* intergenic region (c.–102C>A) was highly associated with XDR-TB versus susceptibility (mutation frequency = 17% versus 0%, $P = 1.5 \times 10^{-7}$). This mutation sits in the promoter of the operon containing *mmpL13a* and *mmpL13b*, which encode transmembrane transport proteins, and thus could influence expression of the encoded proteins[6].

Lineage-specific and compensatory mechanisms. We conducted GWAS stratified by 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 mutations associated with drug resistance 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 PZA (*pncB1*) through analysis of lineage 1, and this locus reached borderline significance for lineage 3.

We applied a systematic approach to identify epistatic interactions between GWAS loci (from Table 2) and explored known compensatory effects using a test of non-random association to detect the frequent co-occurrence of mutations in pairs of loci (Fisher's exact test, P-value cutoff $< 1 \times 10^{-8}$) (Supplementary Table 5). Deep phylogenetic mutations were removed to increase robustness. This approach proved to be successful at identifying well-known compensatory relationships between the *rpoB* and *rpoC* loci (RIF)[7], the *rpoB* and *rpoA* loci (RIF)[22], and the *katG* and *oxyR´–ahpC* loci (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 these mutations are therefore unlikely to represent a compensatory mechanism. New epistatic relationships included *pncA* with *pncB2* (PZA) and *thyA* with *thyX–hsdS.1* (PAS). The *pncB2* epistatic effect with *pncA* appeared to be specific to lineage 4 (Supplementary Table 6). The other nicotinamide cofactor, *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 an effect of *pyrG* (lineage 4, $P = 0.00016$)[25] and

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

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

This table shows loci (protein- and RNA-coding regions, intergenic regions) associated with MDR-TB and XDR-TB ($P<1\times10^{-5}$). The PhyC test additionally detected the *folC*, *pncA–Rv2044c* and *whiB6–Rv3863* loci when comparing MDR-TB against the susceptible group; *eis–Rv2417c*, *gyrB*, *rrs*, *folC*, *alr*, *gid* and the *thyX–hsdS.1* intergenic region when comparing XDR-TB against MDR-TB; and the *alr*, *gyrB*, *pyrG*, *rpoA* and *thyX–hsdS.1* loci when comparing XDR-TB against the susceptible group. Similarly, GWAS using SNPs additionally identified *embC–embA* for MDR-TB vs. the susceptible group (1 SNP), *rrs* and *ubiA* for XDR-TB vs. MDR-TB (each 1 SNP), and the *ubiA* gene for XDR-TB vs. the susceptible group (2 SNPs). [a]The number of nonsynonymous SNPs in the genes. [b]The number of small indels in the genes; those resulting in frameshifts are shown in parentheses. [c]The number of SNPs identified by GWAS. [d]The number of homoplastic SNPs identified using the PhyC test.

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 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 it identified new ones that warrant functional validation in future studies.

GWAS and PhyC tests for individual drugs. As resistance-conferring 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.

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 homoplastic SNP analysis. As previously reported, nonsynonymous 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 INH 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 INH[30], and the *Rv1482c–fabG1* intergenic region harboring 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 INH-resistant strains. For ETH, the *ethA* locus, encoding the

Table 2 | Individual-drug gene-based associations in the complete dataset

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

This table shows loci (protein- and RNA-coding regions, intergenic regions) associated with resistance to individual drugs ($P<1\times10^{-5}$). The GWAS additionally detected a significant association of a SNP (p.Cys213Arg) in the *Rv2688c* locus (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*) and rifampicin (*rpoA*). PAS, *para*-aminosalicylic acid. [a]The number of nonsynonymous SNPs in the genes. [b]The number of small indels in the genes; those resulting in frameshifts are shown in parentheses. [c]The number of SNPs identified by GWAS. [d]The number of homoplastic SNPs identified using the PhyC test.

drug-metabolizing enzyme, was found to be associated with resistance as described previously[31]. A total of 153 nonsynonymous 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 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* locus 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, also demonstrating a role in clinical samples across all four lineages. Two novel loci were identified: *Rv2820c*, thought to enhance mycobacterial virulence ex vivo and in vivo, and *Rv3300c*, encoding a conserved protein of 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 that 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]; accordingly, we observed high odds ratios indicative of high penetrance in association signals in this locus (Fig. 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* versus *rrs* or *gid* Wilcoxon $P=0.03$; *rpsL* versus *gid* Wilcoxon $P=0.04$).

Fluoroquinolones and second-line injectables. The gene- and SNP-based GWAS analyses identified the *gyrA* locus, which encodes

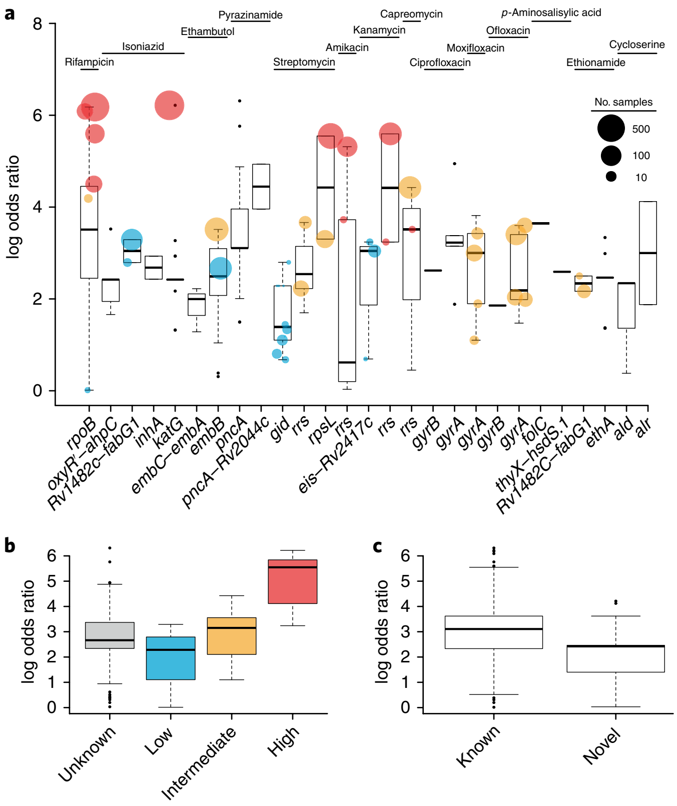

Fig. 3 | Log odds ratios from SNP–drug resistance associations are a potential surrogate for resistance level. a, For each drug, box plots for the log odds ratios ($P<1\times10^{-5}$) for each gene are arranged by increasing median value (as represented by the horizontal line in each box) to show relative effects on resistance. Mutations known to confer low, intermediate or high levels of resistance (Methods) are represented by points colored blue, yellow and red, respectively, and point size is proportional to mutation frequency. Overall, higher levels of resistance are reflected by higher odds ratios; one exception is for *rrs* and CAP, where the c.1484G>C or c.1484G>T mutation (high-level resistance) has a lower odds ratio than c.1401A>G (intermediate level) owing to its low frequency. A similar effect is seen for the same c.1484G>C and c.1484G>T mutations in KAN resistance. **b**, The distribution of log odds ratios ($P<1\times10^{-5}$) for mutations within unknown ($n=167$) or known loci conferring 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-conferring mutations ($P<1\times10^{-5}$). All box plots consist of boxes (median and interquartile range) and whiskers that extend to the most extreme data points that were no more than 1.5 times the interquartile range from the box.

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* mutation encoding p.Cys213Arg was associated with MOX and FLQ resistance, but did not reach statistical significance for OFL. The antibiotic transport ATP-binding protein encoded by *Rv2688c* is a known FLQ efflux pump[39]. As expected, the strongest gene- and SNP-based association signals for resistance across AMK, KAN and CAP were 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 *rrs* mutations, potentially providing evidence that *rrs* mutations confer higher levels of KAN resistance[40], this was not statistically significant owing to small sample size (differences in odds ratios Wilcoxon $P=0.24$) (Fig. 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 harbored multiple homoplastic mutations, including those encoding p.Phe4Leu, p.Lys113Arg and p.Met343Thr. In a previous study, the mutation encoding p.Met343Thr was detected in an XDR-TB strain that had been exposed to CYS treatment, was predicted to alter the protein structure of Alr and was therefore hypothesized to be involved in CYS resistance[42]. To further understand the functional impact of the mutations found in *alr*, we modeled the effect of these variants using the available crystal structure for the protein (PDB 1XFC; Supplementary Fig. 3). Variants 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 p.Met343Thr substitution (found in 12 susceptible and 2 resistant isolates) was predicted to have a more drastic effect on protein structure than p.Lys113Arg, corresponding to the most frequent mutation among CYS-resistant isolates (in 7 susceptible and 23 resistant isolates). There appears to be a balance between the fitness costs associated with mutations and mutation frequency (Supplementary Table 7). The mutation encoding p.Met343Thr appeared independently throughout the phylogenetic tree but did not reach statistical significance for association with 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 prodrug that is converted into its active form by ThyA—a thymidylate synthase, which is encoded by a gene essential for Mtb survival. The candidate drug resistance loci are 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*) were found to be associated with PAS drug resistance in both gene- and SNP-based GWAS analyses. Notably, it has been shown that the c.–16G>A SNP found in our study increases *thyX* expression by 18-fold relative to the wild-type promoter, although no link with PAS resistance was made[18]. Of the three PAS-resistant strains with the c.–16G>A mutation in the *thyX* promoter, two also had a *thyA* mutation (p.Pro145Leu and p.His207Arg), further supporting the idea that upregulation of *thyX* is involved in resistance to PAS[26] or has a compensatory role. The c.–16G>A *thyX* mutation is a homoplastic mutation and is 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 (Fig. 3b; log odds ratios: linear regression trend $P=1.5\times10^{-9}$, high versus intermediate $P=5.2\times10^{-5}$; intermediate versus low $P=5.8\times10^{-10}$). This analysis demonstrates the potential utility of odds ratios and their statistical significance as an indicator of 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 $P=8.3\times10^{-5}$), reflecting the ability of the GWAS to discover effect sizes of lower magnitude (Fig. 3c). A pathway analysis comparing MDR-TB and XDR-TB strains to susceptible strains identified only one significant annotation cluster with 17.7-fold enrichment for antibiotic resistance and response to antibiotics ($P=1.6\times10^{-7}$), 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 and Supplementary Data 1). The candidate genes differed in their abundance of small indels, reflecting their essentiality for survival: drug targets had a lower density of indels, whereas drug-metabolizing 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 1 bp in length and caused frameshifts, the indels in *rpoB* inserted or deleted whole codons; that is, 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 resistance and XDR-TB. 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 center) and the codon 427–434 region in *rpoB*.

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

Effects on resistance prediction using GWAS variants. We sought to establish whether any of the mutations found in association and homoplastic analyses increased the predictability of resistance phenotypes for individual drugs (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% for 8 of 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 nine drugs while maintaining specificities of at least 90%, except for ETH, which was 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 thus far, 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. While resistant phenotypes may be imputed from established resistance-causing mutations, 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-conferring loci and compensatory relationships, thereby confirming the authenticity and robustness of the approach. It also identified several recently discovered loci (*folC*, *ubiA*, *thyX–hsdS.1*, *thyA*, *alr*, *ald* and *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 prodrug 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 in comparison to those for 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 sensitivity to second-line drugs. However, because of the large effect that causal mutations have on drug resistance phenotypes, 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 owing to the paucity of phenotypic susceptibility data.

The analysis highlighted the importance of indels to drug resistance, particularly their high density in drug-metabolizing genes, in

Table 3 | Impact on drug resistance prediction (percentage) 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-conferring SNPs and indels (TBDR; http://tbdr.lshtm.ac.uk/)[9,31] when predicting phenotypic drug resistance (TBDR panel columns). The SNPs in TBDR contribute 100% to the stated sensitivity, except for 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 bold. MDR-TB, multidrug-resistant TB; PAS, *para*-aminosalicylic acid; Sens., sensitivity; Spec., specificity; XDR-TB, extensively drug-resistant TB.

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 epigenetics and gene expression[44].

Mtb strains are usually classified as drug resistant or susceptible on the basis of 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 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 identified 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, as 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 five continents and sampled multiple geographic regions, allowing us to observe differences in the prevalence of drug-resistance-conferring mutations and mechanisms. Some drug resistance and compensatory/epistatic relationships were found to vary across geographic populations and bacterial lineages, implying that regional variation should be considered to fully characterize genotype–phenotype relationships. The differential lineage effects could impact on relative virulence between strain types. Enhanced understanding of the genetic basis of phenotypic antituberculosis 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 UK, it remains to be trialed in resource-poor settings that are representative of most patients with TB worldwide.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-017-0029-0.

Received: 3 February 2017; Accepted: 1 December 2017;
Published online: 22 January 2018

References

1. Dheda, K. et al. Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *Lancet Respir. Med.* **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. *Clin. Infect. Dis.* **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. *Nat. Commun.* **5**, 4812 (2014).
6. Black, P. A. et al. Energy metabolism and drug efflux in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **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. *Antimicrob. Agents Chemother.* **57**, 827–832 (2013).
8. Coll, F. et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7**, 51 (2015).
9. Farhat, M. R. et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **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. *Nat. Genet.* **45**, 1255–1260 (2013).
11. Desjardins, C. A. et al. Genomic and functional analyses of *Mycobacterium tuberculosis* strains implicate *ald* in D-cycloserine resistance. *Nat. Genet.* **48**, 544–551 (2016).
12. Earle, S. G. et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 16041 (2016).
13. Chewapreecha, C. et al. Comprehensive identification of single nucleotide polymorphisms associated with β-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.* **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. *Nat. Genet.* **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 Med.* **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 Med.* **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*. *Antimicrob. Agents Chemother.* **40**, 2452–2454 (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. *Antimicrob. Agents Chemother.* **59**, 1320–1324 (2015).
20. Bisson, G. P. et al. Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, *rpoB* mutant *Mycobacterium tuberculosis*. *J. Bacteriol.* **194**, 6441–6452 (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. *Nat. Genet.* **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. *Nat. Genet.* **45**, 1190–1197 (2013).
25. Mori, G. et al. Thiophenecarboxamide derivatives activated by EthA kill *Mycobacterium tuberculosis* by inhibiting the CTP synthetase PyrG. *Chem. Biol.* **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. *Nat. Genet.* **46**, 279–286 (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*. *Mol. Microbiol.* **46**, 453–466 (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. III Ethionamide activation and sensitivity in multidrug-resistant *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **97**, 9677–9682 (2000).
32. Telenti, A. et al. The *emb* operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol. *Nat. Med.* **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. *Nat. Med.* **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*. *Emerg. Microbes Infect.* **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. *Clin. Microbiol. Infect.* **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. *Antimicrob. Agents Chemother.* **38**, 773–780 (1994).
38. Kocagöz, T. et al. Gyrase mutations in laboratory-selected, fluoroquinolone-resistant mutants of *Mycobacterium tuberculosis* H37Ra. *Antimicrob. Agents Chemother.* **40**, 1768–1774 (1996).
39. Pasca, M. R. et al. Rv2686c–Rv2687c–Rv2688c, an ABC fluoroquinolone efflux pump in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **48**, 3175–3178 (2004).
40. Zaunbrecher, M. A., Sikes, R. D. Jr., Metchock, B., Shinnick, T. M. & Posey, J. E. Overexpression of the chromosomally encoded aminoglycoside acetyltransferase *eis* confers kanamycin resistance in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **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*. *N. Engl. J. Med.* **369**, 290–292 (2013).
43. Schön, T. et al. *Mycobacterium tuberculosis* drug-resistance testing: challenges, recent developments and perspectives. *Clin. Microbiol. Infect.* **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 resistance analysis: a call to action. *Int. J. Infect. Dis.* **56**, 130–135 (2017).

Acknowledgements

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 the 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. Perdigão received a Fundação para a Ciência e a Tecnologia (Portugal) postdoctoral fellowship fund (SFRH/BPD/95406/2013). The Calouste Gulbenkian Foundation, the Institute Gulbenkian in Lisbon and the European Society of Clinical Microbiology and Infectious Diseases supported the research of C.P., J. Perdigão, I.P. and M.V. J. Phelan is funded by a BBSRC PhD studentship. T.G.C. is funded by the Medical Research Council UK (grant numbers MR/K000551/1, MR/M01360X/1, MR/N010469/1 and MC_PC_15103). N.F. is funded by the Medical Research Council UK (grant number MR/K020420/1). 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 N. Mistry (Foundation for Medical Research, Mumbai) for contributing Mtb archived strains and drug sensitivity testing data. We wish to thank G. Moniz at the Laboratorio Central de Saúde Pública for supporting the collection of samples in Brazil and the South African National Health Laboratory Service for their contribution providing access to 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 interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-017-0029-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.M. or A.P. or T.G.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Sequence data and variant calling. Sequence data for 6,465 *M. tuberculosis* complex clinical isolates were generated as part of a collaborative global drug resistance project ($n=2,637$; http://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 (1–4) (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 summarized in Supplementary Fig. 5. The new and archived raw sequence data were aligned to the H37Rv reference genome (Genbank accession NC_000962.3) using the BWA mem algorithm[46] (settings: --c 100 --T 50). SAMtools/BCFtools[47] (default settings) and GATK[48] software were used to call SNPs and small indels. The GATK parameters used were -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 poor-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 localized de novo assembly, and those found in association analysis (*dfrA*/*thyA*, *pncA*, *ethA*/*ethR* and *katG*) were confirmed using PCR.

Phenotypic drug susceptibility testing. Drug susceptibility data were obtained from WHO-recognized testing protocols[51]. The Mtb isolates that provided sequence data included in this study are summarized 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 (for example, India, collected before 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 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 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 recognized testing protocols[51]. Isolates from Malawi were shipped to the UK'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 RIF and INH using the Microscopic Observation Drug Susceptibility (MODS) assay[54] at the Universidad Peruana Cayetano Heredia (UPCH) before transfer to the national reference laboratory for further testing. In Peru, susceptibility to PZA was assessed by the Wayne assay, a colorimetric biochemical test in which PZA is hydrolyzed 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]. PZA sensitivity was determined using BACTEC 7H12 liquid medium, pH 6.0, at 100 μg/ml (BACTEC PZA test medium, Becton Dickinson). When testing on agar, the critical drug concentrations used were as follows: RIF, 1 μg/ml; INH, 0.2 μg/ml; STM, 2 μg/ml; EMB, 5 μg/ml; CIP, 2 μg/ml; AMK, 5 μg/ml; CAP, 10 μg/ml; KAN, 5 μg/ml (Pakistan, 6 μg/ml); ETH, 5 μg/ml; PAS, 2 μg/ml[53]. For Lowenstein–Jensen testing, the drug concentrations used were as follows: STM, 4.0 μg/ml; INH, 0.2 μg/ml; RIF, 40.0 μg/ml; EMB, 2.0 μg/ml; CAP, 40.0 μg/ml; KAN, 30.0 μg/ml (China) or 20.0 μg/ml (Vietnam); OFL, 2.0 μg/ml; ETH, 40 μg/ml; thioacetone, 10 μg/ml; PZA, 200 μg/ml; CYS, 30 μg/ml; PAS, 0.5 μg/ml[55].

Phylogenetic tree and association analysis. The best scoring maximum-likelihood phylogenetic tree rooted on *M. canetti* (Genbank accession 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 were determined using lineage-specific SNPs[5]. Further population structure assessment was performed using principal-components analysis (Supplementary Fig. 2), which clustered samples by genotype congruently with the phylogenetic tree. The principal components were calculated from a SNP pairwise distance matrix between each sample, and the first five components (summarizing 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 (resistant versus susceptible) and the aggregate number of mutations (SNPs, indels or large deletions) by coding regions, 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, sublineage and outbreak-based population structure[49]. 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 main Mtb lineage and sublineage 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 minimize 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 antituberculosis drugs and the stepwise 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 again performed GWAS analysis (10,000 times). We report all findings that are below a calculated permutation threshold of $P<1\times10^{-5}$. All statistical analyses were performed using R software. To identify SNPs enriched for convergent evolution and provide further evolutionary evidence, the phylogenetics-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 result in spurious associations. Mutations truly conferring drug resistance often originate multiple times independently in the phylogeny. Mutations that originated once in the tree (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 loci if they (i) were associated with drug resistance, (ii) harbored 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 gene encoding the drug target or drug-activating enzyme. In the fourth analysis, deep phylogenetic and synonymous SNPs were removed before calculating the number of samples with nonsynonymous SNPs at genes of interest (for example, p.Ala1075Ala at *rpoB* or p.Glu1092Asp at *rpoC*). The significance of differences between studied genes was calculated using Fisher's exact test (cutoff of $P<1\times10^{-8}$).

Protein mutation modeling. Apo crystal structures for Alr were downloaded from the Protein Data Bank (PDB 1XFC[61]) and then subjected to modeling of missing residues, WinCOOT regularization 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 was used to quantify effects on protein–protein interactions and mCSM-Lig was used to quantify effects on drug binding[62–64]. For ligand binding, D-cycloserine was modeled in the active site using UCSF Chimera v1.11[65] from the coordinates of the closest holohomolog *Clostridium difficile* 630 (PDB 4LUT)[66].

Statistical analyses. The statistical mixed models used for association analysis are described above. The terms 'low', 'intermediate' and 'high' levels of resistance referred to in the text and Fig. 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 that 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.

URLs. TubercuList knowledge base, https://mycobrowser.epfl.ch/; Tuberculosis Database, https://mycobrowser.epfl.ch/; R statistical software, https://www.r-project.org/; TB Global Drug Resistance Collaboration, http://pathogenseq.lshtm.ac.uk/#tuberculosis.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. All raw sequencing data are available; 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.

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. *Nat. Genet.* **43**, 491–498 (2011).
49. Phelan, J. et al. *Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **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. *Guidelines for Surveillance of Drug Resistance in Tuberculosis* (World Health Organization, Geneva, 2009).
52. Kubica, G. & Kent, K. *Public Health Mycobacteriology: A Guide for the Level III Laboratory* (Centers for Disease Control, US Department of Health and Human Services, Atlanta, 1985).
53. Canetti, G. et al. Mycobacteria: laboratory methods for testing drug sensitivity and resistance. *Bull. World Health Organ.* **29**, 565–578 (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. *Lancet Infect. Dis.* **10**, 688–698 (2010).
55. Wayne, L. G. Simple pyrazinamidase and urease tests for routine identification of mycobacteria. *Am. Rev. Respir. Dis.* **109**, 147–151 (1974).
56. Pfyffer, G.E., Palicova, F. & Rüsch-Gerdes, S. Testing of susceptibility of *Mycobacterium tuberculosis* to pyrazinamide with the nonradiometric BACTEC MGIT 960 system. *J. Clin. Microbiol.* **40**, 1670–1674 (2002).
57. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).
58. Coll, F. et al. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics* **28**, 2991–2993 (2012).
59. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
60. Alam, M. T. et al. Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. *Genome Biol. Evol.* **6**, 1174–1185 (2014).
61. Velankar, S. et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.* **44** (D1), 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. *Sci. Rep.* **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 Res.* **42**, W314–W319 (2014).
65. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
66. Asojo, O. A. et al. Structural and biochemical analyses of alanine racemase from the multidrug-resistant *Clostridium difficile* strain 630. *Acta Crystallogr. D, Biol. Crystallogr.* **70**, 1922–1933 (2014).
67. Wong, S. Y. et al. Mutations in *gidB* confer low-level streptomycin resistance in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **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. *Antimicrob. Agents Chemother.* **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. *Int. J. Mycobacteriol.* **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).

nature research

Corresponding author(s): Taane G. Clark

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Maximum number of samples possible.

2. Data exclusions

Describe any data exclusions.

Samples with low quality sequencing data were excluded.

3. Replication

Describe whether the experimental findings were reliably reproduced.

N/A

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

N/A

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

N/A

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	□	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

BWA mem, SAMtools/BCFtools, GATK, Delly2, RAxML, SpolPred, phyC, GEMMA, mCSM, DUET, R

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

N/A