A data compendium of *Mycobacterium tuberculosis* antibiotic resistance

Alice Brankin[4,*,**] and Kerri M Malone[19,*,**], Ivan Barilar[29], Simone Battaglia[1], Emanuele Borroni[1], Angela Pires Brandao[2,3], Andrea Maurizio Cabibbe[1], Joshua Carter[5], Daniela Maria Cirillo[1], Pauline Claxton[6], David A Clifton[4], Ted Cohen[7], Jorge Coronel[8], Derrick W Crook[4], Viola Dreyer[29], Sarah G Earle[4], Vincent Escuyer[9], Lucilaine Ferrazoli[3], Philip W Fowler[4], George Fu Gao[10], Jennifer Gardy[11], Saheer Gharbia[12], Kelen Teixeira Ghisi[3], Arash Ghodousi[1,13], Ana Luíza Gibertoni Cruz[4], Louis Grandjean[33], Clara Grazian[14], Ramona Groenheit[44], Jennifer L Guthrie[15,16], Wencong He[10], Harald Hoffmann[17,18], Sarah J Hoosdally[4], Martin Hunt[19,4], Zamin Iqbal[19], Nazir Ahmed Ismail[20], Lisa Jarrett[21], Lavania Joseph[20], Ruwen Jou[22], Priti Kambli[23], Rukhsar Khot[23], Jeff Knaggs[19,4], Anastasia Koch[24], Donna Kohlerschmidt[9], Samaneh Kouchaki[4,25], Alexander S Lachapelle[4], Ajit Lalvani[26], Simon Grandjean Lapierre[27], Ian F Laurenson[6], Brice Letcher[19], Wan-Hsuan Lin[22], Chunfa Liu[10], Dongxin Liu[10], Ayan Mandal[28], Mikael Mansjö[44], Daniela Matias[21], Graeme Meintjes[24], Flávia de Freitas Mendes[3], Matthias Merker[29], Marina Mihalic[18], James Millard[30], Paolo Miotto[1], Nerges Mistry[28], David Moore[31,8], Kimberlee A Musser[9], Dumisani Ngcamu[20], Hoang Ngoc Nhung[32], Stefan Niemann[29, 48], Kayzad Soli Nilgiriwala[28], Camus Nimmo[33], Nana Okozi[20], Rosangela Siqueira Oliveira[3], Shaheed Vally Omar[20], Nicholas Paton[34], Timothy EA Peto[4], Juliana Maira Watanabe Pinhata[3], Sara Plesnik[18], Zully M Puyen[35], Marie Sylvianne Rabodoarivelo[36], Niaina Rakotosamimanana[36], Paola MV Rancoita[13], Priti Rathod[21], Esther Robinson[21], Gillian Rodger[4], Camilla Rodrigues[23], Timothy C Rodwell[37,38], Aysha Roohi[4], David Santos-Lazaro[35], Sanchi Shah[28], Thomas Andreas Kohl[29], Grace Smith[21,12], Walter Solano[8], Andrea Spitaleri[1,13], Philip Supply[39], Utkarsha Surve[23], Sabira Tahseen[40], Nguyen Thuy Thuong Thuong[32], Guy Thwaites[32,4], Katharina Todt[18], Alberto

Trovato[1], Christian Utpatel[29], Annelies Van Rie[41], Srinivasan Vijay[42], Timothy M Walker[4,32], A Sarah Walker[4], Robin Warren[43], Jim Werngren[44], Maria Wijkander[44], Robert J Wilkinson[45,46,26], Daniel J Wilson[4], Penelope Wintringer[19], Yu-Xin Xiao[22], Yang Yang[4], Zhao Yanlin[10], Shen-Yuan Yao[20], Baoli Zhu[47]

Institutions

[1] IRCCS San Raffaele Scientific Institute, Milan, Italy

[2] Oswaldo Cruz Foundation, Rio de Janeiro, Brazil

[3] Institute Adolfo Lutz, São Paulo, Brazil

[4] University of Oxford, Oxford, UK

[5] Stanford University School of Medicine, Stanford, USA

[6] Scottish Mycobacteria Reference Laboratory, Edinburgh, UK

[7] Yale School of Public Health, Yale, USA

[8] Universidad Peruana Cayetano Heredia, Lima, Perú

[9] Wadsworth Center, New York State Department of Health, Albany, USA

[10] Chinese Center for Disease Control and Prevention, Beijing, China

[11] Bill & Melinda Gates Foundation, Seattle, USA

[12] UK Health Security Agency, London, UK

[13] Vita-Salute San Raffaele University, Milan, Italy

[14] University of New South Wales, Sydney, Australia

[15] The University of British Columbia, Vancouver, Canada

[16] Public Health Ontario, Toronto, Canada

[17] SYNLAB Gauting, Munich, Germany

[18] Institute of Microbiology and Laboratory Medicine, IMLred, WHO-SRL Gauting, Germany

[19] EMBL-EBI, Hinxton, UK

[20] National Institute for Communicable Diseases, Johannesburg, South Africa

[21] Public Health England, Birmingham, UK

[22] Taiwan Centers for Disease Control, Taipei, Taiwan

[23] Hinduja Hospital, Mumbai, India

[24] University of Cape Town, Cape Town, South Africa

[25] University of Surrey, Guildford, UK

[26] Imperial College, London, UK

[27] Université de Montréal, Canada

[28] The Foundation for Medical Research, Mumbai, India

[29] Research Center Borstel, Borstel, Germany

[30] Africa Health Research Institute, Durban, South Africa

[31] London School of Hygiene and Tropical Medicine, London, UK

[32] Oxford University Clinical Research Unit, Ho Chi Minh City, Viet Nam

[33] University College London, London, UK

[34] National University of Singapore, Singapore

[35] Instituto Nacional de Salud, Lima, Perú

[36] Institut Pasteur de Madagascar, Antananarivo, Madagascar

[37] FIND, Geneva, Switzerland

[38] University of California, San Diego, USA

[39] Institut Pasteur de Lille, Lille, France

[40] National TB Reference Laboratory, National TB Control Program, Islamabad, Pakistan

[41] University of Antwerp, Antwerp, Belgium

[42] University of Edinburgh, Edinburgh, UK

[43] Stellenbosch University, Cape Town, South Africa

[44] Public Health Agency of Sweden, Solna, Sweden

[45] Wellcome Centre for Infectious Diseases Research in Africa, Cape Town, South Africa

[46] Francis Crick Institute, London, UK

[47] Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

[48] German Center for Infection Research (DZIF), Hamburg-Lübeck-Borstel-Riems, Germany

[*]equal contribution authors

[**]co-corresponding authors

Abstract

The Comprehensive Resistance Prediction for Tuberculosis: an International Consortium (CRyPTIC) presents here a global collection of 15,211 *Mycobacterium tuberculosis* clinical isolates, all of which have undergone whole genome sequencing and have had their minimum inhibitory concentrations to 13 antitubercular drugs measured. The isolates represent five major *M. tuberculosis* lineages originating from 23 countries across four continents. 6,814 isolates were found resistant to at least one drug, including 2,129 samples fully satisfy the clinical definitions of RR/MDR, pre-XDR or XDR. Resistance status to eight antitubercular drugs can be accurately predicted using a genetic mutation catalogue for over 90% of the isolates. Furthermore, we show the presence of suspected resistance conferring mutations for isolates resistant to the newly introduced drugs bedaquiline, clofazimine, delamanid and linezolid. Finally, a case study of rifampicin mono-resistance is presented to showcase how this compendium could be used to advance our genetic understanding of rare resistance

phenotypes and evaluate the likely performance of a widely used molecular diagnostic tool. It is hoped that this compendium, the largest *M. tuberculosis* matched phenotypic and genotypic dataset to date, will facilitate and inspire new research projects for years to come.

Introduction

Tuberculosis (TB) is a curable and preventable disease; 85% of those afflicted can be successfully treated with a six-month regimen. Despite this, TB is the world's top infectious disease killer with 10 million new cases and 1.2 million deaths estimated in 2019 alone [1]. Furthermore, drug resistant TB (DR-TB) is a continual threat; almost half a million people were estimated to have developed resistance to the first-line drug rifampicin (RR-TB), with three quarters of these cases estimated to be multidrug-resistant (MDR-TB, resistant to first-line drugs isoniazid and rifampicin) [1]. Worryingly, only 44% of DR-TB cases were officially notified and just over half of these cases were successfully treated (57%) [1].

To address these issues, the World Health Organisation (WHO) is encouraging the development of better, faster and more targeted diagnostic and treatment strategies through its EndTB campaign. Of particular interest is universal drug susceptibility testing (DST). Conventionally, DST relies on lengthy (4 weeks minimum) culture-based methods that require strict biosafety conditions for *Mycobacterium tuberculosis.* The development of rapid genetics-based assays has decreased diagnostic time to as little as 2 hours through the detection of specific resistance conferring mutations *e.g.* the Cepheid Xpert® MTB/RIF test [2, 3]. However, assay bias towards specific genic regions can result in misdiagnosis of resistance, the

123 prescription of ineffective treatment regimens and subsequent spread of multi-drug
124 resistant disease, as seen during an outbreak in Eswatini of an MDR strain harbouring
125 an *rpoB* I491F mutation not detected by the Xpert® MTB/RIF assay [4-6].
126
127 Furthermore, detection of rifampicin resistance with the Xpert® MTB/RIF line
128 probe assay (LPA) assay is used to infer MDR-TB as epidemiologically, rifampicin
129 resistance tends to coincide with resistance to isoniazid [7]. While this *modus operandi*
130 is successful at pragmatically identifying potential MDR cases quickly and effectively,
131 it is not generally true that a single path towards developing MDR or extensively drug
132 resistant TB (XDR = MDR/RR + resistance to at least one fluoroquinolone and either
133 bedaquiline or linezolid) exists. For example, *M. tuberculosis* strains that are sensitive
134 to rifampicin but resistant to multiple other anti-tubercular drugs would be missed by
135 current LPA diagnostic testing and remain undetected in the community. A better
136 understanding of how various combinations of drug resistance in different
137 backgrounds (nonMDR/MDR) arises is essential to not only prevent the spread of
138 resistant strains and to promote the use of effective treatment regimens, but also to
139 facilitate the development of suitable diagnostic assays unbiased in their approaches
140 to detect *any* and *all* resistance where it occurs.
141
142 Whole-genome sequencing (WGS) has the potential to reveal the entirety of
143 the *M. tuberculosis* genetic resistance landscape for any number of drugs
144 simultaneously whilst enabling a more rapid turnaround time and affordable cost in
145 comparison to DST culture methods [8]. The success of WGS as a diagnostic method
146 wholly depends on a comprehensive and accurate catalogue of resistance-conferring
147 mutations for each drug. Recent advances have shown that genotypic predictions of

148 resistance correlate well with DST results for first-line drugs [7]. However, the
149 mechanisms of resistance to second-line, new and re-purposed drugs (NRDs) are less
150 well understood despite their increased administration in clinics as MDR cases climb
151 [9]. To align with the goals of the EndTB campaign, further research is vital to construct
152 an extensive catalogue of genetic variants associated with resistance, and indeed
153 susceptibility, if WGS-based diagnostics are to become routine practice for DR-TB
154 management.
155
156 The Comprehensive Resistance Prediction for Tuberculosis: an International
157 Consortium (CRyPTIC) has collected *M. tuberculosis* clinical isolates worldwide to
158 survey the genetic variation associated with resistance to 13 antitubercular drugs,
159 specifically the first-line drugs rifampicin, isoniazid, ethambutol, the second-line drugs
160 amikacin, kanamycin, rifabutin, levofloxacin, moxifloxacin, ethionamide, and the new
161 and re-purposed drugs bedaquiline, clofazimine, delamanid and linezolid. Here, we
162 present a data compendium of 15,211 global clinical isolates, each of which has had
163 its genomic sequence determined and DST profile measured [10]. This compendium
164 is the largest drug screening effort to date for *M. tuberculosis* in a 'one isolate – one
165 microscale assay' format across defined compound concentration ranges. Finally, this
166 matched genotypic-phenotypic dataset will prove an invaluable resource to accelerate
167 and improve AMR diagnostic development for TB, both by the enrichment of mutation
168 catalogues for WGS resistance prediction and the identification of important diagnostic
169 gaps and drug resistance patterns.
170

171 ## Methods

172 ### *Ethics*

173 Approval for CRyPTIC study was obtained by Taiwan Centers for Disease Control IRB
174 No. 106209, University of KwaZulu Natal Biomedical Research Ethics Committee
175 (UKZN BREC) (reference BE022/13) and University of Liverpool Central University
176 Research Ethics Committees (reference 2286), Institutional Research Ethics
177 Committee (IREC) of The Foundation for Medical Research, Mumbai (Ref nos.
178 FMR/IEC/TB/01a/2015 and FMR/IEC/TB/01b/2015), Institutional Review Board of
179 P.D. Hinduja Hospital and Medical Research Centre, Mumbai (Ref no. 915-15-CR
180 [MRC]), scientific committee of the Adolfo Lutz Institute (CTC-IAL 47-J / 2017) and in
181 the Ethics Committee (CAAE: 81452517.1.0000.0059) and Ethics Committee review
182 by Universidad Peruana Cayetano Heredia (Lima, Peru) and LSHTM (London, UK).
183
184 ***Sample collection***
185 Participating collection centres varied in their isolate collection approaches, but
186 the consortium collectively aimed to oversample for *M. tuberculosis* isolates with drug
187 resistance and multi-drug resistance.
188
189 ***Plate assay***
190 The CRyPTIC consortium designed two variants of the Sensititre MYCOTB
191 plate (Thermo Fisher Scientific Inc., USA) named the “UKMYC5” and “UKMYC6”
192 microtitre plates [10, 11]. These plates contain five to ten doubling dilutions of 13
193 antibiotics (rifampicin (RIF), rifabutin (RFB), isoniazid (INH), ethambutol (EMB),
194 levofloxacin (LEV), moxifloxacin (MXF), amikacin (AMI), kanamycin (KAN),
195 ethionamide (ETH), clofazimine (CFZ), linezolid (LZD), delamanid (DLM), and
196 bedaquiline (BDQ)). Delamanid and bedaquiline were provided by Otsuka
197 Pharmaceutical Co., Ltd. and Janssen Pharmaceutica respectively. The UKMYC5

198 plate also contained para-aminosalicylic acid (PAS), however the MICs were not
199 reproducible and hence it was excluded from the UKMYC6 plate design and is not
200 included in any subsequent analysis [10].

201 A standard operating protocol for sample processing was defined by CRyPTIC
202 as previously described [10, 11]. Clinical samples were sub-cultured using 7H10 agar
203 plates, Lowenstein-Jensen tubes or MGIT tubes. Bacterial cell suspensions (0.5
204 McFarland standard, saline Tween) prepared from (no later than) 14-day old colonies
205 were diluted 100X in 10 ml enriched 7H9 broth prior to plate inoculation. A semi-
206 automated Sensititre Autoinoculator (Thermo Fisher, Scientific Inc., USA) was used to
207 inoculate 100 µl prepared cell suspensions (1.5 x 10^5 CFU/ml [5 x 10^4 CFU/ml - 5 x
208 10^5 CFU/ml]) into each well of a UKMYC5/6 microdilution plate. The plate was sealed
209 and incubated for 14 days at 37°C. Quality control runs were performed periodically
210 using *M. tuberculosis* H37Rv ATCC 27294, which is sensitive to all drugs on the plates.
211

212 ***Minimum Inhibitory Concentration (MIC) measurement***

213 As stated in reference [10], minimum inhibitory concentrations (MICs) for each
214 drug were read after incubation for 14 days by a laboratory scientist using a Thermo
215 Fisher Sensititre™ Vizion™ digital MIC viewing system. The Vizion apparatus was
216 also used to take a high contrast photograph of the plate with a white background,
217 from which the MIC was measured again using the Automated Mycobacterial Growth
218 Detection Algorithm (AMyGDA) software [12]. The AMyGDA software detects the
219 boundaries of each well using a Hough transform for circles and measures growth as
220 the number of dark pixels within the area contained by this boundary. All images where
221 the MICs measured by Vizion and AMyGDA were different were uploaded to a citizen

222 science project, BashTheBug, on the Zooniverse platform [13]. Each image was then
223 classified by ≥11 volunteers and the median classification taken.
224 MICs were then classified as high (at least two methods concur on the MIC),
225 medium (either there is no plate image, or Vizion and AMyGDA disagree and there is
226 no BashTheBug measurement), or low (all three methods disagree) quality. When
227 considering the pattern of resistance in a sample we took the MIC from the Vizion
228 reading, if it was annotated as having medium or low quality, to ensure adequate data
229 coverage.
230
231 ***Genomic data processing and variant calling***
232 Sequencing arrangements differed slightly between each CRyPTIC
233 participating laboratory. Short-read, paired end libraries were sequenced on Illumina
234 machines and the resulting FASTQ files were processed using the bespoke pipeline
235 Clockwork (v0.8.3, https://github.com/iqbal-lab-org/clockwork). Briefly, all raw
236 sequencing files are indexed into a relational database with which Clockwork
237 proceeds. Human, nasopharyngeal flora and human immunodeficiency virus related
238 reads are removed and remaining reads are trimmed (adapters and low quality ends)
239 using Trimmomatic and mapped with BWA-MEM to the *M. tuberculosis* H37Rv
240 reference genome (NC000962.3) [14, 15]. Read duplicates are removed. Genetic
241 variants are called independently using Cortex and SAMtools, two variant callers with
242 orthogonal strengths (SAMtools a high sensitivity SNP caller, and Cortex a high
243 specificity SNP and indel caller) [16, 17]. The two call sets are merged to produce a
244 final call set, using the Minos adjudication tool to resolving loci where the two callers
245 disagreed, by remapping reads to an augmented genome containing each alternative
246 allele [18]. Default filters of a minimum depth of 5x and a Fraction of Supporting Reads

247 of 0.9 (minos) and a Genotype confidence percentile (GCP) filter of 0.5 were applied.
248 The GCP filter is, roughly speaking, a normalised likelihood ratio test, giving a measure
249 of confidence in the called allele compared with the other alternatives, and is described
250 in [18]. This produced one VCF file per sample, each only describing positions where
251 that sample differed from the reference.
252
253 These filtered VCFs were then combined, to produce a single non-redundant
254 list of all variants seen in the cohort. All samples were then processed a second time
255 with Minos, remapping reads to a graphical representation of all the segregating
256 variation within the cohort, generating VCF files which had an entry at all variable
257 positions (thus for all samples, most positions would be genotyped as having the
258 reference allele).
259
260 In order to remove untrustworthy loci, a genome mask was applied to the
261 resulting VCF files (regions identified with self-blast matches in [19] comprising of
262 324,971 bp of the reference genome. Furthermore, positions with less than 90% of
263 total samples passing default Clockwork/Minos variant call filters (described above)
264 were filtered out, comprising 95,703 bp of the genome, of which 55,980 bp intersect
265 with the genome mask.
266
267 ***Phylogenetic tree construction***
268 A pairwise genetic distance matrix was constructed for 15,211 isolates by
269 comparing each two variant call format (VCF) files (relative to the reference *M.*
270 *tuberculosis* H37Rv genome, see [11] and the above methods section for more details)
271 under consideration position-wise. A neighbourhood-joining tree was constructed from

272 the distance matrix using *quicktree* [20]. Tree visualisation and annotation was
273 performed using the R library *ggtree* [21]. *M. tuberculosis* lineages were assigned
274 using Mykrobe and are represented by the coloured dots at the branch termini of the
275 tree [18]. For isolates that had 'mixed' lineage classification (*i.e.,* 2 lineages were found
276 present in the sample by Mykrobe, *n* = 225, 1.5%), the first of the two lineages was
277 assigned to the isolate. *ggtree* was also used to construct the trees depicting
278 bedaquiline, clofazimine and delamanid resistant isolates.
279
280 ***Data analysis and visualisation***
281 All data can be found in <ftp address: tbc> and were analysed and visualised
282 using either R or python3 libraries and packages. See <github: tbc> for codebase.
283

284 # Results

285 ***15,211 M. tuberculosis clinical isolates***
286 The CRyPTIC dataset comprises isolates from 27 countries collected by 14
287 partner sites worldwide. The compendium presented here contains 15,211 isolates for
288 which both genomic and phenotypic data was collected by 23 of the partner countries
289 across the continents of Asia, Africa, South America and Europe (Fig. 1). Where the
290 origin of an isolate was not known, the isolate was assumed to originate in the same
291 country as the contributing laboratory (269 isolates in Germany, 17 isolates in India, 6
292 isolates in Peru, 885 isolates in Italy, 510 isolates in South Africa, 357 isolates in
293 Sweden, 208 isolates in Taiwan, 1 isolate in Brazil and 4 isolates in the UK). The
294 largest number of isolates were contributed by India (*n* = 4,004), Peru (*n* = 2,691),
295 South Africa (*n* = 2,155), Vietnam (*n* = 1,288) and China (*n* = 1,121).
296

297 Lineage assignment revealed that all isolates (except 41, 0.3%) belong to the
298 four main *M. tuberculosis* lineages (L1-L4). Like previous studies, we see a strong
299 association between geolocation and lineage (Pearson's chi-squared test, $p < 2.2e-$
300 16, Fig. S3)[22, 23]. The pie-charts in Fig. 1 and supplemental tables show the
301 breakdown of isolate number -v- lineages (Table S2) and sub-lineages (Table S3) for
302 each location. The phylogenetic tree in Fig. 2 further highlights the strong population
303 structure of this collection of isolates, with isolates clustering according to lineage.
304
305 Isolates of the ancient Indo-oceanic lineage/L1 (*n* = 1,150) were mostly
306 contributed by India (*n* = 676 isolates) and Vietnam (*n* = 283 isolates). 85% of the L1
307 Indian isolates belong to sub-lineages 1.1.2 and 1.2.2 while 66% of the Vietnamese
308 isolates are sub-lineage 1.1.1.1. No L1 isolates were contributed by 10 of the 23
309 countries in this study with only 2 isolates collected in South America (one in each of
310 Brazil and Peru).
311
312 There are 5,598 L2 (East Asian) isolates, making it the second largest group in
313 the dataset. L2 was found most prevalent in Asia and Europe with the largest
314 proportion found in amongst isolates contributed by China (*n* = 722, 64% of isolates)
315 and India (*n* = 1,573, 39% of isolates). Sub-lineages 2.2 and 2.2.7 dominate the L2
316 isolates (*n* = 1,421 and 1,249 respectively); 2.2 was found mostly amongst Peruvian
317 and South African isolates (*n* = 231 and 161 respectively) apart from those contributed
318 by the Asian countries of Vietnam (*n* = 271), China (*n* = 284) and India (*n* = 272), while
319 85% of sub-lineage 2.2.7 isolates were contributed by South Africa (*n* = 206), Vietnam
320 (*n* = 164) and India (*n* = 691). 70% of sub-lineage 2.2.1 originated from South Africa
321 (10% of isolates found here) and has recently been associated with more favourable

322 transmission rates [24]. Lastly, 86% and 72% of isolates contributed by Kyrgyzstan
323 and Turkmenistan respectively belong to L2 with sub-lineage 2.2.10 dominating (16/24
324 and 75/86 isolates for both countries respectively). 2.2.10 has been previously
325 described as restricted to Central Asia and this is reflected in the compendium [25].
326
327 The majority (1184/1850, 65%) of L3 (East African/Indian) isolates were
328 contributed by India, followed by 19.6% (363/1850) isolates from Pakistan. L3 is
329 typically under-sampled and under-studied in current databases and biobanks in
330 comparison to L2 and L4; the L3 isolates in this study are the largest collected to date
331 in a single study [26].
332
333 L4 (Euro-American) is the largest lineage group in the compendium (*n* = 6,572).
334 Isolates donated from Peru dominate; 87% of all L4 isolates are Peruvian with 4.1.2.1
335 and 4.3.3 being the most prevalent sub-lineages (24% and 22% respectively). There
336 are 34 different L4 sub-lineages in the dataset, making L4 the most diverse in
337 comparison to the other lineage groupings.
338
339 There were no L5 isolates found in the compendium and only 6 L6 isolates were
340 identified. Animal-restricted pathogenic mycobacterial isolates are also rare in the
341 compendium; only 16 cases were identified (*n* = 15 *M. bovis* and *n* = 1 *M. caprae*).

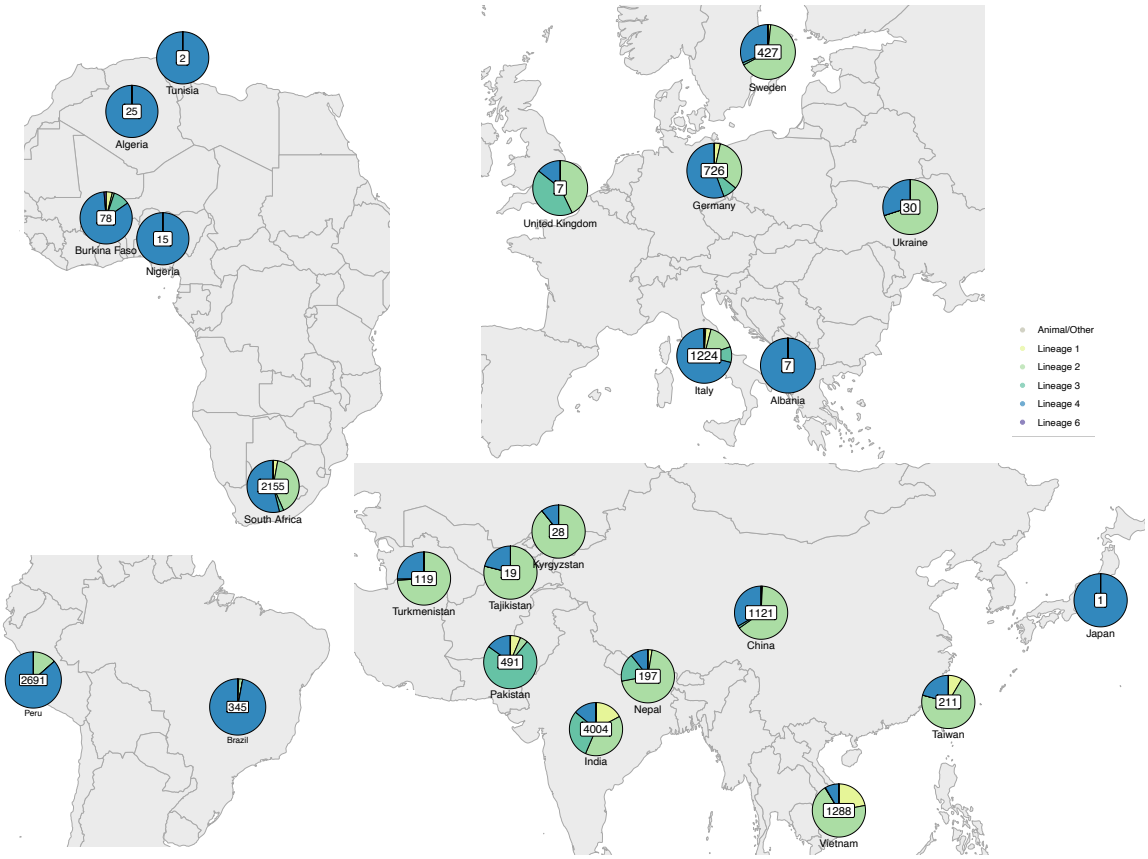


Figure 1: ***Geographical distribution of CRyPTIC M. tuberculosis clinical isolates.*** The total number of isolates contributed by each country is depicted, with pie charts representing the proportion of *M. tuberculosis* lineages. Where the origin of an isolate was not known, the collection site identity was assigned (269 isolates in Germany, 17 isolates in India, 6 isolates in Peru, 885 isolates in Italy, 510 isolates in South Africa, 357 isolates in Sweden, 208 isolates in Taiwan, 1 isolate in Brazil and 4 isolates in the UK).

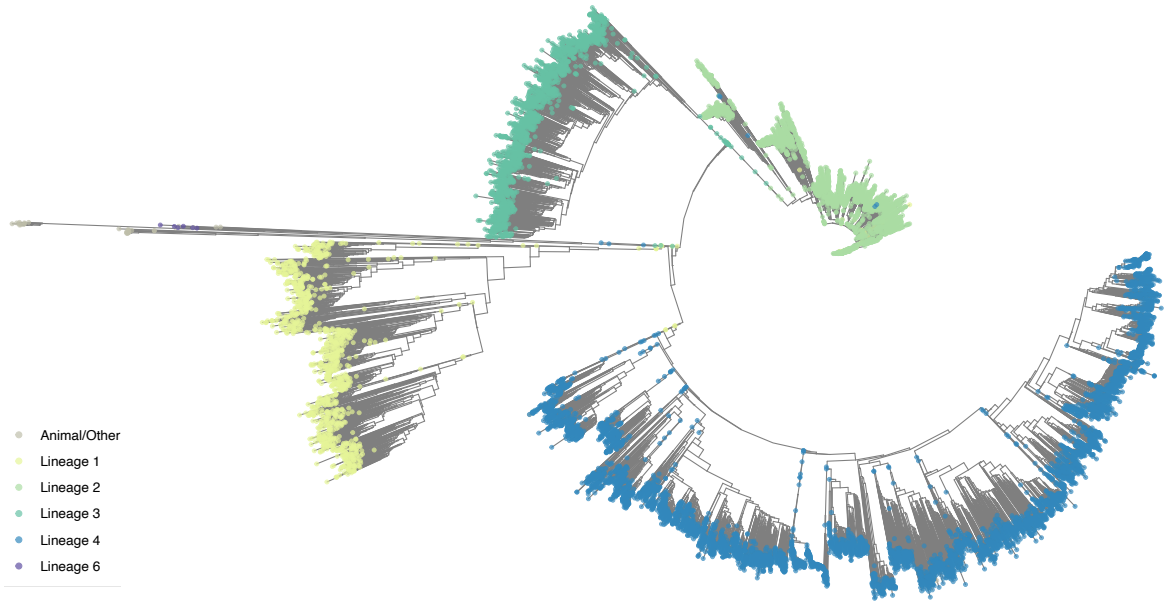


Figure 2: ***Phylogenetic tree of CRyPTIC M. tuberculosis clinical isolates.*** A phylogenetic cladogram of 15,211 *M. tuberculosis* clinical isolates. A neighbour-joining tree was constructed from a pairwise distance matrix using *quicktree* [20]. Coloured dots at the branch termini represent the lineage assigned to each isolate. "Animal/Other" includes 16 isolates that were assigned the following lineages: *M. caprae, M. bovis, unknown* along with 17 quality control isolates representing specific sub-lineages [27].

Quality assurance of the drug minimum inhibitory concentrations

Regular quality assurance checks detected problems with plate inoculation and reading in two laboratories during collection of the 15,211 isolates. In addition, some plates were not able to be read, usually because of inadequate growth in the positive control wells or extensive contamination. These anomalies were investigated, and samples were removed from the dataset where appropriate, resulting in 2,922 isolates being held out of the final dataset. Further, due to skip wells and other phenomena that prevent an MIC being measured, 11.9% of isolates did not return a phenotype for all 13 drugs on the plate. The compendium therefore has a total of 157,401 MIC measurements for 12,289 isolates (Table 1).

	MIC MEASUREMENTS	HIGH QUALITY	MEDIUM QUALITY	LOW QUALITY
INH	12070	9519	1351	1200
RIF	12099	8955	1356	1788
EMB	12158	7506	1355	3297
LEV	12163	7774	1354	3035
MXF	12194	6785	1353	4056
AMI	12072	8973	1350	1749
KAN	12130	9333	1355	1442
BDQ	12068	8536	1355	2177
CFZ	12049	7763	1352	2934
DLM	11927	8095	1349	2483
LZD	12189	7141	1355	3693
ETH	12132	8821	1355	1956
RFB	12150	10042	1352	756
TOTAL	**157401**	**109243**	**17592**	**30566**

Table 1: Q*uality metrics for phenotype data* Stated for each drug is the total number of MIC readings ("PHENOTYPED") stratified into "high" quality (at least two MIC measurement methods agree), "medium" quality (either Vizion and AMyGDA disagree, or there is no plate picture) or "low" quality (all three MIC measurements methods disagree) phenotype classifications as described in Methods. Drug acronyms: INH = isoniazid, RIF = rifampicin, EMB = ethambutol, LEV = levofloxacin, MXF = moxifloxacin, AMI = amikacin, KAN = kanamycin, BDQ = bedaquiline, CFZ = clofazimine, DLM = delamanid, LZD = linezolid, ETH = ethionamide, RFB = rifabutin.

Resistance classification and distribution

Binary drug resistant/susceptible phenotypes were calculated from the MICs by applying an epidemiological cut-off value (ECOFF); samples with MICs at or below the ECOFF are, by definition, wild-type and hence assumed to be susceptible to the drug in question [11] (Fig. S1). Samples with MICs above the ECOFF are therefore assumed to be resistant.

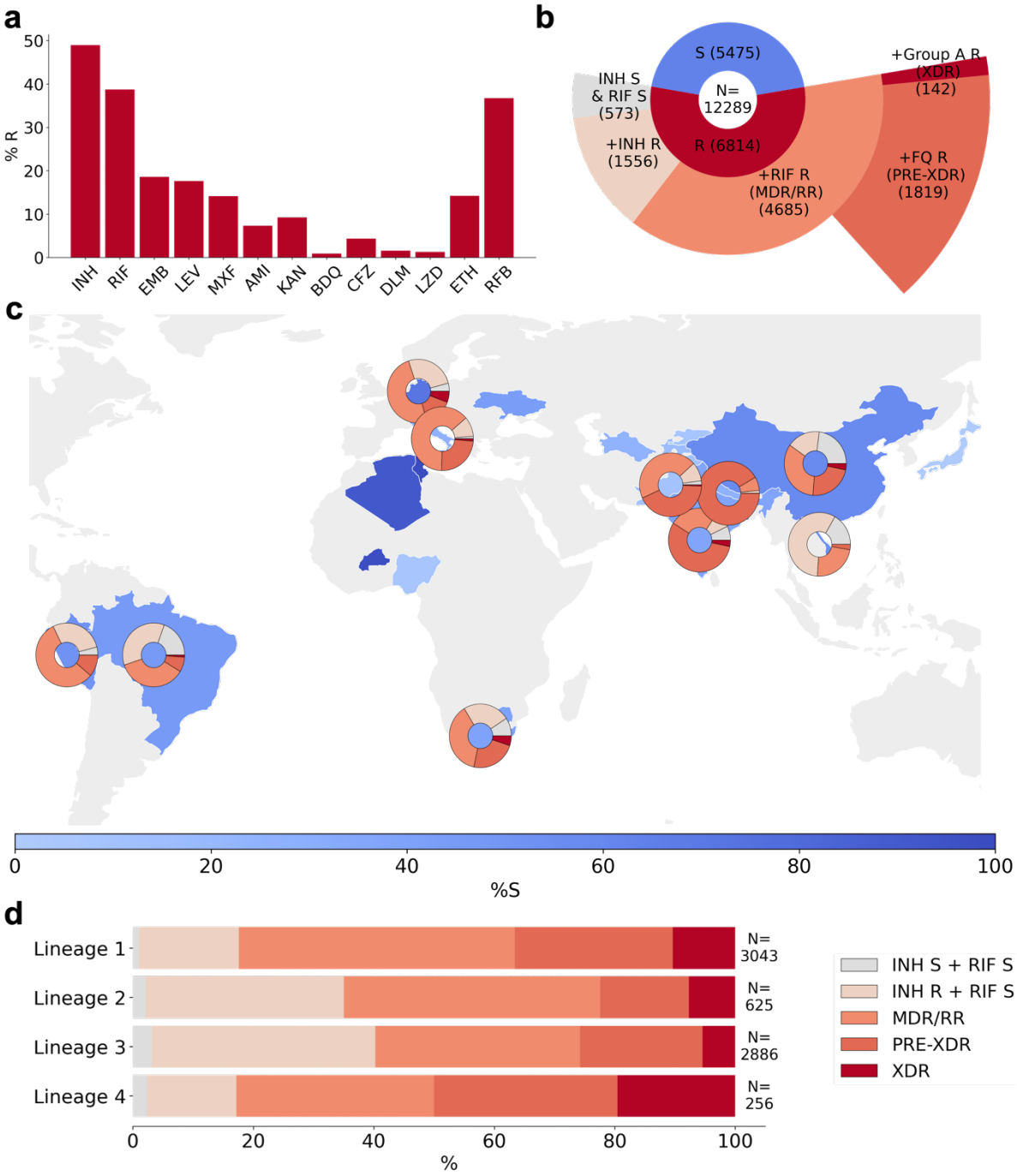
Unsurprisingly given its size, resistance to each of the 13 drugs is represented within the CRyPTIC compendium (Fig. 3a). The drugs with the highest percentage of resistance are the first line drugs isoniazid and rifampicin (49.0% and 38.7% respectively). Of the second line drugs, levofloxacin had the highest proportion of

390 resistant isolates in the dataset (17.6%) and amikacin the lowest (7.3%). Reassuringly,
391 a low proportion of isolates were resistant to the NRDs, bedaquiline (0.9%),
392 clofazimine (4.4%), delamanid (1.6%) and linezolid (1.3%).
393
394 Of the 12,289 isolates which returned a binary phenotype for at least one drug,
395 6,814 (55.4%) were 'resistant' to at least one drug (Fig. 3b). For the purpose of
396 describing the broader resistance categories present in the dataset, we assumed that
397 all MICs that could not be read had susceptible phenotypes in this section.
398 Consequently, the calculated prevalence of MDR, XDR etc in the dataset (Fig. 3) are
399 likely under-estimates.
400
401 At least 68.8% (*n* = 4,685) of drug resistant isolates were either MDR or RR
402 (Fig. 3b) and of these 38.8% (*n* = 1,819) were pre-XDR (MDR/RR + fluoroquinolone
403 resistance) and 3.0% were XDR (*n* = 142). Two of the XDR isolates returned a
404 resistant phenotype to all 13 of the drugs assayed (Table S1) and therefore could be
405 reasonably described as totally drug resistant (TDR). One such isolate belonged to L4
406 and was contributed by South Africa, and the other belonged to L2 with an unknown
407 country of origin.
408
409 Of the 2,129 resistant isolates that did not fit into either of the MDR or RR
410 categories, more were categorised as isoniazid resistant (73.1%, *n* = 1,556) than were
411 categorised as susceptible to both isoniazid and rifampicin and resistant to at least
412 one other drug (26.9%, *n* = 573).
413

414 Whilst many partner laboratories oversampled for resistance, others used a
415 less biased collection protocol (Fig. 3c). Therefore, the proportions of resistant
416 phenotypes are not necessarily representative of the overall distribution of resistance
417 in the country and are not comparable between countries. Isolates with unknown
418 country of origin were excluded from this analysis. In all countries that contributed
419 more than 100 resistant isolates, each of the five phenotypic resistance categories in
420 Fig. 3b were identified, except for Peru, Vietnam and Nepal which did not contribute
421 any XDR isolates. Vietnam and Brazil sampled a particularly high proportion of non-
422 MDR/RR resistant phenotypes; 73.9% and 55.1% of resistant isolates contributed by
423 these countries, respectively, were neither MDR nor RR.
424
425 For Nepal and India, an especially high proportion of the MDR/RR isolates
426 contributed were fluoroquinolone resistant (92.9% (*n* = 132/142 isolates) and 69.8%
427 (*n* = 538/770 isolates) respectively), as has been previously observed for this
428 geographical region [28, 29].
429
430 All five categories of resistance were represented in the four major *M.*
431 *tuberculosis* lineages. We note that these relative proportions will have been
432 influenced by the different sampling approaches at each partner laboratory since
433 lineage distributions are typically geographically distinct (Fig. 1). Bearing this in mind,
434 we observe that in the compendium, L3 isolates contained the most MDR/RR isolates
435 as a proportion of resistant isolates (77.6%), followed by L2 (74.3%), L4 (63.4%), and
436 L1 (50.0%). L2 is known to be associated with MDR/RR and XDR and although it did
437 not contain the highest proportion of MDR/RR isolates, it did have the most pre-XDR
438 isolates as a proportion of MDR/RR isolates (54.2%) [30, 31]. L1 and L2 contained the

439 most XDR isolates as a proportion of MDR/RR isolates (4.7% and 4.2% of isolates

440 respectively).



441

442

443 **Figure 3:** ***Drug phenotype data for the CRyPTIC compendium*** **a)** Prevalence of
444 resistance to each of 13 drugs in the CRyPTIC dataset. The total number of isolates with a binary
445 phenotype (of any quality) for the corresponding drug is presented in Table 1. **b)** Phenotypes of
446 12, 289 CRyPTIC isolates with a binary phenotype for at least one drug. **c)** Geographical
447 distribution of phenotypes of 12, 289 CRyPTIC isolates. Intensity of blue shows the percentage of
448 isolates contributed that were categorised as susceptible to all 13 drugs. Donut plots show the

449 proportions of resistant phenotypes identified in (b) for countries contributing >=100 isolates with
450 drug resistance. **d)** Proportions of resistance phenotypes in the 4 major *M. tuberculosis* lineages.
451 N is the number of isolates of the lineage called resistant to at least one of the 13 drugs. Drug
452 acronyms: INH = isoniazid, RIF = rifampicin, EMB = ethambutol, LEV = levofloxacin, MXF =
453 moxifloxacin, AMI = amikacin, KAN = kanamycin, BDQ = bedaquiline, CFZ = clofazimine, DLM =
454 delamanid, LZD = linezolid, ETH = ethionamide, RFB = rifabutin.
455

456

457 ***Co-occurrence of drug resistance amongst the CRyPTIC isolates***

458 Isolates with all possible two-drug resistant combinations were present in this
459 dataset (Fig. 4a). Resistance to any of the remaining 11 drugs was associated with
460 resistance to both isoniazid and rifampicin. Isoniazid resistance was the most strongly
461 associated bar the following two exceptions, both of which arise because drugs from
462 the same class are being compared: a greater proportion of isolates resistant to
463 rifabutin were also resistant to rifampicin (96.8% had rifabutin - rifampicin co-
464 resistance versus 93.3% rifabutin - isoniazid co-resistance) and a greater proportion
465 of isolates resistant to moxifloxacin were associated with levofloxacin resistance
466 (97.6% had moxifloxacin – levofloxacin co-resistance versus 95.0% moxifloxacin –
467 isoniazid co-resistance). Of the second line drugs, levofloxacin and moxifloxacin were
468 more commonly seen as a second resistant phenotype than the injectable drugs
469 kanamycin and amikacin.

470

471 Despite this, resistance to both drugs in the aminoglycoside class was common
472 the dataset; 90.4% of amikacin resistant isolates were also resistant to kanamycin
473 although significantly fewer kanamycin resistant isolates were resistant to amikacin
474 (72.0%, p <0.00001). In a similar fashion, a smaller proportion of rifampicin resistant
475 isolates were resistant to rifabutin than rifabutin resistant isolates that were resistant
476 to rifampicin (91.3%, 96.8%, p <0.00001) while a smaller proportion of levofloxacin

477 resistant isolates were resistant to moxifloxacin than moxifloxacin resistant isolates
478 that were resistant to levofloxacin (78.5%, 97.6%, p <0.00001). One possible
479 explanation is that the cut off, in this case the ECOFFs, have been inconsistently
480 defined for the different compounds, however here all ECOFFs have been determined
481 in the same way using a much larger dataset than is typically the case [11]. Differences
482 in drugs of the same class are also well documented by *in vitro* studies [32-34].
483
484 Bedaquiline, clofazimine, delamanid and linezolid were les commonly seen as
485 a second resistance phenotype and the smallest proportional resistance combinations
486 involved the NRDs (*e.g.* 1.5% of isoniazid resistant isolates were bedaquiline resistant
487 and 1.7% were delamanid resistant). Within the NRDs however, co-occurrence of
488 resistance was proportionally higher; bedaquiline resistance was most seen with
489 clofazimine resistance (52.4%), clofazimine resistance was most seen with
490 bedaquiline resistance (10.6%), delamanid resistance was most seen with clofazimine
491 resistance (26.3%) and linezolid resistance was most seen with clofazimine resistance
492 (34.2%).
493
494
495 ***Additional antibiotic resistance in isolates with MDR or non-MDR phenotypic***
496 ***backgrounds***
497 We next examined in more detail the correlation structure of phenotypes by
498 conditioning on different phenotypic "backgrounds". (Fig. 4b-f). We found that a
499 greater proportion of isolates that were susceptible to isoniazid and rifampicin were
500 resistant to second-line drugs than the first line drug ethambutol (15.9%) (Fig. 4b). The
501 proportion of isolates resistant to clofazimine or levofloxacin was particularly high

502 (32.9% and 24.1%, respectively), and more isolates were resistant to these two drugs
503 than ethambutol in an isoniazid resistant and rifampicin susceptible background but
504 not in MDR/RR isolates (Fig. 4c-f).
505
506 MDR/RR isolates (asides from rifabutin) were most commonly resistant to the
507 first line drug ethambutol (46.3%), closely followed by levofloxacin (41.4%), then
508 moxifloxacin (34.1%). A greater proportion of isolates resistant to levofloxacin than
509 moxifloxacin was also seen in all other backgrounds and as expected, the proportion
510 of fluoroquinolone resistance was higher in MDR/RR isolates than non-MDR isolates
511 [35]. For the aminoglycosides, a greater percentage of MDR/RR isolates were
512 kanamycin resistant (21.8%) than amikacin resistant (18.1%), a trend seen in all other
513 backgrounds.
514
515 For isolates with an XDR phenotype, a higher proportion were also resistant to
516 linezolid than bedaquiline (66.7% compared to 44.6%) and 11.3% of XDR isolates
517 were resistant to both bedaquiline and linezolid (Fig. 4f). XDR isolates were also
518 resistant to the other NRDs, clofazimine (41.3%) and delamanid (18.8%). For isolates
519 with an MDR/RR or pre-XDR background, the most common NRD resistance seen
520 was clofazimine (6.0% and 9.3%, respectively) followed by linezolid (2.3% and 4.7%
521 respectively), bedaquiline (1.8% and 3.2%, respectively), and delamanid (1.7% and
522 2.3%, respectively). In non-MDR/RR backgrounds: clofazimine resistance remained
523 the most common followed by delamanid resistance, linezolid resistance and
524 bedaquiline resistance (Fig. 4b-c).
525
526

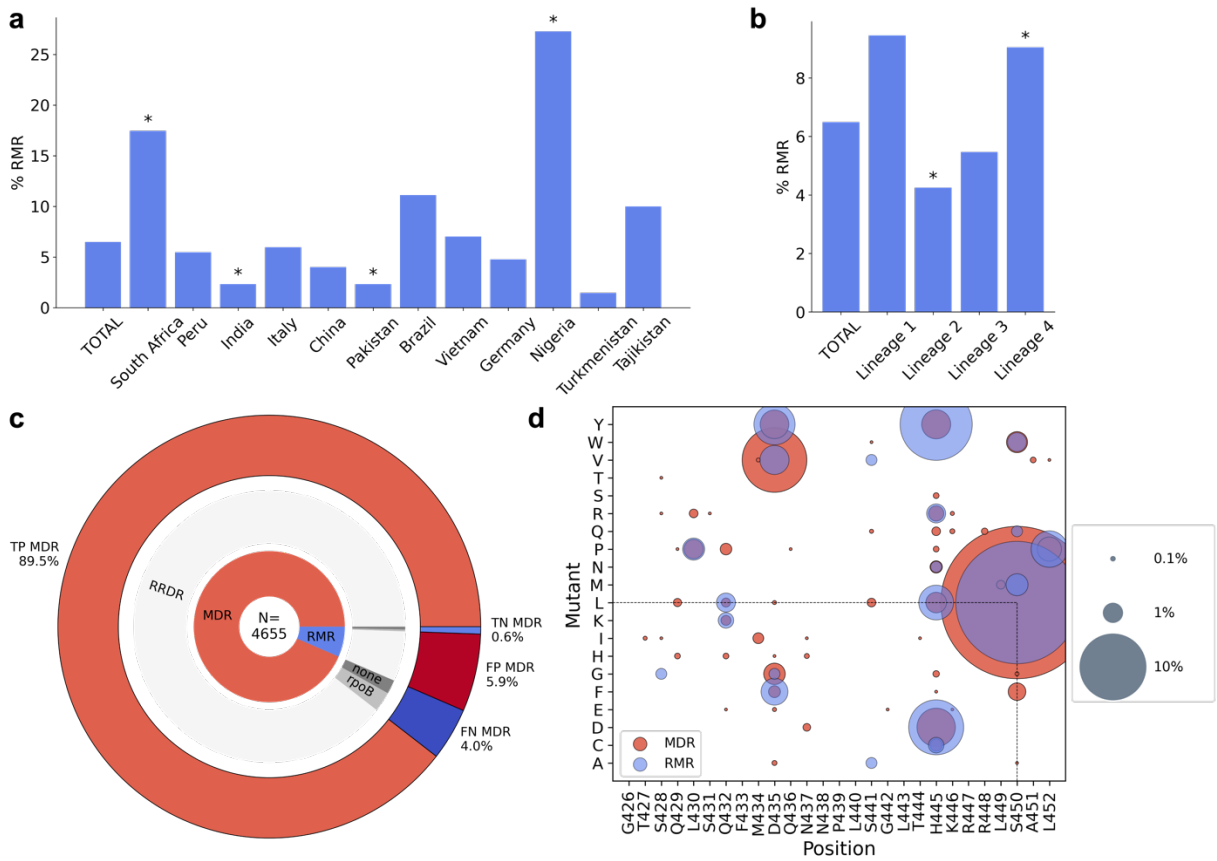


Figure 4: ***Co-occurrence of antibiotic resistance in CRyPTIC M. tuberculosis isolates*** **a)** The heatmap shows the probability of an isolate being resistant to Drug 2 if it is resistant to Drug 1, percentages are given in Table S4. **(b-g)** Percentage of isolates that are resistant to another of the 13 drugs in a background of **(b)** isoniazid susceptible + rifampicin susceptible, **(c)** isoniazid resistant + rifampicin susceptible, **(d)** MDR/RR **(e)** Pre-XDR, **(f)** XDR. Only samples with definite phenotypes for RIF in MDR backgrounds and RIF and INH in non-MDR backgrounds and the additional drug are included. Drug acronyms: INH = isoniazid, RIF = rifampicin, EMB = ethambutol, LEV = levofloxacin, MXF = moxifloxacin, AMI = amikacin, KAN = kanamycin, BDQ = bedaquiline, CFZ = clofazimine, DLM = delamanid, LZD = linezolid, ETH = ethionamide, RFB = rifabutin.

Genetic determinants of resistance

In order to establish a baseline measure of how well resistance and susceptibility could be predicted based on the state of the art prior to the CRyPTIC project, we constructed a hybrid catalogue of genetic variants associated with resistance to the first-line compounds isoniazid, rifampicin & ethambutol and the second-line compounds levofloxacin, moxifloxacin, amikacin, kanamycin, ethionamide based on existing catalogues [7, 36]. (We note in passing that we specifically did not

547 use the recent WHO catalogue in order to avoid circularity and over-training, as that
548 catalogue was developed (via prior literature, expert rules and a heuristic algorithm)
549 based partially on these isolates [37]. We applied our hybrid catalogue to predict the
550 susceptibility and resistance of all isolates in this compendium to these compounds,
551 and the resulting genetic-based predictions were then compared with the binary
552 phenotypes derived from the MICs (Table 2). Since these data were not collected
553 prospectively or randomly, but indeed are enriched for resistance, the calculated error
554 rates are not representative of how well such a method would perform in routine clinical
555 use. We also did not apply the approach used in [7] of refusing to make a prediction if
556 a novel mutation was detected in a known resistance gene, as we simply wanted to
557 measure how well a pre-CRyPTIC catalogue could predict resistance in the
558 compendium. The results were broadly in line with prior measurements on a smaller
559 (independent) set [18]. The hybrid catalogue does not make predictions for rifabutin,
560 linezolid, bedaquiline, delamanid or clofazimine; indeed, this is one of the main aims
561 of the consortium and new catalogues published by CRyPTIC and the WHO will begin
562 to address this shortcoming (Fig. 5) [19].
563
564
565 Table 3 shows the top mutations found amongst isolates phenotypically
566 resistant to first- and second- line drugs. As expected, *rpoB* S450L is the most
567 prevalent mutation associated with rifampicin resistance, with 62.2% of all
568 phenotypically rifampicin-resistant isolates carrying it, while 70% of isoniazid resistant
569 isolates have *katG* S315T mutations. Mutations in *gyrA* dominate amongst
570 fluoroquinolone resistant isolates; D94G and A90V are the two most frequently
571 occurring mutations for levofloxacin and moxifloxacin, with just under 77% and just
572 over 94% of phenotypically resistant isolates carrying at least one mutation in *gyrA*

573 respectively. Mutations in *rpoC* are amongst those most prevalent for rifabutin
574 resistant isolates; these mutations are typically described as compensatory, while
575 G594E was subsequently defined as a lineage-specific marker for L4 LAM and
576 Haarlem lineages [38].
577

	TP	FP	TN	FN	VME	ME	PPV	NPV
INH	5493	142	5622	224	0.039	0.025	0.961	0.975
RIF	4535	435	6669	107	0.023	0.061	0.977	0.939
EMB	1919	513	6702	111	0.055	0.071	0.945	0.929
LEV	1689	255	8104	184	0.098	0.031	0.902	0.969
MXF	1358	504	9022	160	0.105	0.053	0.895	0.947
AMI	632	84	10117	163	0.205	0.008	0.795	0.992
KAN	735	124	9043	197	0.211	0.014	0.789	0.986
ETH	971	114	9183	511	0.345	0.012	0.655	0.988

578
579 **Table 2: *Predicting phenotypic resistance using genetics***
580 TP: the number of phenotypically resistant samples are that correctly identified as resistant ("true
581 positives"); FP, the number of phenotypically susceptible samples that are falsely identified as
582 resistant ("false positives"); TN, the number of phenotypically susceptible samples that are
583 correctly identified as susceptible ("true negatives"); FN, the number of phenotypically resistant
584 samples that are incorrectly identified as susceptible ("false negative"); VME, very major error rate
585 (false-negative rate), 0-1; ME, major error rate (false-positive rate), 0-1; PPV, positive predictive
586 value, 0-1; NPV, negative predictive value, 0-1.
587
588
589
590
591
592
593
594
595
596
597

	GENE	VARIANT	*n*	%
RIF	*rpoB*	S450L	2914	62.2
	rpoB	D435V	506	10.8
	rpoB	H445D	202	4.3
	rpoB	H445Y	147	3.1
	rpoB	D435Y	112	2.4
INH	*katG*	S315T	3668	70.
	fabG1	c-15t	829	15.8
	fabG1	g-17t	176	3.4
	fabG1	t-8c	154	2.9
	inhA	I194T	56	1.1
EMB	*embB*	M306V	1131	50
	embB	M306I	1001	44.3
	embB	Q497R	449	19.9
	embB	G406A	164	7.2
	embB	G406D	105	4.6
KAN	*rrs*	a1401g	660	58.9
	eis	c-14t	70	6.2
	eis	g-10a	53	4.7
AMI	*rrs*	a1401g	660	74.7
	rrs	g1484t	7	0.8
LEV	*gyrA*	D94G	783	36.5
	gyrA	A90V	487	22.7
	gyrA	D94N	157	7.3
	gyrA	D94A	133	6.2
	gyrA	S91P	92	4.3
MXF	*gyrA*	A90V	487	28.2
	gyrA	D94G	783	45.4
	gyrA	D94N	157	9.1
	gyrA	D94A	133	7.7
	gyrA	D94Y	70	4.1
ETH	*fabG1*	c-15t	829	48
	fabG1	L203L	124	7.2

Table 3: ***The top mutations associated with phenotypic drug resistance*** A genetic resistance catalogue was used for genotypic prediction of resistance for each drug in column 1 [7, 36]. "VARIANT": non-synonymous amino acid mutations are denoted by upper case letters while nucleotide substitutions for non-coding sequences are denoted by lower case letters. Negative numbers denote substitutions in promoter regions; "GENE": genic region of interest in which "Variant" can be found; "*n*": number of phenotypically resistant isolates with "VARIANT"; "%": percentage of total phenotypically resistant isolates with "VARIANT". Drug acronyms: INH = isoniazid, RIF = rifampicin, EMB = ethambutol, AMI = amikacin, KAN = kanamycin, LEV = levofloxacin, MXF = moxifloxacin, ETH = ethionamide.

609

610 ***Resistance to new and re-purposed drugs***

611 As previously stated, relatively few isolates are resistant to the NRDs,
612 bedaquiline (*n* = 109), clofazimine (*n* = 525), delamanid (*n* = 186) and linezolid (*n* =
613 156). South Africa contributed the greatest number of isolates resistant to bedaquiline,
614 clofazimine and linezolid (Fig. 5a), while China and India contributed the most isolates
615 resistant to delamanid. Since the collection protocol differed between laboratories it is
616 not possible to infer any differences in the relative prevalence of resistance to the
617 NRDs in these countries. The results of a survey of all non-synonymous mutations in
618 genes known or suspected to be involved in resistance to these four drugs (*e.g.*
619 *rv0678, mmpL5, pepQ, ddn, rplC, rrl etc.*) are depicted in Fig. 5b-e [39-43]. Mutations
620 known to be associated with sensitivity were ignored, along with mutations that
621 occurred at a frequency of ≥ 5% amongst all isolates. In contrast to first- and second-
622 line drugs, there are no mutations within a single gene/small group of genes that can
623 fully explain resistance to any NRD. Note that the role of most of these mutations in
624 resistance remains undetermined. Work is ongoing within the consortium to unravel
625 the association of these (and novel) mutations with resistance to the NRDs along with
626 the relevant mechanisms of resistance.

627

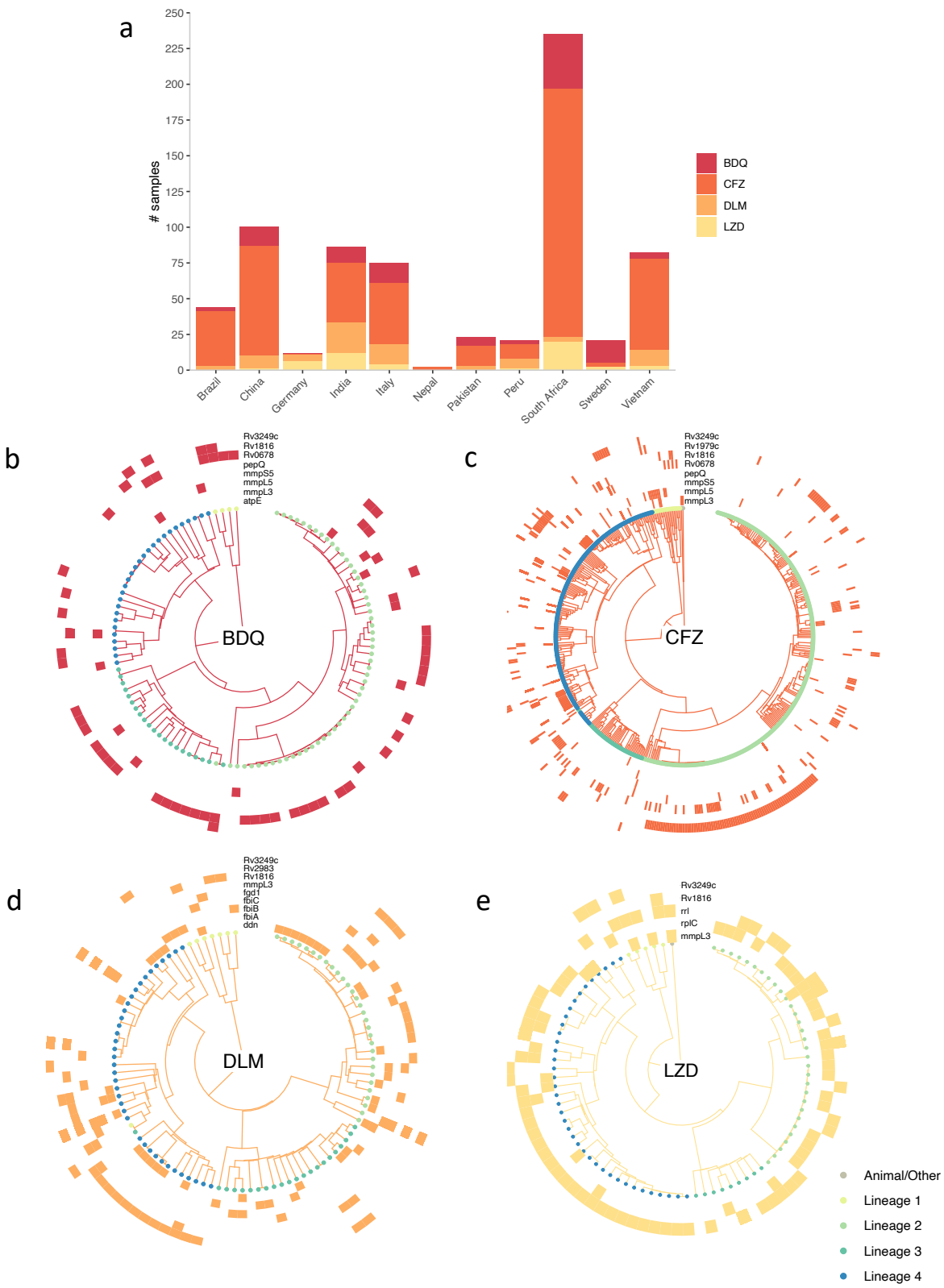


Figure 5: ***Resistance to bedaquiline, clofazimine, delamanid and linezolid amongst M. tuberculosis CRyPTIC isolates*** **a)** The prevalence of resistance to bedaquiline (BDQ), clofazimine (CFZ), delamanid (DLM) and linezolid (LZD) per country or origin or collection site. Phylotrees are shown for isolates phenotypically resistant to b) BDQ, c) CFZ, d) DLM and e) LZD. Tip point colours denote lineage. Each outer track represents a gene thought to

634 be associated with resistance and coloured blocks denote the presence of a non-synonymous
635 mutation in the relevant gene for a given isolate. Mutations in these genes that are either
636 associated with sensitivity or present in >5% of the collection of isolates as a whole were ignored.
637
638
639 ***Case study on rifampicin mono-resistance***

640 Around 1% of TB cases are rifampicin mono-resistant (RMR) and the frequency
641 is increasing [1, 44]. The WHO does not recommend isoniazid for RMR treatment,
642 despite it being effective; this is likely due to the reliance on the Xpert® MTB/RIF assay
643 which cannot distinguish between RMR and MDR. Use of isoniazid could improve
644 treatment outcomes for RMR patients which are currently similar to that of MDR TB,
645 including a higher risk of death compared to drug susceptible infections [45, 46]. Due
646 to its low natural prevalence, RMR has been poorly studied to date but increasingly
647 large clinical TB datasets, such as the one presented here, make its study now
648 feasible.
649
650 For this case study, we defined RMR as any isolate that was rifampicin resistant
651 and isoniazid susceptible, discounting isolates with no definite phenotype for either
652 isoniazid and/or rifampicin. Of the 4,655 rifampicin resistant isolates in the CRyPTIC
653 database that also had a phenotype for isoniazid, 302 (6.5%) were RMR. These
654 isolates were contributed by South Africa, Peru, India, Italy, China, Pakistan, Brazil,
655 Vietnam, Germany, Nigeria, Turkmenistan and Tajikistan (Fig. 6a). South Africa and
656 Nigeria contributed a significantly higher number of RMR isolates, as a proportion of
657 rifampicin resistant isolates, than that of the total dataset at 17.5% (p <0.00001) and
658 27.3% (p = 0.00534) respectively. We note that countries may have oversampled for
659 solo resistance, but the higher contribution of RMR isolates from South Africa is
660 consistent with previous studies [44].

661

662 RMR has evolved in all four major lineages, with L4 having a significantly higher
663 RMR prevalence than that of the total rifampicin resistant dataset at 9.4% (p =
664 0.00031) and L2 significantly lower RMR prevalence at 4.2% (p = 0.00026). Although
665 these numbers will have been somewhat influenced by differences in sample
666 collection bias at different geographical sites, we would expect L2 to have a lower
667 proportion of RMR isolates due to its association with MDR phenotypes [31, 47] (Fig.
668 6b).

669

670 ***Rifampicin mono-resistance is incorrectly predicted by current diagnostics***

671 A widely used, WHO-endorsed diagnostic tool, the Xpert® MTB/RIF assay,
672 uses a proxy whereby *any* SNP detected in the "rifampicin-resistance determining
673 region" (RRDR) of *rpoB* results in a prediction of MDR. However, the suitability of the
674 proxy is dependent upon prevalence of RMR in the population [44]. We tested the
675 reliability of this on the 4,655 rifampicin resistant isolates in our dataset that had a
676 phenotype for isoniazid (Fig. 6c).

677

678 Of these isolates, 4,353 (93.5%) were MDR and 302 (6.5%) were RMR. 187 of
679 the MDR isolates had no RRDR mutation and therefore 4.0% of isolates in this study
680 would be predicted as false negative MDR by the Xpert® MTB/RIF assay. 276 of the
681 RMR isolates had a mutation in the RRDR of *rpoB* and so the Xpert® MTB/RIF assay
682 proxy would incorrectly predict 5.9% of the rifampicin resistant isolates as false
683 positive MDR cases. However, overall, the Xpert® MTB/RIF assay proxy correctly
684 predicts 89.5% of the rifampicin resistant isolates as MDR and 0.6% of the isolates as
685 non-MDR in this dataset, which suggests it is a reasonably successful diagnostic tool

686 with >90% accuracy for MDR classification of rifampicin resistant isolates. As most
687 sites oversampled for resistance, our dataset likely contains a higher prevalence of
688 RMR than the global average and hence the Xpert® MTB/RIF assay is likely to perform
689 better on more representative data. However, the analysis shows how the increasing
690 global levels of RMR TB cases could increase the number of false positive MDR
691 diagnoses by the Xpert® MTB/RIF assay, denying isoniazid treatment to a greater
692 number of patients who would then be moved on to less effective drugs.
693
694 ***There are genetic differences between rifampicin mono-resistant and***
695 ***multidrug resistant isolates***
696 We have analysed our matched phenotypic and genotypic data to examine
697 whether there were any differences in the genetic determinants of rifampicin
698 resistance between RMR and MDR isolates. The proportion of RMR isolates with no
699 *rpoB* mutation (5.3%, Fig. 6c) was significantly higher than that of MDR isolates (1.8%,
700 p <0.00001). This suggests that non-target-mediated resistance mechanisms, such as
701 upregulation of rifampicin specific efflux pumps, could be more influential in providing
702 protection against rifampicin in RMR isolates than in MDR isolates.
703
704 The majority of RMR and MDR isolates contained one or more SNPs in *rpoB*,
705 with the majority having at least one mutation in the RRDR. To date, several non-
706 synonymous RRDR mutations have been found in RMR *M. tuberculosis* isolates,
707 including the resistance conferring mutations S450L, H445D and D435Y, which are
708 also seen in MDR isolates [48, 49]. For both RMR and MDR isolates in this dataset,
709 the most common *rpoB* RRDR mutation seen was S450L (63.6% and 41.1% of
710 isolates respectively, Fig. 6d). Five mutations were present in RMR isolates that were

711 not seen in MDR isolates: S428G, S441A, S441V, S450M and S450Q, however these
712 were seen at low prevalence (less than 2%) of RMR isolates. We found more RMR
713 isolates had His445 mutated than MDR isolates (27.8% of RMR and 9.5% of MDR, p
714 <0.00001), and mutations to amino acids Ser450 and Asp435 were more prevalent in
715 MDR isolates than RMR isolates (43.7% of RMR and 65.8% of MDR (p <0.00001),
716 and 9.3% of RMR and 15.5% of MDR (p = 0.00328) respectively).
717
718 Relatively little difference in the mutational profiles for *rpoB* mutations falling
719 outside the RRDR was found; many RMR mutations were also seen in MDR isolates
720 and those found exclusively in RMR were at <2% prevalence (Fig. S2). The most
721 common non-RRDR mutation in both MDR and RMR isolates was a cytosine to
722 thymine mutation 61 bases upstream of the *rpoB* start codon (10.1% and 8.6% of
723 isolates respectively). The resistance conferring mutations, V695L and V170F, were
724 seen at low proportions with no significant difference between MDR and RMR isolates
725 (V695L was seen in 0.5% MDR isolates and 1.3% of RMR isolates, and V170F was
726 seen in 0.8% of MDR isolates and 1.7% of RMR isolates).

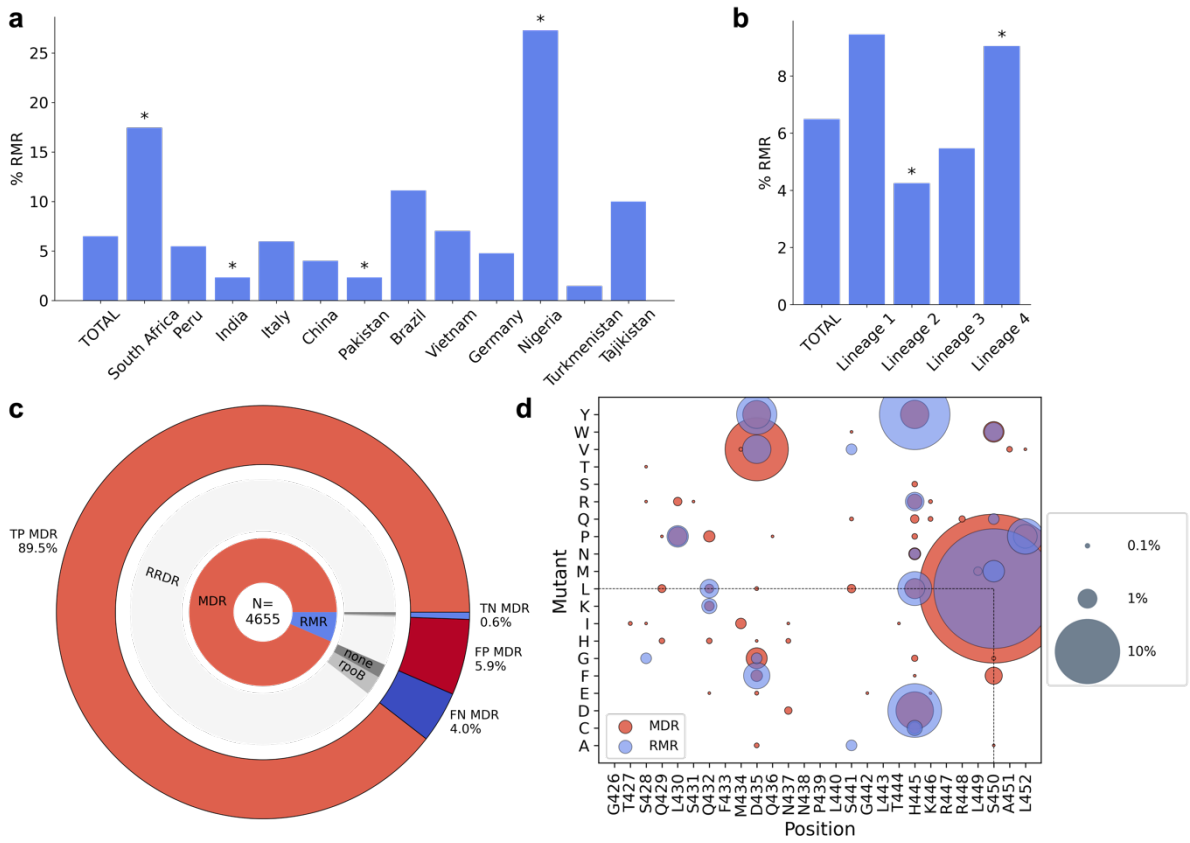


Figure 6: *Rifampicin mono resistance* a) Percentage of rifampicin resistant isolates that are rifampicin mono-resistant (RMR) by country of isolate origin. * indicates RMR proportions that were significantly different from that of the total dataset using a two tailed z-test with 95% confidence. **b)** RMR prevalence by lineage. * indicates RMR proportions that were significantly different from that of the total dataset using a two tailed z-test with 95% confidence. **c)** MDR predictions for rifampicin resistant isolates made using the Xpert® MTB/RIF assay proxy. N is the total number of rifampicin resistant isolates. The inner ring shows the proportion of rifampicin resistant isolates that are MDR and RMR. The middle ring represents the proportions of MDR and RMR isolates that have a SNP (synonymous or non-synonymous) in the RRDR of *rpoB* (RRDR), no RRDR SNP but a SNP elsewhere in the *rpoB* gene (rpoB) and no *rpoB* mutations (none). The outer ring shows the expected true positive (TP), true negative (TN), false positive (FP) and false negative (FN) MDR predictions of Xpert® MTB/RIF assay, based on the SNPs present in the rifampicin resistant isolates. **d)** Non-synonymous mutations found in the RRDR of *rpoB* in RMR isolates and MDR isolates. Presence of a coloured spot indicates that the mutation was found in RMR/MDR isolates and spot size corresponds to the proportion of RMR or MDR isolates carrying that mutation.

751 ## Discussion

752 This compendium of *M. tuberculosis* clinical isolates is the result of an extensive
753 global effort by the CRyPTIC consortium to better map the genetic variation associated
754 with drug resistance. Through its sheer size and by oversampling for resistance, the
755 compendium gives an unparalleled view of resistance patterns in the panel of 13
756 antitubercular compounds studied, providing more nuance to the traditional
757 classification (nonMDR/MDR/XDR) nomenclature. These insights could be used to
758 inform future treatment guidelines and steer the development of improved diagnostics.
759

760 For first-line drugs, molecular based diagnostic assays have vastly improved
761 the detection of and the speed at which we find drug resistant TB cases, resulting in
762 improved quality of care for patients. However, relying solely on these diagnostic
763 methods has several drawbacks. Aside from the Xpert® MTB/RIF assay potentially
764 increasing false positive MDR diagnoses as discussed earlier in the RMR case study,
765 the assay assumes isoniazid resistance upon detection of rifampicin resistance. Thus,
766 less is known about the prevalence of mono isoniazid resistance or 'true' cases of
767 MDR (confirmed rifampicin *and* isoniazid resistance) (REFWHO). A greater amount of
768 isoniazid-resistant and rifampicin-susceptible (Hr-TB) isolates were contributed by
769 CRyPTIC countries than RMR isolates *(n* = 1470 versus *n* = 302), a pattern also
770 recently observed in a global prevalence study [50]. A modified 6-month treatment
771 regimen is now recommended for Hr-TB (rifampicin, ethambutol, levofloxacin and
772 pyrazinamide), and as a result of inadequate diagnosis many of the 1.4 million global
773 Hr-TB estimated cases would have received inadequate and unnecessarily longer
774 treatment regimens [1, 50]. Encouragingly, CRyPTIC isolates with a Hr-TB
775 background exhibited relatively low levels of resistance to other antitubercular drugs,

776 including those in the augmented regimen (Fig. 4c). However, without the tools
777 appropriate to assess and survey this, we will continue to misdiagnose and infectively
778 treat these clinical cases.
779
780 In 2018, CRyPTIC and the 100,000 Genomes project demonstrated using WGS
781 that genotypic prediction correlates well with culture-based phenotype for first-line
782 drugs, which is reflected in our summary of the genetic catalogue applied to this
783 dataset (Table 3) [7]. While predictions can be made to a high level of sensitivity and
784 specificity, there is still more to learn, as exemplified by the isolates in the compendium
785 that despite being resistant to rifampicin and isoniazid could not be described
786 genetically (Table 2). This shortfall, along with the limitations of molecular based
787 diagnostic assays, highlights the need for continual genetic surveillance as drug
788 resistant cases rise and shines a favourable light on a WGS-led approach.
789
790 One of the particular strengths of this compendium lies with the data collated
791 for second-line drugs. A greater proportion of drug resistant isolates had additional
792 resistance to fluoroquinolones than second line injectable drugs (Fig. 4a). This could
793 be because of more widespread use of fluoroquinolones to treat bacterial infections in
794 general as well as their ease of administration, hence being recommended over
795 injectables for longer MDR treatment regimens [1]. Concerningly, we found that
796 resistance to levofloxacin and moxifloxacin, and kanamycin and amikacin, were more
797 common than resistance to the mycobacterial specific drug ethambutol in an isoniazid
798 and rifampicin susceptible background (Fig. 4b) which suggests a level of pre-existing
799 second line resistance. This concurs with a systematic review that found patients
800 previously prescribed fluoroquinolones were 3 times more likely to have

801 fluoroquinolone resistant TB [51]. Careful stewardship of fluoroquinolones both in TB
802 and other infectious diseases will be paramount for the success of treatment regimens.
803 Despite the variability in sample collection, the observed high proportions of
804 fluoroquinolone resistant MDR/RR isolates from some countries also suggests that
805 MDR treatment regimens could be improved by optimisation on a geographic basis.
806
807 Further improvement could also be made by the selection of appropriate drugs
808 from each class; for example, the WHO recommends switching from kanamycin to
809 amikacin when treating MDR TB patients [52]. The compendium supports such
810 recommendations as we saw more resistance to kanamycin than amikacin in not just
811 MDR/RR but also all other phenotypic backgrounds. For fluoroquinolones, more
812 isolates were resistant to levofloxacin than moxifloxacin in all phenotypic backgrounds
813 suggesting moxifloxacin may by the most appropriate fluoroquinolone to recommend.
814 However, the amenability of drugs to catalogue-based genetic diagnostics is also an
815 important consideration, and our data suggest levofloxacin resistance could be
816 predicted more reliably than moxifloxacin with fewer false positives predicted (Table
817 2).
818
819 Testing for fluoroquinolone resistance using molecular diagnostic tests remains
820 limited. However, the limited global data from the past 15 years suggests that the
821 proportion of MDR/RR TB cases resistant to fluoroquinolones sits at around 20%, with
822 these cases primarily found in regions of high MDR-TB burden [1]. While tools such
823 as the Cepheid Xpert® MTB/XDR cartridge, which has just been approved for use by
824 the WHO, will permit both isoniazid and fluoroquinolone testing to be increased, the
825 same pitfalls are to be encountered regarding targeted diagnostic assays [53]. In

826 contrast, the genetic survey in this study demonstrates the potential of WGS for
827 genetic prediction of resistance to second-line drugs and studies within the consortium
828 to investigate this are underway.
829
830 The CRyPTIC compendium has facilitated the first global survey of resistance
831 to NRDs. Reassuringly, prevalence of resistance to the NRDs was lower than for first-
832 and second-line agents in the dataset as a whole (Fig. 3a), and resistance to the new
833 drugs bedaquiline and delamanid was less common than the repurposed drugs
834 clofazimine and linezolid in an MDR/RR background (Fig. 4c). However, the presence
835 of higher levels of delamanid and clofazimine resistance than ethambutol resistance
836 in the isoniazid and rifampicin susceptible background does suggest some pre-existing
837 propensity towards NRD resistance (Fig. 4b).
838
839 Co-resistance between all two-drug combinations of NRDs was seen in isolates
840 in the compendium, the most common being isolates resistant to both bedaquiline and
841 clofazimine. This link is well documented and has been attributed to shared resistance
842 mechanisms such as non-synonymous mutations in *rv0678* which were found in both
843 clofazimine and bedaquiline resistant isolates in the compendium [42] (Fig. 5b,c).
844 Increased clofazimine use could further increase the prevalence of *M. tuberculosis*
845 isolates with clofazimine and bedaquiline co-resistance, limiting MDR treatment
846 options such as using bedaquiline as the backbone of a shorter MDR regimen [54].
847 Therefore, proposed usage of clofazimine for other infectious diseases, including
848 COVID-19, should be carefully considered [55].
849

850 The WHO recommends not using bedaquiline and delamanid in combination to
851 prevent the development of co-resistance, which could occur relatively quickly [56];
852 the rates of spontaneous evolution of delamanid resistance *in vitro* has been shown
853 to be comparable to that of isoniazid, and likewise bedaquiline resistance arises at a
854 comparable rate to rifampicin resistance [57]. In this compendium, 12.9% bedaquiline
855 resistant isolates were resistant to delamanid and 7.1% delamanid resistant isolates
856 were resistant to bedaquiline. Several scenarios could account for this, including the
857 presence of shared resistance mechanisms. For example, as bedaquiline targets
858 energy metabolism within the cell, changes to cope with energy/nutrient imbalances
859 upon the acquisition of resistance-associated ATPase pump mutations may result in
860 cross resistance to delamanid in a yet unknown or unexplored mechanism [58].
861 Additionally, off-target mutations in the *mmpl3* gene (involved in export), are
862 associated with resistance to delamanid, bedaquiline, clofazimine and linezolid, and
863 were seen in isolates resistant to each of these drugs, although we note this could be
864 artefactual (Fig. 5b-e). In any case, it is imperative that genetic determinants of
865 resistance are fully explored for our last resort NRDs, with special attention given to
866 those mechanisms that could be shared with other agents. In the meantime, careful
867 stewardship and phenotypic and genotypic surveillance of the NRDs should be
868 implemented, including linezolid and clofazimine which are now group A and B drugs
869 respectively for MDR treatment [1].
870
871 Several research avenues are being actively explored by the CRyPTIC
872 consortium that make use of this compendium, including: *i)* relating genetic mutations
873 to quantitative changes in the minimum inhibitory concentrations of different drugs
874 [58]; *ii)* genome-wide association studies [59]; *iii)* training machine learning models

875 that can predict resistance [60]; *iv*) exploration of the genetic determinants of
876 resistance to second line and NRDs [61]. Collectively these studies share the same
877 aim of facilitating the implementation of WGS-directed resistance prediction in the
878 clinic. Finally, we urge other researchers to explore this large dataset of *M.*
879 *tuberculosis* clinical isolates and hope it will lead to a wave of new and inciteful studies
880 that will positively serve the TB community for years to come.
881
882
883
884
885
886 **Author contributions**
887 D.W.C, T.E.A.P, S.H, A.L.G.C, A.W.S, T.M.W, P.W.F, D.M.C designed the CRyPTIC
888 study and all contributing laboratories collected samples and provided data. MIC data
889 and genetic information was retrieved and processed by P.W.F, S.H, A.L.G.C, Z.I, M.H
890 and J.K. A.B and K.M.M designed and performed all analyses for this manuscript. A.B
891 and K.M.M wrote the manuscript with feedback from CRyPTIC partners.
892
893
894 **Acknowledgements**
895 We thank Faisal Masood Khanzada and Alamdar Hussain Rizvi (NTRL, Islamabad,
896 Pakistan), Angela Starks and James Posey (Centers for Disease Control and
897 Prevention, Atlanta, USA), and Juan Carlos Toro and Solomon Ghebremichael (Public
898 Health Agency of Sweden, Solna, Sweden), Iñaki Comas and Álvaro Chiner-Oms
899 (Instituto de Biología Integrativa de Sistemas, Valencia, Spain; CIBER en

Epidemiología y Salud Pública, Valencia, Spain; Instituto de Biomedicina de Valencia, IBV-CSIC, Valencia, Spain).

Funding bodies

This work was supported by Wellcome Trust/Newton Fund-MRC Collaborative Award (200205/Z/15/Z); and Bill & Melinda Gates Foundation Trust (OPP1133541). Oxford CRyPTIC consortium members are funded/supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health, and the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, a partnership between Public Health England and the University of Oxford, the views expressed are those of the authors and not necessarily those of the NIHR, Public Health England or the Department of Health and Social Care. J.M. is supported by the Wellcome Trust (203919/Z/16/Z). Z.Y. is supported by the National Science and Technology Major Project, China Grant No. 2018ZX10103001. K.M.M. is supported by EMBL's EIPOD3 programme funded by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska Curie Actions. T.C.R. is funded in part by funding from Unitaid Grant No. 2019-32-FIND MDR. R.S.O. is supported by FAPESP Grant No. 17/16082-7. L.F. received financial support from FAPESP Grant No. 2012/51756-5. B.Z. is supported by the National Natural Science Foundation of China (81991534) and the Beijing Municipal Science & Technology Commission (Z201100005520041). N.T.T.T. is supported by the Wellcome Trust International Intermediate Fellowship (206724/Z/17/Z). G.T. is funded by the Wellcome Trust. R.W. is supported by the South African Medical Research Council.

925 J.C. is supported by the Rhodes Trust and Stanford Medical Scientist Training
926 Program (T32 GM007365). A.L. is supported by the National Institute for Health
927 Research (NIHR) Health Protection Research Unit in Respiratory Infections at Imperial
928 College London. S.G.L. is supported by the Fonds de Recherche en Santé du Québec.
929 C.N. is funded by Wellcome Trust Grant No. 203583/Z/16/Z. A.V.R. is supported by
930 Research Foundation Flanders (FWO) under Grant No. G0F8316N (FWO Odysseus).
931 G.M. was supported by the Wellcome Trust (098316, 214321/Z/18/Z, and
932 203135/Z/16/Z), and the South African Research Chairs Initiative of the Department
933 of Science and Technology and National Research Foundation (NRF) of South Africa
934 (Grant No. 64787). The funders had no role in the study design, data collection, data
935 analysis, data interpretation, or writing of this report. The opinions, findings and
936 conclusions expressed in this manuscript reflect those of the authors alone. L.G. was
937 supported by the Wellcome Trust (201470/Z/16/Z), the National Institute of Allergy and
938 Infectious Diseases of the National Institutes of Health under award number
939 1R01AI146338, the GOSH Charity (VC0921) and the GOSH/ICH Biomedical
940 Research Centre (www.nihr.ac.uk). A.B. is funded by the NDM Prize Studentship from
941 the Oxford Medical Research Council Doctoral Training Partnership and the Nuffield
942 Department of Clinical Medicine. D.J.W. is supported by a Sir Henry Dale Fellowship
943 jointly funded by the Wellcome Trust and the Royal Society (Grant No. 101237/Z/13/B)
944 and by the Robertson Foundation. A.S.W. is an NIHR Senior Investigator. T.M.W. is a
945 Wellcome Trust Clinical Career Development Fellow (214560/Z/18/Z). A.S.L. is
946 supported by the Rhodes Trust. R.J.W. receives funding from the Francis Crick
947 Institute which is supported by Wellcome Trust, (FC0010218), UKRI (FC0010218),
948 and CRUK (FC0010218). T.C. has received grant funding and salary support from US
949 NIH, CDC, USAID and Bill and Melinda Gates Foundation. The computational aspects

of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. Parts of the work were funded by the German Center of Infection Research (DZIF). The Scottish Mycobacteria Reference Laboratory is funded through National Services Scotland. The Wadsworth Center contributions were supported in part by Cooperative Agreement No. U60OE000103 funded by the Centers for Disease Control and Prevention through the Association of Public Health Laboratories and NIH/NIAID grant AI-117312. Additional support for sequencing and analysis was contributed by the Wadsworth Center Applied Genomic Technologies Core Facility and the Wadsworth Center Bioinformatics Core. SYNLAB Holding Germany GmbH for its direct and indirect support of research activities in the Institute of Microbiology and Laboratory Medicine Gauting. N.R. thanks the Programme National de Lutte contre la Tuberculose de Madagascar.

Wellcome Trust Open Access

This research was funded in part, by the Wellcome Trust/Newton Fund-MRC Collaborative Award [200205/Z/15/Z]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

This research was funded, in part, by the Wellcome Trust [214321/Z/18/Z, and 203135/Z/16/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Competing Interest

977 E.R. is employed by Public Health England and holds an honorary contract with
978 Imperial College London. I.F.L. is Director of the Scottish Mycobacteria Reference
979 Laboratory. S.N. receives funding from German Center for Infection Research,
980 Excellenz Cluster Precision Medicine in Chronic Inflammation, Leibniz Science
981 Campus Evolutionary Medicine of the LUNG (EvoLUNG)tion EXC 2167. P.S. is a
982 consultant at Genoscreen. T.R. is funded by NIH and DoD and receives salary support
983 from the non-profit organization FIND. T.R. is a co-founder, board member and
984 shareholder of Verus Diagnostics Inc, a company that was founded with the intent of
985 developing diagnostic assays. Verus Diagnostics was not involved in any way with
986 data collection, analysis or publication of the results. T.R. has not received any
987 financial support from Verus Diagnostics. UCSD Conflict of Interest office has
988 reviewed and approved T.R.'s role in Verus Diagnostics Inc. T.R. is a co-inventor of a
989 provisional patent for a TB diagnostic assay (provisional patent #: 63/048.989). T.R. is
990 a co-inventor on a patent associated with the processing of TB sequencing data
991 (European Patent Application No. 14840432.0 & USSN 14/912,918). T.R. has agreed
992 to "donate all present and future interest in and rights to royalties from this patent" to
993 UCSD to ensure that he does not receive any financial benefits from this patent. S.S.
994 is working and holding ESOPs at HaystackAnalytics Pvt. Ltd. (Product: Using whole
995 genome sequencing for drug susceptibility testing for *Mycobacterium tuberculosis*).
996 G.F.G. is listed as an inventor on patent applications for RBD-dimer-based CoV
997 vaccines. The patents for RBD-dimers as protein subunit vaccines for SARS-CoV-2
998 have been licensed to Anhui Zhifei Longcom Biopharmaceutical Co. Ltd, China.
999
1000
1001

Supplemental

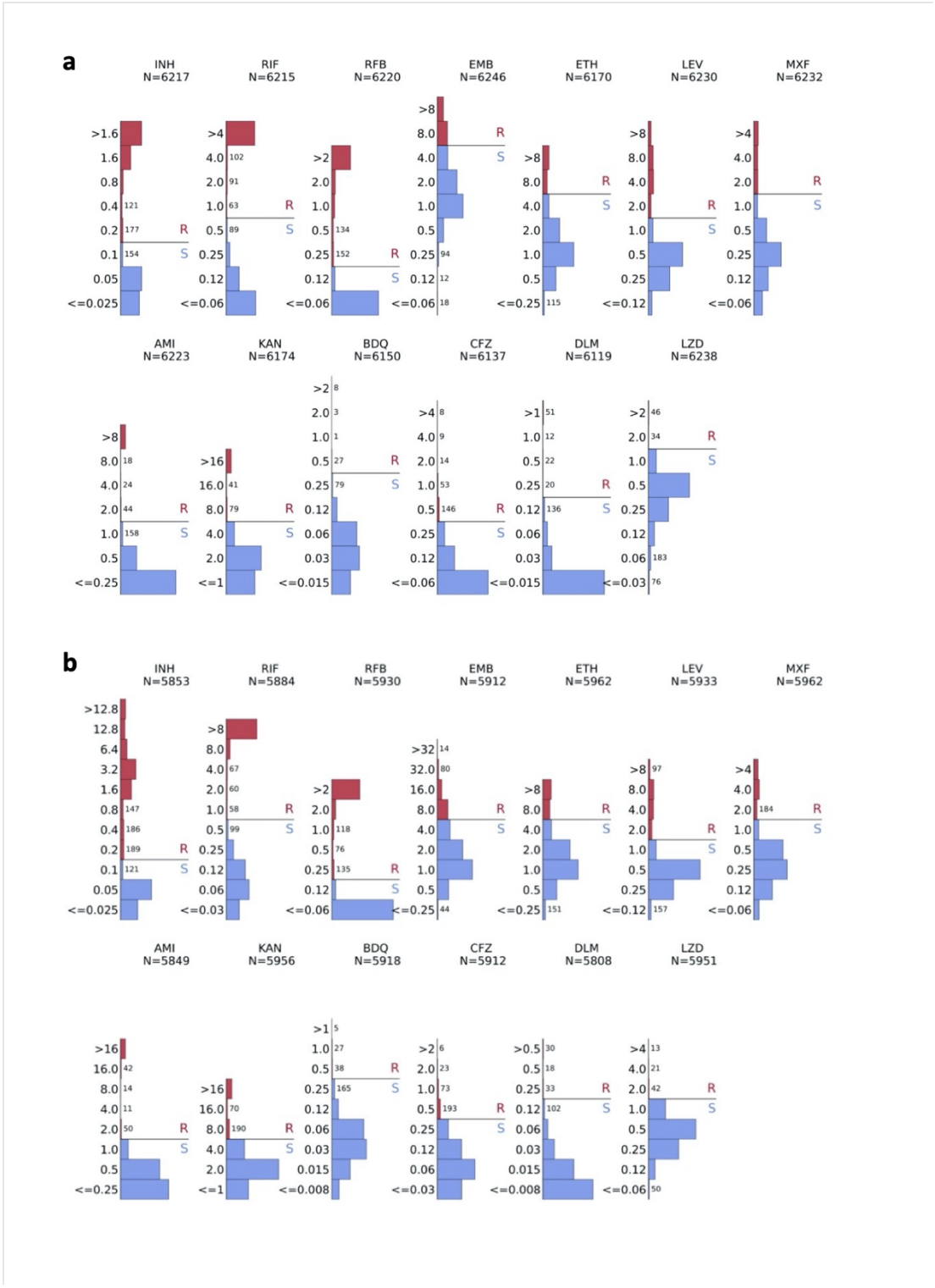


Figure S1: ***Per drug MIC distributions of isolates plated on CRyPTIC designed variations on the Thermo Fischer Sensititre MYCOTB MIC plate; UKMYC5 (a) and UKMYC6 (b).*** The solid black line represents the ecological resistance cut-off (ECOFF) for each drug as determined by [11]. Isolates with an MIC above the cut-off are considered resistant. N denotes the total number of isolates tested on each plate that returned a phenotype for each drug.

UNIQUEID	COUNTRY OF ORIGIN	LINEAGE
site.11.subj.XTB-18-224.lab.XTB-18-224.iso.1	UNKNOWN	Lineage 2
site.10.subj.YA00026182.lab.YA00026182.iso.1	ZAF	Lineage 4

Table S1: ***Sample information for isolates classified as resistant to all 13 CRyPTIC drugs tested.*** The country of origin is specified using the 3-letter country codes (alpha-3) defined by ISO 3166-1.

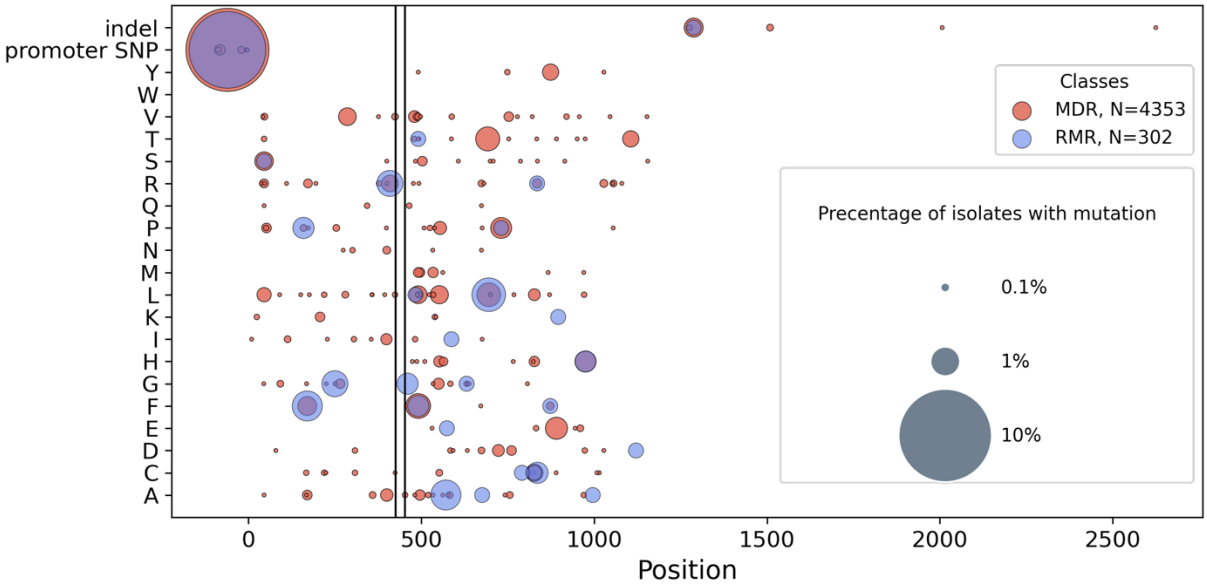


Figure S2: ***Non-synonymous mutations found outside the RRDR of rpoB in RMR isolates and MDR isolates.*** Presence of a coloured spot indicates that the mutation was found in RMR/MDR isolates and spot size corresponds to the proportion of RMR or MDR isolates carrying that mutation.

	Animal/ other	L1	L2	L3	L4	L6	Total
Albania	0	0	0	0	7	0	7
Algeria	0	0	0	0	25	0	25
Brazil	1	1	8	0	335	0	345
Burkina Faso	0	3	1	8	65	1	78
China	7	2	722	16	374	0	1121
Germany	1	25	235	61	403	1	726
India	11	676	1573	1184	560	0	4004
Italy	11	38	192	113	866	4	1224
Japan	0	0	0	0	1	0	1
Kyrgyzstan	0	0	25	0	3	0	28
Nepal	0	5	137	34	21	0	197
Nigeria	0	0	0	0	15	0	15
Pakistan	1	30	23	363	74	0	491
Peru	2	1	360	0	2328	0	2691
South Africa	1	61	874	57	1162	0	2155
Sweden	0	7	280	6	134	0	427
Taiwan	0	18	149	0	44	0	211
Tajikistan	0	0	15	0	4	0	19
Tunisia	0	0	0	0	2	0	2
Turkmenistan	0	0	88	1	30	0	119
Ukraine	0	0	21	0	9	0	30
UK	0	0	3	3	1	0	7
Vietnam	0	283	892	4	109	0	1288
Total	35	1150	5598	1850	6572	6	15211

Table S2: ***Lineages –v- geographical location of origin/contribution for CRyPTIC isolates***

ALB	ALG	BRZ	BFA	CHN	GER	IND	ITL	JPN	KGZ	NPL	NGA	PAK	PER	ZAF	SWE	TWN	TJK	TUN	TKM	UKR	UK	VNM

1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.1.1	0	0	0	2	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38
1.1.1.1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	188
1.1.2	0	0	0	1	1	12	316	16	0	0	2	0	23	0	0	1	0	0	0	0	0	0	0
1.1.3	0	0	1	0	0	1	52	4	0	0	1	0	1	0	6	0	0	0	0	0	0	0	2
1.2.1	0	0	0	0	0	8	12	11	0	0	0	0	0	1	0	3	16	0	0	0	0	0	24
1.2.2	0	0	0	0	1	2	259	5	0	0	2	0	4	0	54	3	1	0	0	0	0	0	1
2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	20	0	0	0	0	0	34
2.2	0	0	5	1	284	19	272	14	0	0	68	0	6	231	181	13	56	0	0	0	0	0	271
2.2.1	0	0	0	0	37	4	2	2	0	0	0	0	0	3	224	1	21	0	0	0	0	0	26
2.2.10	0	0	0	0	1	117	1	84	0	16	0	0	3	0	0	70	0	5	0	75	9	1	0
2.2.2	0	0	0	0	32	8	5	11	0	1	0	0	1	0	0	1	5	0	0	0	1	0	8
2.2.3	0	0	2	0	172	3	290	4	0	0	21	0	2	1	3	1	10	0	0	0	0	0	100
2.2.4	0	0	0	0	3	3	4	8	0	0	0	0	0	0	37	1	1	0	0	0	0	0	70
2.2.5	0	0	0	0	59	1	12	5	0	0	0	0	2	84	1	1	13	0	0	0	0	0	45
2.2.6	0	0	1	0	22	8	226	2	0	0	9	0	7	4	204	1	4	0	0	0	0	0	154
2.2.7	0	0	0	0	97	4	691	5	0	1	38	0	0	20	206	5	18	0	0	0	0	0	164
2.2.8	0	0	0	0	4	0	42	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	2
2.2.9	0	0	0	0	0	64	0	51	0	6	0	0	0	2	0	174	0	10	0	11	11	2	0
3	0	0	0	8	15	37	828	81	0	0	28	0	346	0	34	5	0	0	0	1	0	2	2
3.1.1	0	0	0	0	0	17	17	25	0	0	1	0	0	0	19	1	0	0	0	0	0	1	2
3.1.2	0	0	0	0	1	5	201	3	0	0	1	0	7	0	2	0	0	0	0	0	0	0	0
3.1.2.1	0	0	0	0	0	2	119	3	0	0	4	0	6	0	0	0	0	0	0	0	0	0	0
4	0	0	1	0	6	4	0	6	0	0	0	0	0	100	1	1	0	0	0	1	0	0	0
4.1	0	0	0	1	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.1.1	0	0	0	0	0	6	5	10	0	0	1	0	0	154	13	1	0	0	0	0	0	0	0
4.1.1.1	0	0	5	0	1	3	20	5	0	0	1	0	2	5	54	1	0	0	0	0	0	0	1
4.1.1.2	0	0	0	0	0	0	4	1	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
4.1.1.3	0	0	1	6	1	0	30	7	1	0	3	0	2	68	138	0	0	0	0	0	0	0	1
4.1.2	0	0	13	0	0	15	3	26	0	1	0	0	0	31	43	0	0	0	0	0	0	0	10
4.1.2.1	1	8	87	10	0	111	56	183	0	0	3	2	4	653	54	5	0	1	0	0	1	0	17
4.1.3	0	0	0	5	0	0	0	6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4.10	4	6	35	5	1	110	166	207	0	1	3	0	28	128	170	4	5	2	1	6	2	0	12
4.2	0	0	0	0	3	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.2.1	0	0	0	0	0	9	3	46	0	0	1	0	2	0	2	4	0	0	0	8	3	1	0
4.2.2	0	1	0	0	53	31	76	51	0	0	5	0	12	2	10	10	1	0	0	0	0	0	11
4.2.2.1	0	0	0	0	0	5	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.3	0	0	17	0	0	0	0	1	0	0	0	0	0	96	1	0	0	0	0	0	0	0	0
4.3.1	0	3	0	1	0	1	2	21	0	0	0	0	0	8	0	1	0	0	0	0	0	0	1
4.3.2	0	1	10	0	0	3	1	8	0	0	0	0	0	245	21	0	0	0	0	0	0	0	0
4.3.2.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	234	0	0	0	0	0	0	0	0
4.3.3	1	4	34	2	0	33	99	81	0	1	1	0	0	576	120	102	2	1	1	11	2	0	1
4.3.4	0	0	1	0	0	0	8	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4.3.4.1	0	0	70	0	0	2	6	3	0	0	0	0	0	60	22	2	1	0	0	0	0	0	0
4.3.4.2	0	0	38	0	0	6	7	27	0	0	0	0	0	132	12	0	0	0	0	0	1	0	0
4.3.4.2.1	0	0	0	0	0	2	1	1	0	0	0	0	0	0	75	0	0	0	0	0	0	0	0
4.4	0	0	0	0	0	0	13	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4.4.1	0	0	10	0	0	0	0	2	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0
4.4.1.1	0	0	4	0	0	16	4	66	0	0	1	0	0	22	180	2	0	0	0	0	0	0	4
4.4.1.2	0	0	0	0	0	1	25	3	0	0	0	0	0	2	0	0	0	0	0	0	0	0	3
4.4.2	0	0	0	0	124	1	4	1	0	0	0	0	0	4	0	0	5	0	0	0	0	0	16
4.5	0	0	0	0	179	28	15	14	0	0	1	0	18	5	1	0	30	0	0	3	0	0	29
4.6	0	1	0	1	0	1	1	7	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
4.6.1.1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.6.1.2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4.6.2	0	0	0	0	0	2	2	14	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
4.6.2.1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
4.6.2.2	1	0	0	34	0	8	1	55	0	0	0	10	1	0	1	0	0	0	0	0	0	0	3
6	0	0	0	1	0	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Bovis	0	0	0	0	0	1	1	10	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
Caprae	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mixed	0	1	9	0	16	3	89	9	0	1	0	0	6	38	25	12	2	0	0	3	0	0	48
Unknown	0	0	1	0	7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Table S3: ***Sub-lineages –v- geographical location of origin/contribution for CRyPTIC isolates***

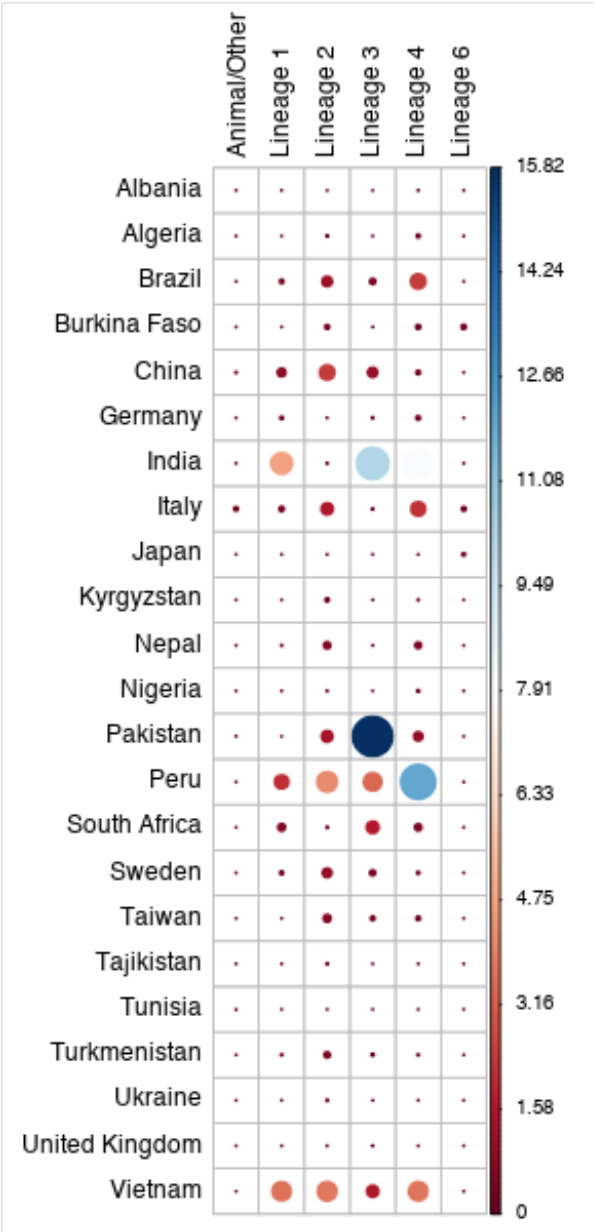


Figure S3: ***A significant association between country and lineage can be seen in the CRyPTIC data.*** Pearson's chi-squared test, X-squared = 7935.2, df = 110, $p < 2.2e-16$. The correlation plot indicates the relative contribution of each row-column pairing to the chi-square test score (%).

	INH	RIF	EMB	LEV	MXF	AMI	KAN	BDQ	CFZ	DLM	LZD	ETH	RFB
INH	100.0	74.8	38.0	34.0	27.8	14.0	17.0	1.5	5.5	1.7	2.0	28.2	70.4
RIF	93.5	100.0	46.3	41.4	34.1	17.2	20.8	1.8	6.0	1.7	2.3	29.4	91.3
EMB	98.5	95.9	100.0	53.9	47.1	23.3	26.4	2.4	8.2	2.3	3.7	35.8	87.3
LEV	93.3	90.2	56.9	100.0	78.5	27.3	31.3	3.1	9.4	2.9	4.7	39.2	87.1
MXF	95.0	92.3	61.9	97.6	100.0	29.9	34.6	3.2	9.9	2.7	5.3	41.9	88.8
AMI	93.0	90.4	58.9	65.3	57.8	100.0	90.4	2.4	13.0	5.2	8.1	40.7	82.8
KAN	89.4	86.7	53.0	59.5	53.2	72.0	100.0	1.8	10.6	3.2	5.7	40.4	80.0
BDQ	79.4	77.8	49.5	60.7	49.5	20.0	18.9	100.0	52.4	12.9	14.0	34.9	76.9
CFZ	61.9	52.8	34.9	38.2	32.1	21.8	22.5	10.6	100.0	9.4	10.2	27.0	53.7
DLM	55.2	44.4	28.2	33.2	23.9	24.4	19.1	7.1	26.3	100.0	19.3	22.1	45.4
LZD	77.6	69.9	53.2	64.7	58.7	46.7	41.2	9.8	34.2	24.1	100.0	42.6	66.7
ETH	96.5	79.4	47.0	48.7	41.9	21.1	26.2	2.2	8.3	2.4	3.9	100.0	77.1
RFB	93.3	96.8	44.5	42.1	34.4	16.6	20.3	1.9	6.4	1.9	2.3	29.9	100.0

Table S4: *Co-occurrence of antibiotic resistance in CRyPTIC M. tuberculosis isolates*. The probability (%) of an isolate being resistant to Drug 2 (top) if it is resistant to Drug 1 (left). Drug acronyms: INH = isoniazid, RIF = rifampicin, EMB = ethambutol, LEV = levofloxacin, MXF = moxifloxacin, AMI = amikacin, KAN = kanamycin, BDQ = bedaquiline, CFZ = clofazimine, DLM = delamanid, LZD = linezolid, ETH = ethionamide, RFB = rifabutin.

References

1. WHO, *World Health Organisation: Global Tuberculosis Report 2020*. 2020.
2. Boehme, C.C., et al., *Feasibility, diagnostic accuracy, and effectiveness of decentralised use of the Xpert MTB/RIF test for diagnosis of tuberculosis and multidrug resistance: a multicentre implementation study.* Lancet, 2011. **377**(9776): p. 1495-505.
3. Shinnick, T.M., et al., *Evaluation of the Cepheid Xpert MTB/RIF assay.* Expert Rev Mol Diagn, 2015. **15**(1): p. 9-22.
4. Sanchez-Padilla, E., et al., *Detection of Drug-Resistant Tuberculosis by Xpert MTB/RIF in Swaziland.* New England Journal of Medicine, 2015. **372**(12): p. 1181-1182.
5. Beckert, P., et al., *MDR M. tuberculosis outbreak clone in Eswatini missed by Xpert has elevated bedaquiline resistance dated to the pre-treatment era.* Genome Medicine, 2020. **12**(1): p. 104.
6. Makhado, N.A., et al., *Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study.* Lancet Infect Dis, 2018. **18**(12): p. 1350-1359.
7. Consortium, T.C., et al., *Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing.* N Engl J Med, 2018. **379**(15): p. 1403-1415.
8. Pankhurst, L.J., et al., *Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study.* Lancet Respir Med, 2016. **4**(1): p. 49-58.

9. Kalokhe, A.S., et al., *Multidrug-resistant tuberculosis drug susceptibility and molecular diagnostic testing.* Am J Med Sci, 2013. **345**(2): p. 143-8.
10. Rancoita, P.M.V., et al., *Validating a 14-Drug Microtiter Plate Containing Bedaquiline and Delamanid for Large-Scale Research Susceptibility Testing of Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2018. **62**(9).
11. Consortium, T.C. and P.W. Fowler, *Epidemiological cutoff values for a 96-well broth microdilution plate for high-throughput research antibiotic susceptibility testing of M. tuberculosis.* medRxiv, 2021: p. 2021.02.24.21252386.
12. Fowler, P.W., et al., *Automated detection of bacterial growth on 96-well plates for high-throughput drug susceptibility testing of Mycobacterium tuberculosis.* Microbiology (Reading), 2018. **164**(12): p. 1522-1530.
13. Fowler, P.W., et al., *BashTheBug: a crowd of volunteers reproducibly and accurately measure the minimum inhibitory concentrations of 13 antitubercular drugs from photographs of 96-well broth microdilution plates.* bioRxiv, 2021: p. 2021.07.20.453060.
14. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
15. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
16. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
17. Iqbal, Z., et al., *De novo assembly and genotyping of variants using colored de Bruijn graphs.* Nat Genet, 2012. **44**(2): p. 226-32.
18. Hunt, M., et al., *Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe.* Wellcome Open Research, 2019. **4**(191).
19. Walker, T.M., et al., *Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study.* Lancet Respir Med, 2014. **2**(4): p. 285-292.
20. Howe, K., A. Bateman, and R. Durbin, *QuickTree: building huge Neighbour-Joining trees of protein sequences.* Bioinformatics, 2002. **18**(11): p. 1546-7.
21. Yu, G., *Using ggtree to Visualize Data on Tree-Like Structures.* Curr Protoc Bioinformatics, 2020. **69**(1): p. e96.
22. Gagneux, S., *Ecology and evolution of Mycobacterium tuberculosis.* Nat Rev Microbiol, 2018. **16**(4): p. 202-213.
23. Coscolla, M. and S. Gagneux, *Consequences of genomic diversity in Mycobacterium tuberculosis.* Semin Immunol, 2014. **26**(6): p. 431-44.
24. Holt, K.E., et al., *Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam.* Nat Genet, 2018. **50**(6): p. 849-856.
25. Chiner-Oms, A. and I. Comas, *Large genomics datasets shed light on the evolution of the Mycobacterium tuberculosis complex.* Infect Genet Evol, 2019. **72**: p. 10-15.
26. Freschi, L., et al., *Population structure, biogeography and transmissibility of Mycobacterium tuberculosis.* bioRxiv, 2020: p. 2020.09.29.293274.
27. Borrell, S., et al., *Reference set of Mycobacterium tuberculosis clinical strains: A tool for research and product development.* PLoS One, 2019. **14**(3): p. e0214088.
28. Sharma, R., et al., *High degree of fluoroquinolone resistance among pulmonary tuberculosis patients in New Delhi, India.* Indian J Med Res, 2019. **149**(1): p. 62-66.
29. Ghimire, S., et al., *Treatment outcomes of patients with MDR-TB in Nepal on a current programmatic standardised regimen: retrospective single-centre study.* BMJ Open Respir Res, 2020. **7**(1).

30. Oppong, Y.E.A., et al., *Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance.* BMC Genomics, 2019. **20**(1): p. 252.
31. Ford, C.B., et al., *Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis.* Nat Genet, 2013. **45**(7): p. 784-90.
32. Berrada, Z.L., et al., *Rifabutin and rifampin resistance levels and associated rpoB mutations in clinical isolates of Mycobacterium tuberculosis complex.* Diagn Microbiol Infect Dis, 2016. **85**(2): p. 177-81.
33. Maitre, T., et al., *Are moxifloxacin and levofloxacin equally effective to treat XDR tuberculosis?* J Antimicrob Chemother, 2017. **72**(8): p. 2326-2333.
34. Dijkstra, J.A., et al., *In Vitro Susceptibility of Mycobacterium tuberculosis to Amikacin, Kanamycin, and Capreomycin.* Antimicrob Agents Chemother, 2018. **62**(3).
35. Ho, J., P. Jelfs, and V. Sintchenko, *Fluoroquinolone resistance in non-multidrug-resistant tuberculosis-a surveillance study in New South Wales, Australia, and a review of global resistance rates.* Int J Infect Dis, 2014. **26**: p. 149-53.
36. Miotto, P., et al., *A standardised method for interpreting the association between mutations and phenotypic drug resistance in Mycobacterium tuberculosis.* Eur Respir J, 2017. **50**(6).
37. Timothy M Walker, P.M., Claudio Köser†, Philip Fowler, Jeff Knaggs, Zamin Iqbal, Martin Hunt, Leonid Chindelevitch, Maha Farhat, Daniela Cirillo, Inaki Comas, Jamie Posey, Shaheed Vally Omar, Tim Peto, Anita Suresh, Swapna Uplekar, Sacha Laurant, Rebecca Colman, Carl-Michael Nathanson, Matteo Zignol, Ann Sarah Walker, The CRyPTIC Consortium, The Seq&Treat Consortium, Derrick Crook, Nazir Ismail and Timothy Rodwell, *The 2021 WHO catalogue of Mycobacterium tuberculosis mutations associated with drug resistance: a new global standard for molecular diagnostics (in prep).* 2021.
38. Comas, I., et al., *Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes.* Nat Genet, 2011. **44**(1): p. 106-10.
39. Almeida, D., et al., *Mutations in pepQ Confer Low-Level Resistance to Bedaquiline and Clofazimine in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2016. **60**(8): p. 4590-9.
40. Degiacomi, G., et al., *In vitro Study of Bedaquiline Resistance in Mycobacterium tuberculosis Multi-Drug Resistant Clinical Isolates.* Front Microbiol, 2020. **11**: p. 559469.
41. Hartkoorn, R.C., S. Uplekar, and S.T. Cole, *Cross-resistance between clofazimine and bedaquiline through upregulation of MmpL5 in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2014. **58**(5): p. 2979-81.
42. Nimmo, C., et al., *Population-level emergence of bedaquiline and clofazimine resistance-associated variants among patients with drug-resistant tuberculosis in southern Africa: a phenotypic and phylogenetic analysis.* Lancet Microbe, 2020. **1**(4): p. e165-e174.
43. Singh, R., et al., *PA-824 kills nonreplicating Mycobacterium tuberculosis by intracellular NO release.* Science, 2008. **322**(5906): p. 1392-5.
44. Mvelase, N.R., et al., *Evolving rifampicin and isoniazid mono-resistance in a high multidrug-resistant and extensively drug-resistant tuberculosis region: a retrospective data analysis.* BMJ Open, 2019. **9**(11): p. e031663.

45. Park, S., et al., *Treatment outcomes of rifampin-sparing treatment in patients with pulmonary tuberculosis with rifampin-mono-resistance or rifampin adverse events: A retrospective cohort analysis.* Respir Med, 2017. **131**: p. 43-48.
46. Villegas, L., et al., *Prevalence, Risk Factors, and Treatment Outcomes of Isoniazid- and Rifampicin-Mono-Resistant Pulmonary Tuberculosis in Lima, Peru.* PLoS One, 2016. **11**(4): p. e0152933.
47. Oppong, Y.E.A., et al., *Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance.* BMC Genomics, 2019. **20**(1): p. 252.
48. Pang, Y., et al., *Study of the rifampin monoresistance mechanism in Mycobacterium tuberculosis.* Antimicrob Agents Chemother, 2013. **57**(2): p. 893-900.
49. Kigozi, E., et al., *Prevalence and patterns of rifampicin and isoniazid resistance conferring mutations in Mycobacterium tuberculosis isolates from Uganda.* PLoS One, 2018. **13**(5): p. e0198091.
50. Dean, A.S., et al., *Prevalence and genetic profiles of isoniazid resistance in tuberculosis patients: A multicountry analysis of cross-sectional data.* PLoS Med, 2020. **17**(1): p. e1003008.
51. Migliori, G.B., et al., *Protecting the tuberculosis drug pipeline: stating the case for the rational use of fluoroquinolones.* Eur Respir J, 2012. **40**(4): p. 814-22.
52. Heysell, S.K., et al., *Hearing loss with kanamycin treatment for multidrug-resistant tuberculosis in Bangladesh.* Eur Respir J, 2018. **51**(3).
53. Organisation, W.H., *Update on the use of nucleic acid amplification tests to detect TB and drug-resistant TB: rapid communication*. 2021.
54. Olaru, I.D., et al., *Bedaquiline-based treatment regimen for multidrug-resistant tuberculosis.* Eur Respir J, 2017. **49**(5).
55. Yuan, S., et al., *Clofazimine broadly inhibits coronaviruses including SARS-CoV-2.* Nature, 2021. **593**(7859): p. 418-423.
56. Falzon, D., et al., *World Health Organization treatment guidelines for drug-resistant tuberculosis, 2016 update.* Eur Respir J, 2017. **49**(3).
57. Huitric, E., et al., *Rates and mechanisms of resistance development in Mycobacterium tuberculosis to a novel diarylquinoline ATP synthase inhibitor.* Antimicrob Agents Chemother, 2010. **54**(3): p. 1022-8.
58. Carter J, The CRyPTIC Consortium, *Quantitative measurement of antibiotic resistance in Mycobacterium tuberculosis reveals genetic determinants of resistance and susceptibility in a target gene approach.* 2021, in preparation.
59. Earle S and Wilson D, The CRyPTIC Consortium, *Genome-wide association studies of global Mycobacterium tuberculosis resistance to thirteen antimicrobials in 10,228 genomes.* 2021, in preparation.
60. Lachapelle AS, The CRyPTIC Consortium, *Predicting Susceptibility to First- and Second-line Tuberculosis Drugs by DNA sequencing and Machine Learning.* 2021, in preparation.
61. Sonnenkalb, L., et al., *Deciphering Bedaquiline and Clofazimine Resistance in Tuberculosis: An Evolutionary Medicine Approach.* bioRxiv, 2021: p. 2021.03.19.436148.