Detection of a historic reservoir of bedaquiline / clofazimine resistance associated variants in *Mycobacterium tuberculosis*

Running title: Emergence of bedaquiline resistance in tuberculosis

Camus Nimmo[1,2,3], Arturo Torres Ortiz[1,4], Juanita Pang[1,2], Mislav Acman[1], Cedric C.S. Tan[1], James Millard[3,5,6], Nesri Padayatchi[7], Alison Grant[3,8], Max O'Donnell[7,9], Alex Pym[3], Ola B Brynildsrud[10], Vegard Eldholm[10], Louis Grandjean[2,11,12], Xavier Didelot[13], François Balloux[1*], Lucy van Dorp[1*]

*These authors contributed equally.

Correspondence: Camus Nimmo (camus.nimmo@crick.ac.uk), François Balloux (f.balloux@ucl.ac.uk) and Lucy van Dorp (lucy.dorp.12@ucl.ac.uk)

1. UCL Genetics Institute, University College London, London, UK
2. Division of Infection and Immunity, University College London, London, UK
3. Africa Health Research Institute, Durban, South Africa
4. Department of Medicine, Imperial College, London, UK
5. Wellcome Trust Liverpool Glasgow Centre for Global Health Research, Liverpool, UK
6. Institute of Infection and Global Health, University of Liverpool, Liverpool, UK
7. CAPRISA MRC-HIV-TB Pathogenesis and Treatment Research Unit, Durban, South Africa
8. TB Centre, London School of Hygiene & Tropical Medicine, London, UK
9. Department of Medicine & Epidemiology, Columbia University Irving Medical Center, New York, NY, USA
10. Division of Infectious Diseases and Environmental Health, Norwegian Institute of Public Health, Oslo, Norway
11. Laboratorio de Investigacion y Enfermedades Infecciosas/Universidad Peruana Cayetano Heredia, Lima, Peru
12. Department of Infection, Immunity and Inflammation, Institute of Child Health, University College London, London, UK
13. School of Life Sciences and Department of Statistics, University of Warwick, Coventry, UK

Keywords: Tuberculosis, phylogenetics, bedaquiline, drug resistance, AMR

Abstract

Drug resistance in tuberculosis (TB) poses a major ongoing challenge to public health. The recent inclusion of bedaquiline into TB drug regimens has improved treatment outcomes, but this advance is threatened by the emergence of strains of *Mycobacterium tuberculosis* (*Mtb*) resistant to bedaquiline. Clinical bedaquiline resistance is most frequently conferred by off-target resistance-associated variants (RAVs) in the *mmpR5* gene (*Rv0678*), the regulator of an efflux pump, which can also confer cross-resistance to clofazimine, another TB drug. We compiled a dataset of 3,682 *Mtb* genomes, including 150 carrying variants in *mmpR5* that have been associated to borderline (henceforth intermediate) or confirmed resistance to bedaquiline. We identified eight cases where RAVs were present in the genomes of strains collected prior to the use of bedaquiline in TB treatment regimes. Phylogenetic reconstruction points to multiple emergence events and circulation of RAVs in *mmpR5*, some estimated to predate the introduction of bedaquiline. However, epistatic interactions can complicate bedaquiline drug-susceptibility prediction from genetic sequence data. Indeed, in one clade of isolates where the RAV Ile67fs is estimated to have emerged prior to the antibiotic era, co-occurrence of mutations in *mmpL5* are found to neutralise bedaquiline resistance. The presence of a pre-existing reservoir of *Mtb* strains carrying bedaquiline RAVs prior to its clinical use augments the need for rapid drug susceptibility testing and individualised regimen selection to safeguard the use of bedaquiline in TB care and control.

Introduction

Drug-resistant tuberculosis (DR-TB) currently accounts for 450,000 of the 10 million new tuberculosis (TB) cases reported annually[1]. Treatment outcomes for multidrug-resistant TB (MDR-TB), resistant to at least rifampicin and isoniazid, have historically been poor, with treatment success rates of only 50-60% in routine programmatic settings[2,3]. The discovery of bedaquiline, a diarylquinoline antimycobacterial active against ATP synthase, which is highly effective against *Mycobacterium tuberculosis* (*Mtb*)[4], was reported in 2004. Following clinical trials, which confirmed reduced time to culture conversion in patients with DR-TB[5], in 2012 bedaquiline received an accelerated Food and Drug Administration (FDA) licence for use in DR-TB[6].

Cohort studies of patients treated with bedaquiline-containing regimens against MDR-TB report success rates of 70-80%[7,8]. Similar results have been achieved for extensively drug-resistant TB (XDR-TB, traditionally defined as MDR-TB strains with additional resistance to fluoroquinolones and injectables), where treatment outcomes without bedaquiline are even worse[9,10]. In light of these promising results, the World Health Organization (WHO) now recommends that bedaquiline be included in all MDR-TB regimens[11]. It has played a central role in the highly successful ZeNix[12] and TB-PRACTECAL[13] trials of bedaquiline, pretomanid and linezolid (+/- moxifloxacin) six-month all-oral regimens for DR-TB. These are now incorporated in WHO guidance. In addition, bedaquiline is positioned as a key drug in multiple phase III clinical trials for drug-susceptible TB (SimpliciTB, ClinicalTrials.gov NCT03338621; TRUNCATE-TB[14]).

Resistance in *Mtb* is typically reported shortly after the introduction of a novel TB drug and often appears sequentially[15,16]. For example, mutations conferring resistance to isoniazid – one of the first antimycobacterials – tend to emerge prior to resistance to rifampicin, the other major first-line drug. These also predate resistance mutations to second-line drugs, so termed because they are used clinically to treat patients infected with strains already resistant to first-line drugs. This was observed, for example, in KwaZulu-Natal, South Africa, where resistance-associated mutations accumulated over

76 decades prior to their identification, leading to a major outbreak of extensively drug-resistant TB (XDR-
77 TB)[16]. Unlike other major drug-resistant bacteria, *Mtb* reproduces strictly clonally and systematically
78 acquires resistance by chromosomal mutations rather than via horizontal gene transfer or
79 recombination[17]. This allows phylogenetic reconstructions, based on whole genome sequencing data,
80 to be used to infer the timings of emergence and subsequent spread of variants in *Mtb* that have been
81 suggested to reduce drug susceptibility, termed resistance-associated variants (RAVs).
82
83 A number of mechanisms have been implicated in conferring bedaquiline resistance. For example,
84 mutations conferring resistance have been selected *in vitro*, located in the *atpE* gene encoding the F1F0
85 ATP synthase, the target of bedaquiline[18]. Off-target resistance-conferring mutations have also been
86 found in *pepQ* in a murine model and potentially in a small number of patients[19]. However, the primary
87 mechanism of resistance observed in clinical isolates has been identified in the context of off-target
88 resistance-associated variants (RAVs) in the *mmpR5 (Rv0678)* gene, a negative repressor of expression
89 of the MmpL5 efflux pump. Loss of function of mmpR5 leads to pump overexpression[20] and increased
90 minimum inhibitory concentrations (MIC) to bedaquiline, along with the recently repurposed
91 antimycobacterial clofazimine, fusidic acid, the azole class of antifungal drugs (which also have
92 antimycobacterial activity), as well as to the novel therapeutic class of DprE1 inhibitors in clinical
93 trials[21,22]. Aligned with this mechanism of resistance, coincident mutations leading to loss of function
94 of the MmpL5 efflux pump can negate the resistance-inducing effect of MmpR5 loss of function[23].
95
96 A range of single nucleotide polymorphisms (SNPs) and frameshift *mmpR5* mutations have been
97 associated with resistance to bedaquiline and are often present as heteroresistant alleles in patients[24-34].
98 In contrast to most other RAVs in *Mtb,* which often cause many-fold increases in MIC and clear-cut
99 resistance, *mmpR5* variants may be associated with normal MICs or subtle increases in bedaquiline
100 MIC, although they may still be clinically important[35]. These increases may not cross the current WHO
101 critical concentrations used to classify resistant versus susceptible strains (0.25µg/mL on Middlebrook
102 7H11 agar, or 1µg/mL in Mycobacteria Growth Indicator Tube [MGIT] liquid media). The first version
103 of the WHO tuberculosis drug resistance catalogue does not contain any bedaquiline RAVs, although a

104 subsequent meta-analysis identified two RAVs (a*tpE* Ala63Pro and *mmpR5* Ile67fs)[34]. Bedaquiline has
105 a long terminal half-life of up to 5.5 months[6], leading to the possibility of subtherapeutic concentrations
106 where adherence is suboptimal or treatment is interrupted, which could act as a further driver of
107 resistance.
108
109 Bedaquiline and clofazimine cross-resistance has now been reported across three continents following
110 the rapid expansion in usage of both drugs[25,30,36,37], and is associated in some cases with poor adherence
111 to therapy and inadequate regimens. However, baseline isolates in 8/347 (2.3%) patients from phase IIb
112 bedaquiline trials demonstrated *mmpR5* RAVs and high bedaquiline MICs in the absence of prior
113 documented use of bedaquiline or clofazimine[38]. This suggests that bedaquiline RAVs may have been
114 in circulation prior to the usage of either of these drugs, which may be expected in the case of mutations
115 which do not have major fitness consequences[39]. While there have been isolated clinical reports from
116 multiple geographical regions, the global extent of bedaquiline resistance emergence and spread has
117 not yet been investigated.
118
119 In this study, we characterise and date the emergence of variants in *mmpR5*, including those implicated
120 as bedaquiline RAVs, in the two global *Mtb* lineage 2 (L2) and lineage 4 (L4) lineages, which include
121 the majority of drug resistance strains[15]. Phylogenetic analyses of two datasets comprising 1,514 *Mtb*
122 L2 and 2,168 L4 whole genome sequences revealed the emergence and spread of multiple *mmpR5*
123 variants associated to resistance or borderline (intermediate) resistance to bedaquiline prior to its first
124 clinical use. This pre-existing reservoir of bedaquiline/clofazimine-resistant *Mtb* strains suggests
125 *mmpR5* RAVs exert a relatively low fitness cost which could be rapidly selected for as bedaquiline and
126 clofazimine are more widely used in the treatment of TB.

127 # Results

128 **The global diversity of *Mtb* lineage L2 and L4**

129 To investigate the global distribution of *Mtb* isolates with variants in *mmpR5*, we curated two large
130 datasets of whole genomes from the two dominant global lineages L2 and L4. Both datasets were
131 enriched for samples with variants in *mmpR5* following a screen for variants in public sequencing
132 repositories (see **Methods**) and retaining those samples uploaded with accompanying full metadata for
133 geolocation and time of sampling (**Figure 1, Supplementary Table S1-S2, Supplementary Figure**
134 **S1)**. The final L2 dataset included 1,514 isolates collected over 24.5 years (between 1994 and 2019)
135 yielding 29,205 SNPs. The final L4 dataset comprised 2,168 sequences collected over 232 years,
136 including three samples from 18th century Hungarian mummies[40], encompassing 67,585 SNPs. Both
137 datasets included recently generated data from South Africa (155 L2, 243 L4)[41,42] and new whole
138 genome sequencing data from Peru (9 L2, 154 L4).

139

140 Consistent with previous studies[43-45], both datasets are highly diverse and exhibit strong geographic
141 structure (**Figure 2**). As a nonrecombining clonal organism, identification of mutations in *Mtb* can
142 provide a mechanism to predict phenotypic resistance from a known panel of genotypes[46,47]. Based on
143 genotypic profiling[47], 911 strains within the L2 dataset were classified as MDR-TB (60%) and 295
144 (20%) as XDR-TB. Within the L4 dataset, 911 isolates were classified as MDR-TB (42%) and 115 as
145 XDR-TB (5%). The full phylogenetic distribution of resistance profiles is provided in **Supplementary**
146 **Figure S2**. As is commonplace with genomic datasets, the proportion of drug-resistant strains exceeds
147 their actual prevalence, due to the overrepresentation of drug-resistant isolates in public sequencing
148 repositories.

149

150 Both the L2 and L4 phylogenetic trees displayed a significant temporal signal following date
151 randomisation (**Supplementary Figure S3**), making them suitable for time-calibrated phylogenetic
152 inference[48]. We estimated the time to the Most Recent Common Ancestor (tMRCA) of both datasets
153 using a Bayesian tip-dating analysis (BEAST2) run on a representative subset of genomes from each

154 dataset (see **Methods**, **Supplementary Table 3**, **Supplementary Figure S4**). For the final temporal
155 calibration of the L2 dataset we applied an estimated clock rate of $7.7x10^{-8}$ ($4.9x10^{-8}$ - $1.03x10^{-7}$)
156 substitutions per site per year, obtained from the subsampled BEAST2[48] analysis, to the global
157 maximum likelihood phylogenetic tree. This resulted in an estimated tMRCA of 1332CE (945CE-
158 1503CE). Using the same approach for the L4 dataset we estimated a clock rate of $7.1x10^{-8}$ ($6.2x10^{-8}$ -
159 $7.9x10^{-8}$) substitutions per site per year resulting in an estimated tMRCA of 853CE (685CE – 967CE)
160 (**Figure 2**). We observed a slightly higher, yet statistically not significant, clock rate in L2 compared to
161 L4 (**Supplementary Table S3**), with all estimated substitution rates falling largely in line with
162 previously published estimates[49].
163
164 **Identification of variants in *mmpR5***
165 Since *atpE* and *pepQ* bedaquiline RAVs are found at low prevalence (1 L2 isolate [0.03%] and 18 L4
166 isolates [0.49%]), we focused on characterising mutations in *mmpR5*. In total we identified the presence
167 of non-synonymous and promoter *mmpR5* variants in 437 sequences (193 L2 [12.8%], 244 L4 [11.3%]).
168 We classified all identified non-synonymous and promoter mutations in *mmpR5*, based on an evaluation
169 of their phenotypic impact through review of published literature, into six phenotypic categories for
170 bedaquiline susceptibility: wild type, hypersusceptible, susceptible, intermediate, resistant, and
171 unknown (full references available in **Supplementary Table S4, Supplementary Figures S5-S7)**.
172 Across both lineages, 148 sequences were considered as bedaquiline resistant (i.e., classified as
173 intermediate or resistant). The most frequently observed variants are listed in **Table 1**.
174
175 We identified a significant relationship between the presence of *mmpR5* variants and drug resistance
176 status in both the L2 and L4 datasets (**Supplementary Figure S8-S9**), though in both cases we
177 identified otherwise fully phenotypically susceptible isolates carrying *mmpR5* RAVs. Notably we
178 identified 24 sequenced isolates carrying nonsynonymous/frameshift variants in *mmpR5* uploaded with
179 collection dates prior to the first clinical trials for bedaquiline in 2007. This comprised ten L2 isolates
180 collected before 2007, of which eight harboured variants previously associated to phenotypic
181 bedaquiline resistance (RAVs). For L4 we identified 15 sequences with *mmpR5* variants predating

182 2007, of which six have been previously classified as carrying mutations conferring a bedaquiline
183 resistance phenotype above wild-type ('intermediate') (**Figure 1c-d, Supplementary Table S5**).
184
185 Within the datasets, we identified one L2 isolate (ERR2677436 sampled in Germany in 2016) which
186 already had two *mmpR5* RAVs at low allele frequency – Val7fs (11%) and Val20Phe (20%) – and also
187 contained two low frequency *atpE* RAVs: Glu61Asp (3.2%) and Alal63Pro (3.7%)[50]. We also identified
188 three isolates obtained in 2007-08 from separate but neighbouring Chinese provinces carrying the
189 *Rv1979c* Val52Gly RAV, which has been suggested to be associated with clofazimine resistance in a
190 study from China[25] but was associated with a normal MIC in another[39], with its role in resistance
191 remaining unclear[31]. Furthermore, frameshift and premature stop mutations in *pepQ* have been
192 previously associated with bedaquiline and clofazimine resistance. In this dataset, we identified 18
193 frameshift mutations in *pepQ* across 11 patients, one of which also had a *mmpR5* frameshift mutation.
194 In one isolate the *pepQ* frameshift occurred at the Arg271 position previously reported to be associated
195 with bedaquiline resistance[19].
196
197 Sixty-three genomes harboured nonsynonymous *mmpR5* variants of unknown phenotypic effect (12 L2,
198 28 L4), corresponding to 23 unique mutations or combinations of mutations. To assess properties
199 associated to RAVs which may be useful predictors of the phenotypic effect of these unknown variants
200 we employed a machine learning approach, providing a foundation for further exploration of genomic
201 features associated to RAV status (see **Supplementary Note 1**).
202
203 **The time to emergence of *mmpR5* variants**
204 To estimate the age of the emergence of different *mmpR5* non-synonymous variants, we identified all
205 nodes in each of the L2 and L4 global time calibrated phylogenies delineating clades of isolates carrying
206 a particular *mmpR5* variant (**Figure 3, Supplementary Table S6, Supplementary Table S7**). For the
207 L2 dataset we identified 58 unique phylogenetic nodes where *mmpR5* RAVs emerged, of which 40
208 were represented by a single genome. The point estimates for these nodes ranged from March 1845 to
209 November 2018. Eight nonsynonymous/frameshift variants in *mmpR5*, including four bedaquiline

210 RAVs (Met139Ile, Cys46fs, Ala59Val, Asn98fs) and one case expected to lead to an intermediate
211 phenotype (Arg90Cys), were estimated to have emergence dates (point estimates) predating the first
212 bedaquiline clinical trial in 2007 (**Supplementary Figure S10**).
213
214 For the L4 dataset we identified 85 unique nodes where *mmpR5* RAVs emerged, of which 59 were
215 represented by a single isolate in the dataset. The point estimates for these nodes ranged from September
216 1701 to January 2019 (**Figure 3, Supplementary Figure S11**). Nineteen *mmpR5* mutations, including
217 six unique bedaquiline RAVs (Gln22Arg, Asn98Asp, Ile67fs x2, Arg96Gly, Met146Thr, Asn98Asp)
218 and two predicted to have an intermediate phenotype (Arg90Cys, Ser53Leu), were estimated to have
219 emerged prior to 2007. Arg90Cys, in particular was estimated to emerge between 1930-1947,
220 suggesting the likely circulation of variants which lead to a response to bedaquiline above wild-type
221 pre-existed the first clinical trials for clofazimine in the 1960s. While we identified no nodes with a
222 second emergence of *mmpR5* nonsynonymous/frameshift mutations across the L4 dataset, eight nodes
223 were identified in the L2 dataset where a clade already carrying a nonsynonymous/frameshift variant
224 in *mmpR5* subsequently acquired a second nonsynonymous/frameshift mutation.
225
226 In the L4 dataset, we noted one large clade of 66 samples, predominantly collected in Peru (henceforth
227 Peruvian clade), which all carry the Ile67fs *mmpR5* resistance associated mutation[36,50,51]. While it is not
228 inconceivable that multiple independent emergences of Ile67fs occurred in this clade, the more
229 parsimonious scenario is a single ancestral emergence. We estimate the time of this emergence to 1702
230 (1657-1732) (**Figure 3, Supplementary Figure S11-S12**). Of significance, we identified a frameshift
231 mutation in the adjacent *MmpL5* efflux pump (Arg202fs) in isolates from this Peruvian clade, the
232 protein whose overexpression mediates bedaquiline resistance following loss-of-function of the
233 MmpR5 regulatory protein. This frameshift, which leads to a premature stop codon at amino acid 206,
234 is expected to counteract the otherwise resistance-conferring mutation. This epistatic interaction
235 restoring bedaquiline susceptibility has recently been described elsewhere[23,39]. The *mmpL5* frameshift
236 mutation was present in all isolates in the Peruvian clade bar one (ERR7339051) which had *mmpL5*
237 Arg202Leu. This event of reading-frame restoration is likely explained by a recent secondary

238 duplication of a T downstream of the initial deletion (777876 GGCAT > GGAT, GGAT > GGATT).
239 We considered the phenotype of this strain as unknown. No other *mmpL5* mutations were found in any
240 isolate containing *mmpR5* mutations within this study though we did identify a low prevalence of
241 variants in *mmpL5* and *mmpS5* independent of *mmpR5* mutations across both lineages (**Supplementary**
242 **Figures S13 and S14**).
243
244 We also noted a tendency for *mmpR5* mutations to emerge in clades that also displayed genetic markers
245 of rifampicin resistance. This was more common in mutations emerging after 2007 (77.2%) than before
246 2007 (58.3%). Most of the oldest Ile67fs Peruvian clade was rifampicin resistant (58/66 samples), with
247 the remaining samples demonstrating only isoniazid resistance.
248
249 **Phenotypic validation of *mmpR5* variants**
250 Given documented epistasis as a modulator of bedaquiline resistance phenotype, we performed MIC
251 testing on a selection of available isolates and identified further MICs that have been recently published
252 as part of the Cryptic consortium using microtitre plates (**Supplementary Table S7**)[52,53]. The
253 epidemiological cut-off (ECOFF, defined as MIC of 95-99% of wild-type isolates) for bedaquiline has
254 been proposed to be 0.12 or 0.25 μg/mL depending on the method used, although the final decision was
255 to use an ECOFF of 0.25 μg/mL[53].
256
257 We were able to identify 30 L4 isolates from Peru (including the aforementioned Peruvian clade) for
258 MIC testing, and a further 9 MICs for L4 that had recently been published by the Cryptic consortium[52].
259 For the oldest dated *mmpR5* mutation emergence – the L4 Ile67fs mutation in Peruvian isolates with an
260 associated MRCA estimated to 1701 - 10/11 (90.9%) had an MIC below the lower proposed ECOFF of
261 0.12 μg/mL, presumably due to the co-existing *mmpL5* loss of function mutation. Hence, we denote
262 isolates from this clade as having a hypersusceptible phenotype. The second oldest predicted resistance
263 mutation (Arg90Cys, dated to 1940) was however associated with MICs ≥0.12 μg/mL in 6/7 (85.7%)
264 instances, and in 3/4 (75%) instances for the third oldest predicted resistance associated mutation for

265 which data were available (Asn98Asp, dated to 1987). These MICs are above the wild-type range, if
266 not formally classified as resistant. Clades with associated MIC confirmation are highlighted in Figure
267 3b.
268
269 **Table 1:** Frequency of all *mmpR5* variants occurring ≥ 5 times in dataset and associated resistance
270 classification. *Where co-existing *mmpL5* mutations were identified this is indicated – only one
271 mutation was found (*mmpL5* Arg202fs) and it was in the presence of *mmpR5* Ile67fs mutations only,
272 with no other co-existing *mmpR5* variants.
273
274

Variant	**Associated phenotype**	**L2**	**L4**	**Total**
C-11A	Hypersusceptible	93		93
Ile67fs + mmpL5 Arg202fs	Hypersusceptible		66	66
Asp5Gly	Susceptible	20	3	23
Met146Thr	Resistant	2	20	22
Ile67fs	Resistant	5	17	22
Leu40Val	Susceptible		19	19
Arg90Cys	|Intermediate	2	9	11
Glu49fs	Resistant	2	8	10
Val20Ala	Intermediate	1	6	7
Ala59Val	Resistant	7		7
Val1Ala	Resistant	6		6
Gly121Arg	Resistant	5		6
Asp141fs	Unknown		6	6
Asn98Asp	Resistant		6	6
Arg96Gly	Resistant		5	5
Arg109Leu + Arg156fs	Resistant		5	5

275

Discussion

Our work establishes that the emergence of variants in *mmpR5*, including bedaquiline RAVs, is not solely driven by the use of bedaquiline. We identified up to 11 cases where RAVs have emerged prior to the first clinical trials of bedaquiline in 2007 and a further four cases of variants emerging prior to the clinical use of bedaquiline which are expected to give rise to an intermediate phenotype. These are highlighted red and orange respectively in Supplementary Table S7, not including the oldest emergence of Ile67fs as its resistant phenotype is negated by the epistatic interaction. Phylogenetic inference estimated the oldest clade containing *mmpR5* mutations, composed mostly of samples from Peru carrying the Ile67fs RAV, to have emerged around 1702 (1657-1732). We identify two further early emergences of *mmpR5* mutations, estimated to 1871 and 1940 (Asp141fs and Arg90Cys; point estimates), with samples from the latter clade confirmed to have MICs above the wild-type range justifying classification of an intermediate phenotype. The phenotypic implications of Asp141fs remains unclear. However, together this suggests the likely circulation of variants exhibiting borderline resistance even prior to the first clinical trials for clofazimine. Our phylogenetic inference method, which points to multiple emergences of *mmpR5* nonsynonymous/frameshift variants predating the use of bedaquiline, is also confirmed by the direct observation of eight *Mtb* genomes carrying *mmpR5* RAVs sampled prior to 2007. We also identified, within the aforementioned Peruvian clade, a consistent frameshift mutation in *mmpL5*, which seemed to counteract the resistance phenotypic through an epistatic interaction (MIC <0.12 μg/mL). While Ile67fs is central for bedaquiline resistance in *Mtb*, and this mutation has clearly emerged well prior to the use of bedaquiline and clofazimine in this clade, its phenotypic impact is influenced by the strain genetic background.

We identified other localised clusters with *mmpR5* mutations, reinforcing the need for concern even in situations where such mutations are globally rare. This included 19 isolates with a Met146Thr mutation found in lineage 4 isolates from Eswatini. Met146Thr mutations have been previously associated with a clade that has a rifampicin-resistance conferring mutation located outside of the canonical rifampicin-resistance determining region, and these isolates exhibit elevated bedaquiline MICs[54]. The emergence

303 of the Met146Thr mutation has previously been dated to have emerged in approximately 2003[23,39,54].
304 This is in reasonable agreement with our analysis on a much larger dataset which inferred an emergence
305 in 2005.6 (95% confidence intervals 2004.8 – 2006.0). The long-standing presence of variants
306 implicated in resistance and borderline resistance to bedaquiline predating the use of the drug and at
307 high prevalence in geographically notable cases is of concern, as it suggests that non-synonymous
308 mutations in *mmpR5* exert little fitness cost.
309
310 Together, our work suggests the existence of pre-existent reservoirs of bedaquiline resistant *Mtb*. These
311 may have been selected for through historic clofazimine use, though we note at least one case of
312 intermediate resistance to bedaquiline emerging as early as 1930-1947. We also note that detected
313 variants in mmpR5 tend to exist in strains already displaying rifampicin-resistance, although also
314 include otherwise fully susceptible strains (**Supplementary Table S7**). Together this suggests the
315 important role of prior drug exposure in selecting for strains with pre-existing (cross-)resistance
316 potential. This reservoir of putatively adaptive variants is expected to expand under drug pressure with
317 the increasing use of bedaquiline and clofazimine in TB treatment. Further, these reservoirs may also
318 pose a threat for other candidate TB agents from different drug classes that are also exported by *mmpS5*
319 and *mmpL5*[19,22,55].
320
321 The identification of resistance variants occurring before the clinical use of a drug is not limited to *M.*
322 *tuberculosi*s and *mmpR5* alone. To illustrate, within *M. bovis*, there is evidence indicating that the *pncA*
323 H57D mutation, which is associated with resistance to pyrazinamide (PZA), emerged approximately
324 900 years ago, providing inherent resistance to PZA in *M. bovis*[56]. Similarly, variations in intrinsic
325 susceptibility to pretomanid have been observed across the MTBC, including *Mtb* lineages, even
326 without prior exposure to nitroimidazoles[57]. It is likely that there are numerous other instances of such
327 loss of function mutations with minimal or no impact on fitness, similar to the case of *mmpR5*.
328 Furthermore, the existence of antimicrobial resistance (AMR) in different forms has persisted
329 throughout the natural history of various bacteria[58].
330

331 Nevertheless, it is crucial to determine the age and diversity of variants that have been implicated in
332 drug resistance to gain a better understanding of the potential for widespread resistance as a
333 contemporary challenge. We identified a large number of different *mmpR5* nonsynonymous/frameshift
334 variants across both of our *Mtb* lineage cohorts; 46 in L2 and 67 in L4. This suggests the mutational
335 target leading to bedaquiline resistance is wider than for most other current TB drugs and raises
336 concerns about the ease with which bedaquiline resistance can emerge during treatment. It is further
337 concerning that resistance to the new class of nitroimidazole drugs, such as pretomanid and delamanid,
338 is also conferred by loss of function mutations in any of at least six genes, suggesting that they may
339 also have a low barrier to resistance[59].
340
341 While we identified many non-synonymous variants in *mmpR5*, only one (Ile67fs) has been previously
342 definitively linked to resistance. We acknowledge that several of our detected variants have no
343 associated MIC values available in the literature and are thus currently not fully phenotypically
344 validated. We hope by presenting these as 'unknown' our work, estimating the age of emergence of
345 non-synonymous mutations, can be of value as further variants are phenotyped in the future. It is
346 however true that determining the phenotypic consequences of *mmpR5* variants that have previously
347 been described is challenging as there are often only limited reports correlating MICs to genotypes.
348 Moreover, at least four different methods are used to determine MICs, some of which do not have
349 associated critical concentrations. Even where critical concentrations have been set, there is an overlap
350 in MICs of isolates that are genetically wild type and those that have mutations likely to cause
351 resistance[35]. We also note that as we purposefully enriched our dataset for *mmpR5* mutations, the
352 sampling precludes estimation of the overall prevalence of these mutations in genome sequencing
353 databases.
354
355 Prediction of phenotypic bedaquiline resistance from genomic data is further complicated by the
356 existence of hypersusceptibility variants. For example, the c-11a variant located in the promoter of
357 mmpR5, which appears to increase susceptibly to bedaquiline[38], was observed to be fixed throughout a
358 large clade within L2. The early emergence of this variant and its geographical concentration in South

359 Africa and Eswatini may suggest the role of non-pharmacological influences on mmpR5 which
360 regulates multiple MmpL efflux systems[20]. Further, analysis of hypersusceptibility is limited by the
361 truncated lower MIC range of the UKMYC microtitre plates, with many isolates giving MICs below
362 the lower end of the measured range. While large-scale genotype/phenotype analyses will likely support
363 the development of rapid molecular diagnostics, targeted or whole genome sequencing, at reasonable
364 depths, may provide the only opportunity to detect all possible *mmpR5* RAVs, and possible co-
365 occurring mutations, in clinical settings.
366
367 Bedaquiline resistance can also be conferred by other RAVs including in *pepQ* (bedaquiline and
368 clofazimine), *atpE* (bedaquiline only)[51] and *Rv1979c* (clofazimine only). We only found *atpE* RAVs at
369 low allele frequency in one patient who also had *mmpR5* variants (sample accession ERR2677436),
370 which is in line with other evidence suggesting they rarely occur in clinical isolates, likely due to a high
371 fitness cost. Likewise, we only identified *Rv1979c* RAVs in three patients in China, although there were
372 other variants in *Rv1979c* for which ability to cause phenotypic resistance has not been previously
373 assessed. Frameshift *pepQ* mutations that are potentially causative of resistance were identified in 11
374 cases, in keeping with its possible role as an additional rare resistance mechanism.
375
376 Our findings, of reservoirs of *mmpR5* RAVs predating the therapeutic use of bedaquiline, are of high
377 clinical relevance as the presence of *mmpR5* variants during therapy in clinical strains has been
378 associated with substantially worse outcomes in patients treated with drug regimens including
379 bedaquiline[36]. Although it is uncertain what the impact of *mmpR5* RAVs are on outcomes when present
380 prior to treatment[60,61], it is imperative to monitor and prevent the wider transmission of bedaquiline
381 resistant clones, particularly in high MDR/XDR-TB settings. Early evaluation of new TB drug
382 candidates entering clinical trials will also be vital given early data suggesting possible cross-resistance
383 for DprE1 inhibitors such as macozinone[22]. The large and disparate set of mutations in *mmpR5* we
384 identified, with differing phenotypes and some having been in circulation historically, adds further
385 urgency to the development of rapid drug susceptibility testing for bedaquiline to inform effective
386 treatment choices and mitigate the further spread of DR-TB.

387 # Materials and methods

388

389 **Sample collection**

390 In this study we curated large representative datasets of *Mtb* whole genome sequences encompassing
391 the global genetic and geographic distribution of lineages 2 (L2) and L4 (**Figure 1, Supplementary**
392 **Tables S1-S2**). The dataset was enriched to include all available sequenced isolates with *mmpR5*
393 variants, which in some cases included isolates with no, or limited, published metadata. In all other
394 cases samples for which metadata on the geographic location and date of collection was available were
395 retained. To ensure high quality consensus alignments we required that all samples mapped with a
396 minimum percentage cover of 96% and a mean coverage of 30x to the H37Rv reference genome
397 (NC_000962.3). We excluded any samples with evidence of mixed strain infection as identified by the
398 presence of lineage-specific SNPs to more than one sublineage[62] or the presence of a high proportion
399 of heterozygous alleles[63]. The total number of samples included in these datasets, and their source is
400 shown in **Supplementary Table S2.** An index of all samples is available in **Supplementary Table S1**.

401

402 A large global dataset of 1,669 L4 *Mtb* sequences has been constructed, which we used as the basis for
403 curating our L4 dataset[44]. We refer to this as the 'base dataset' for L4. For L2, we constructed a 'base
404 dataset' by screening the Sequence Read Archive (SRA) and European Nucleotide Archive (ENA)
405 using BIGSI[64] for the *rpsA* gene sequence containing the L2 defining variant *rpsA* a636c[62] with a 100%
406 match. This search returned 6,307 *Mtb* genomes, of which 1,272 represented unique samples that had
407 the minimum required metadata. Metadata from three studies were also added manually as they were
408 not included in their respective SRA submissions but were available within published studies[65-67].

409

410 For isolates with only information on the year of sample collection, we set the date to be equal to the
411 middle of the year. For those with information on the month but not the date of collection we set the
412 date of collection to the first of the month. For sequenced samples which were missing associated
413 metadata (32 L2 genomes and 19 L4 genomes) we attempted to estimate an average time of sample

414 collection in order to impute a sampling date. To do so we computed the average time between date of
415 collection and sequence upload date for all samples with associated dates separately in each of the L2
416 and L4 datasets (**Supplementary Figure S1**). For L2 we estimated a mean lag time of 4.7 years (0.5–
417 12.6 years 95% CI). For L4, having excluded three sequences obtained from 18th Century mummies
418 from Hungary[40], we estimated a mean lag time of 6.9 years (0.6-19.1 years 95% CI). The estimated
419 dates, where required, are provided in Supplementary Table S1.
420
421 To enrich the datasets for isolates with *mmpR5* variants, we included further sequences from our own
422 studies in KwaZulu-Natal, South Africa[41,42], other studies of drug-resistant TB in southern Africa[16,44,68-
423 71], and Peru[72,73]. We additionally supplement the Peruvian data with 163 previously unpublished
424 isolates. In these cases, and to facilitate the most accurate possible estimation of the date of resistance
425 emergence, we included samples with *mmpR5* variants as well as genetically related sequences without
426 *mmpR5* variants.
427
428 To identify further published raw sequencing data with *mmpR5* variants from studies where
429 bedaquiline/clofazimine resistance may have been previously unidentified, we screened the NCBI
430 Sequencing Read Archive (SRA) for sequence data containing 85 previously published *mmpR5*
431 variants[28-30,41,42,74,75] with BIGSI[64]. BIGSI was employed against a publicly available indexed database
432 of complete SRA/ENA bacterial and viral whole genome sequences current to December 2016
433 (available here: http://ftp.ebi.ac.uk/pub/software/bigsi/nat_biotech_2018/all-microbial-index-v03/),
434 and also employed locally against an updated in-house database which additionally indexed SRA
435 samples from January 2017 until January 2019. Samples added using this approach are flagged 'BIGSI'
436 in **Supplementary Table S1**. We also used the PYGSI tool[76] to interrogate BIGSI with the *mmpR5*
437 sequence adjusted to include every possible single nucleotide substitution. In each instance we included
438 30 bases upstream and downstream of the gene as annotated on the H37Rv *Mtb* reference genome. For
439 the purpose of this study we only considered coding region, non-synonymous substitutions and
440 insertions and deletions. Samples added following the PYGSI screen are flagged 'PYGSI' in

441 **Supplementary Table S1**. A breakdown of the different datasets used is provided in **Supplementary**
442 **Table S2**.
443
444 **Reference mapping and variant calling**
445 Original fastq files for all included sequences were downloaded and paired reads mapped to the H37Rv
446 reference genome with bwa mem v0.7.17[77]. Mapped reads were sorted and de-duplicated using Picard
447 Tools v2.20 followed by indel realignment with GATK v3.8[78]. Alignment quality and coverage was
448 recorded with Qualimap v2.21[79]. Variant calling was performed using bcftools v1.9, based on reads
449 mapping with a minimum mapping quality of 20, base quality of 20, no evidence of strand or position
450 bias, a minimum coverage depth of 10 reads, and a minimum of four reads supporting the alternate
451 allele, with at least two of them on each strand. Moreover, SNPs that were less than 2bp away from an
452 indel were excluded from the analysis. Similarly, only indels 3bp apart of other indels were kept.
453
454 All sites with insufficient coverage to identify a site as variant or reference were excluded (marked as
455 'N'), as were those in or within 100 bases of PE/PPE genes, or in insertion sequences or phages. SNPs
456 present in the alignment with at least 90% frequency were used to generate a pseudoalignment of equal
457 length to the H37Rv. Samples with more than 10% of the alignment represented by ambiguous bases
458 were excluded. Those positions with more than 10% of ambiguous bases across all the samples were
459 also removed. In order to avoid bias on the tree structure, positions known to be associated with drug
460 resistance were not included.
461
462 A more permissive variant calling pipeline was used to identify *mmpR5* variants, as they are often
463 present at <100% frequency with a high incidence of frameshift mutations. Here we instead employed
464 FreeBayes v1.2[85] to call all variants present in the *mmpR5* gene (or up to 100 bases upstream) that were
465 present at ≥5% frequency (alternate allele fraction *-F* 0.05) and supported by at least four reads
466 including one on each strand. Using this more permissive variant calling strategy we also systematically

467 screened for all mutations in the efflux pump proteins mmpS5-mmpL5 operon (**Supplementary**
468 **Figures S13 and S14**).
469
470 **Classification of resistance variants**
471 All raw fastq files were screened using the rapid resistance profiling tool TBProfiler[47,80] against a
472 curated whole genome drug resistance mutations library. This allowed rapid assignment of
473 polymorphisms associated with resistance to different antimycobacterial drugs and categorisation of
474 MDR and XDR *Mtb* status (**Supplementary Figure S2, Supplementary Figures S5-S9**). Resistance
475 profiles of sequences containing *mmpR5* variants are listed in Supplementary Table S7 as either "S" for
476 susceptible, "RR" for rifampicin-resistant and "preXDR" for fluoroquinolone-resistant.
477
478 **Classification of *mmpR5* variants**
479 The diverse range of *mmpR5* variants and paucity of widespread MIC testing means that there are
480 limited data from which to infer the phenotypic consequences of identified *mmpR5* variants. This was
481 true aside from a subset of data sampled in Peru for which 30 L4 isolates from Peru were subjected to
482 MIC testing using the UKMYC6 plate and a further nine were evaluated for MICs reported by the
483 Cryptic consortium[52]. The approach we used was to assign whether nonsynonymous variants confer a
484 normal or raised MIC based on published phenotypic tests for strains carrying that variant. A full list
485 of the literature reports used for each mutation is provided in **Supplementary Table S4**. We also
486 introduced an intermediate category to describe isolates with MICs at the critical concentration (e.g.,
487 0.25µg/mL on Middlebrook 7H11 agar), where there is an overlap of the MIC distributions of *mmpR58*
488 mutated and wild type isolates with uncertain clinical implications[35]. We assumed that all other
489 disruptive frameshift and stop mutations would confer resistance in light of the role of *mmpR5* as a
490 negative repressor, where loss of function should lead to efflux pump overexpression, unless evidence
491 exists in the literature to suggest otherwise. This allowed us to identify two frameshifts of currently
492 unclear effect (**Supplementary Table S4**). All other promoter and previously unreported missense
493 mutations were categorised as unknown (**Supplementary Table S4**). Where *mmpR5* mutations were

494 accompanied by an *mmpS5* or *mmpL5* loss of function mutation, we assumed that would confer
495 susceptibility (or hypersusceptibility) to bedaquiline[23].
496
497 **Global phylogenetic inference**
498 The alignments for phylogenetic inference were masked for the *mmpR5* region using bedtools v2.25.0.
499 All variant positions were extracted from the resulting global phylogenetic alignments using snp-sites
500 v2.4.1[81], including a L4 outgroup for the L2 alignment (NC_000962.3) and a lineage 3 (L3) outgroup
501 for the L4 alignment (SRR1188186). This resulted in a 67,585 SNP alignment for the L4 dataset and
502 29,205 SNP alignment for the L2 dataset. A maximum likelihood phylogenetic tree was constructed for
503 both SNP alignments using RAxML-NG v0.9.0[82] specifying a GTR+G substitution model, correcting
504 for the number of invariant sites using the ascertainment flag (ASC_STAM) and specifying a minimum
505 branch length of $1x10^{-9}$ reporting 12 decimal places (--precision 12).
506
507 **Estimating the age of emergence of *mmpR5* variants**
508 To test whether the resulting phylogenies can be time-calibrated we first dropped the outgroups from
509 the phylogeny and rescaled the trees so that branches were measured in unit of substitutions per genome.
510 We then computed a linear regression between root-to-tip distance and the time of sample collection
511 using BactDating[83], which additionally assesses the significance of the regression based on 10,000 date
512 randomisations. We obtained a significant temporal correlation for both the L2 and L4 phylogenies,
513 both with and without imputation of dates for samples with missing metadata (**Supplementary Figure**
514 **3**).
515
516 We employed the Bayesian method BactDating v1.01[83], run without updating the root (updateRoot=F),
517 a mixed relaxed gamma clock model and otherwise default parameters to both global datasets. The
518 MCMC chain was run for $1x10^7$ iterations and $3x10^7$ iterations. BactDating results were considered
519 only when MCMC chains converged with an Effective Sample Space (ESS) of at least 100. The analysis
520 was applied to the datasets both with and without considering imputed and non-imputed collection dates
521 (**Supplementary Table 3**).

522

523 To independently infer the evolutionary rates associated with each of our datasets, we sub-sampled both
524 the L4 and L2 datasets to 200 isolates, selected so as to retain the maximal diversity of the tree using
525 Treemmer v0.3[84]. As before, we excluded all variants currently implicated in drug resistance from the
526 alignments. This resulted in a dataset for L4 comprising 25,104 SNPs and spanning 232 years of
527 sampling and for L2 comprising 8,221 SNPs and spanning 24 years of sampling. In both cases the L3
528 sample SRR1188186 was used as an out-group given this has an associated collection date. Maximum
529 likelihood trees were constructed using RaXML-NG v0.9.0[82], as previously described, and a significant
530 temporal regression was obtained for both sub-sampled datasets (**Supplementary Figure S4**).

531

532 BEAST2 v2.6.0[48] was run on both subsampled SNP alignments allowing for model averaging over
533 possible choices of substitution models[85]. All models were run with either a relaxed or a strict prior on
534 the evolutionary clock rate for three possible coalescent demographic models: exponential, constant
535 and skyline. To speed up the convergence, the prior on the evolutionary clock rate was given as a
536 uniform distribution (limits 0 to 10) with a starting value set to 10^{-7}. In each case, the MCMC chain was
537 run for 500,000,000 iterations, with the first 10% discarded as burn-in and sampling trees every 10,000
538 chains. The convergence of the chain was inspected in Tracer 1.7 and through consideration of the ESS
539 for all parameters (ESS>200). The best-fit model to the data for these runs was assessed through a path
540 sampling analysis[86] specifying 100 steps, 4 million generations per step, alpha = 0.3, pre-burn-in = 1
541 million generations, burn-in for each step = 40%. For both datasets, the best supported strict clock
542 model was a coalescent Bayesian skyline analysis. The rates (mean and 95% HPD) estimated under
543 these subsampled analyses (L2 7.7×10^{-8} [4.9×10^{-8} - 1.03×10^{-7}] substitutions per site per year; L4 $7.1\times10^{-}$
544 ^{8} [6.2×10^{-8} - 7.9×10^{-8}] substitutions per site per year) were used to rescale the maximum likelihood
545 phylogenetic trees generated across the entire L2 and L4 datasets, by transforming all branch lengths of
546 the tree from per unit substitution to per unit substitutions per site per year using the R package Ape
547 v5.3[87]. This resulted in an estimated tMRCA of 1332CE (945CE-1503CE) for L2 and 853CE (685CE
548 – 967CE) for L4 (**Figure 2**).

549

550 The resulting phylogenetic trees were visualised and annotated for place of geographic sampling and
551 *mmpR5* variant status using ggtree v1.14.6[88]. All nonsynonymous/frameshift mutations in *mmpR5* were
552 considered, with the phenotypic status assigned in **Supplementary Table S4**. For the purpose of this
553 analysis, and to be conservative, 'unknown' variants classified using XGBoost were still considered
554 'unknown' (**Supplementary Note 1**). Clades carrying shared variants in *mmpR5* were identified and
555 the distributions around the age of the node (point estimates – mean - and 95% HPDs) were extracted
556 from the time-stamped phylogeny. For isolated samples (single emergences) exhibiting variants in
557 *mmpR5*, the time of sample collection was extracted together with the date associated with the upper
558 bound on the age of the next closest node of the tree, allowing for the mutation to have occurred
559 anywhere over the length of the terminal branch (**Figure 3, Supplementary Figures S11-S12**). For the
560 Peruvian clade Bayesian skyline analysis was implemented through the skylineplot analysis
561 functionality available in Ape v5.3[87].
562
563 **Data availability**
564 Raw sequence data and full metadata for all newly generated isolates are available on NCBI Sequencing
565 Read Archive under BioProject ID: PRJEB39837.
566

567 # Footnotes

568
569 **Author Contributions**
570 LvD, CN and FB conceived and designed the study. JM, NP, AG, MO, AP, OBB, VE and LG provided
571 sequence data. ATO, JP, MA, CCST and XD performed and advised on computational analyses. LvD,
572 CN and FB wrote the manuscript with input from all co-authors. All authors read and approved the final
573 manuscript.
574
575 **Acknowledgments**
576 CN and JM are supported by the Wellcome Trust (203583/Z/16/Z and 203919/Z/16/Z, respectively).
577 LvD is supported by a UCL Excellence Fellowship. F.B. acknowledges support from the BBSRC

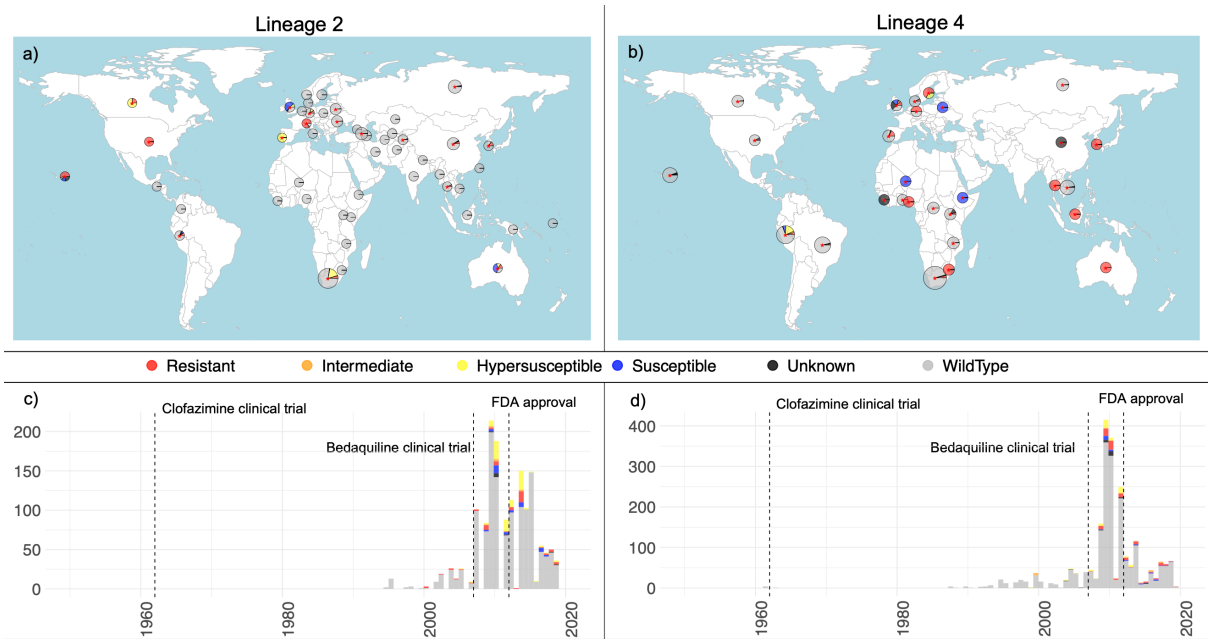
(equipment grant BB/R01356X/1). FB additionally acknowledges the National Institute for Health Research University College London Hospitals Biomedical Research Centre. M.A. was supported by a Ph.D. scholarship from University College London. All authors acknowledge UCL Biosciences Big Data equipment grant from BBSRC (BB/R01356X/1).

Competing interests

The authors declare no competing financial interests. AP is currently employed by Janssen. Dr Pym's involvement with the research described herein precedes his employment at Janssen.

586 # Figures

587

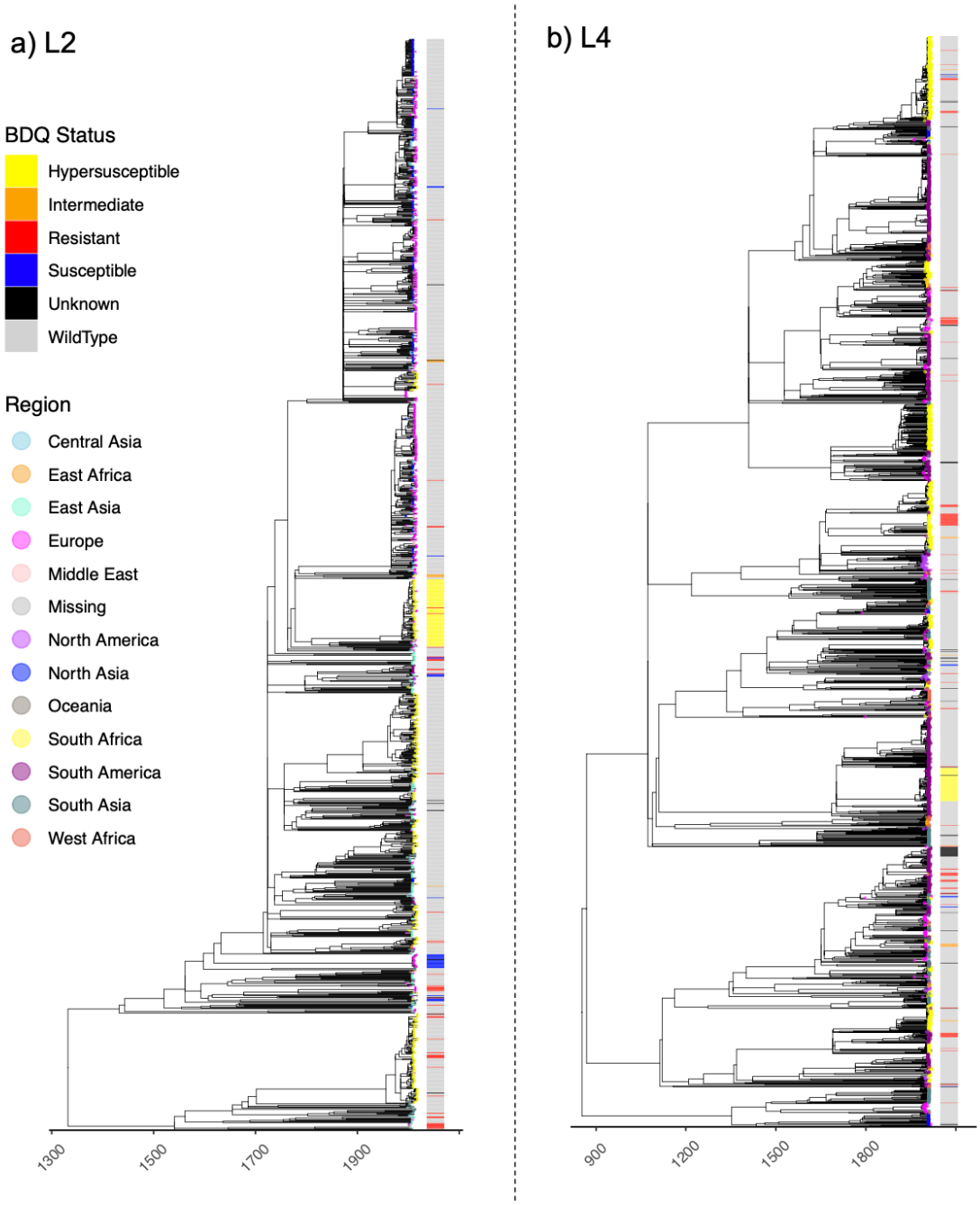


588
589
590 **Figure 1: Compiled global *Mtb* genomic datasets.**

591 Panels a) and b) provide the geographic location of isolates included in the lineage 2 and lineage 4
592 datasets respectively. Pies are scaled by the number of samples per country (raw data available in
593 **Supplementary Table S1**) with the colours providing the fraction of genomes with any
594 nonsynonymous/frameshift variants detected in *mmpR5* (coloured as per the legend). Countries
595 comprising samples with known RAVs are highlighted with a red asterisk. Genomic data for which no
596 associated metadata on the geographic location of sampling was available are shown in the Pacific
597 Ocean. Panels c) and d) provide the collection dates associated to each genome in the lineage 2 and
598 lineage 4 datasets respectively highlighting those with any variants in *mmpR5* (colour, as per legend).
599 Lineage 4 *Mtb* obtained from 18th century mummies are excluded from this plot but included in all
600 analyses. The vertical dashed lines indicate the dates of the first clinical trials for clofazimine,
601 bedaquiline and FDA approval of bedaquiline for clinical use.

602

603



604

605 **Figure 2: Global time calibrated *Mtb* phylogenies.**

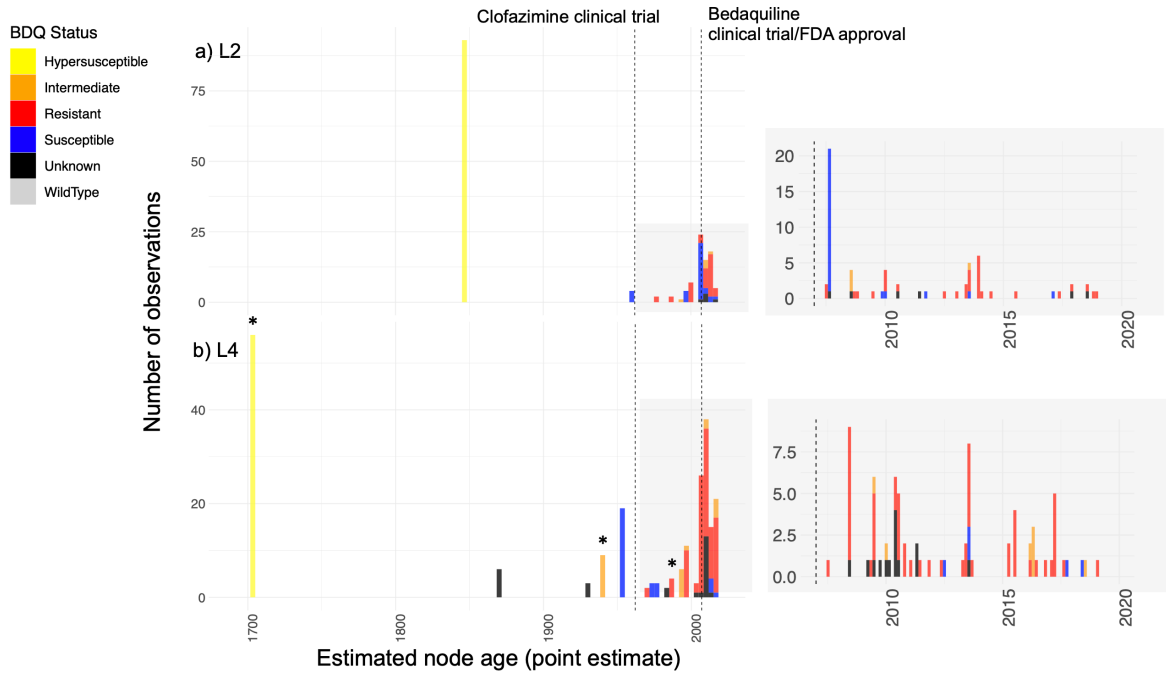
606 Inferred dated phylogenies (x-axis) for the a) lineage 2 and b) lineage 4 datasets. Tips are coloured by

607 the geographic region of sampling as given in the legend. The bar provides the assessed phenotype

608 (colour) based on assignment of nonsynonymous/frameshift variants in *mmpR5*.

609

610



611

612 **Figure 3: Estimated age of emergence of *mmpR5* nonsynonymous/frameshift variants.**

613 Inferred point estimates for the age of emergence of clades with *mmpR5* variants for the lineage 2 (a)

614 and lineage 4 (b) datasets, including a zoomed in reproduction of the period from 2007-2020. Y-axis

615 provides the absolute number of sequences descending from the identified and dated nodes. The *mmpR5*

616 RAV status is given by the colour as defined in the legend at bottom. *indicates phenotypic data

617 available for considered isolates that are supportive of MIC classification (see text).The full mutation

618 timelines are provided in **Supplementary Figures 11-12** and **Supplementary Table S7**.

619

620 **References**

621 1 Organization, W. H. Global Tuberculosis Report 2022. (WHO, 2022).
622 2 Cegielski, J. P. *et al.* Multidrug-Resistant Tuberculosis Treatment Outcomes in
623 Relation to Treatment and Initial Versus Acquired Second-Line Drug Resistance. *Clin*
624 *Infect Dis* **62**, 418-430 (2016). https://doi.org:10.1093/cid/civ910
625 3 Organization, W. H. Global Tuberculosis Report 2019. (2019).
626 4 Andries, K. *et al.* A diarylquinoline drug active on the ATP synthase of *Mycobacterium*
627 *tuberculosis*. *Science* **307**, 223-227 (2005). https://doi.org:10.1126/science.1106753
628 5 Diacon, A. H. *et al.* Multidrug-resistant tuberculosis and culture conversion with
629 bedaquiline. *N Engl J Med* **371**, 723-732 (2014).
630 https://doi.org:10.1056/NEJMoa1313865
631 6 Spring, S. Sirturo (bedaquiline).
632 7 Borisov, S. E. *et al.* Effectiveness and safety of bedaquiline-containing regimens in the
633 treatment of MDR- and XDR-TB: a multicentre study. *Eur Respir J* **49** (2017).
634 https://doi.org:10.1183/13993003.00387-2017
635 8 Guglielmetti, L. *et al.* Long-term outcome and safety of prolonged bedaquiline
636 treatment for multidrug-resistant tuberculosis. *Eur Respir J* **49** (2017).
637 https://doi.org:10.1183/13993003.01799-2016
638 9 Olayanju, O. *et al.* Long-term bedaquiline-related treatment outcomes in patients with
639 extensively drug-resistant tuberculosis from South Africa. *Eur Respir J* **51** (2018).
640 https://doi.org:10.1183/13993003.00544-2018
641 10 Ndjeka, N. *et al.* High treatment success rate for multidrug-resistant and extensively
642 drug-resistant tuberculosis using a bedaquiline-containing treatment regimen. *Eur*
643 *Respir J* **52** (2018). https://doi.org:10.1183/13993003.01528-2018
644 11 Organization, W. H. *Module 4: treatment - drug-resistant tuberculosis treatment, 2022*
645 *update*. (2022).
646 12 Conradie, F. *et al.* Bedaquiline-Pretomanid-Linezolid Regimens for Drug-Resistant
647 Tuberculosis. *N Engl J Med* **387**, 810-823 (2022).
648 https://doi.org:10.1056/NEJMoa2119430
649 13 Berry, C. *et al.* TB-PRACTECAL: study protocol for a randomised, controlled, open-
650 label, phase II-III trial to evaluate the safety and efficacy of regimens containing
651 bedaquiline and pretomanid for the treatment of adult patients with pulmonary
652 multidrug-resistant tuberculosis. *Trials* **23**, 484 (2022). https://doi.org:10.1186/s13063-
653 022-06331-8
654 14 Paton, N. I., Cousins, C. & Suresh, C. Treatment Strategy for Rifampin-Susceptible
655 Tuberculosis. Reply. *N Engl J Med* **388**, 2298 (2023).
656 https://doi.org:10.1056/NEJMc2304776
657 15 Manson, A. L. *et al.* Genomic analysis of globally diverse *Mycobacterium tuberculosis*
658 strains provides insights into the emergence and spread of multidrug resistance. *Nat*
659 *Genet* **49**, 395-402 (2017). https://doi.org:10.1038/ng.3767
660 16 Cohen, K. A. *et al.* Evolution of Extensively Drug-Resistant Tuberculosis over Four
661 Decades: Whole Genome Sequencing and Dating Analysis of *Mycobacterium*
662 *tuberculosis* Isolates from KwaZulu-Natal. *PLoS Med* **12**, e1001880 (2015).
663 https://doi.org:10.1371/journal.pmed.1001880
664 17 Eldholm, V. & Balloux, F. Antimicrobial Resistance in *Mycobacterium tuberculosis:*
665 The Odd One Out. *Trends Microbiol* **24**, 637-648 (2016).
666 https://doi.org:10.1016/j.tim.2016.03.007

18 Huitric, E. *et al.* Rates and mechanisms of resistance development in *Mycobacterium tuberculosis* to a novel diarylquinoline ATP synthase inhibitor. *Antimicrob Agents Chemother* **54**, 1022-1028 (2010). https://doi.org:10.1128/AAC.01611-09

19 Almeida, D. *et al.* Mutations in pepQ Confer Low-Level Resistance to Bedaquiline and Clofazimine in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **60**, 4590-4599 (2016). https://doi.org:10.1128/AAC.00753-16

20 Andries, K. *et al.* Acquired resistance of *Mycobacterium tuberculosis* to bedaquiline. *PLoS One* **9**, e102135 (2014). https://doi.org:10.1371/journal.pone.0102135

21 Hartkoorn, R. C., Uplekar, S. & Cole, S. T. Cross-resistance between clofazimine and bedaquiline through upregulation of *MmpL5* in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **58**, 2979-2981 (2014). https://doi.org:10.1128/AAC.00037-14

22 Poulton, N. C., Azadian, Z. A., DeJesus, M. A. & Rock, J. M. Mutations in rv0678 Confer Low-Level Resistance to Benzothiazinone *DprE1* Inhibitors in *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **66**, e0090422 (2022). https://doi.org:10.1128/aac.00904-22

23 Vargas, R., Jr. *et al.* Role of Epistasis in Amikacin, Kanamycin, Bedaquiline, and Clofazimine Resistance in *Mycobacterium tuberculosis* Complex. *Antimicrob Agents Chemother* **65**, e0116421 (2021). https://doi.org:10.1128/AAC.01164-21

24 Bloemberg, G. V. *et al.* Acquired Resistance to Bedaquiline and Delamanid in Therapy for Tuberculosis. *N Engl J Med* **373**, 1986-1988 (2015). https://doi.org:10.1056/NEJMc1505196

25 Xu, J. *et al.* Primary Clofazimine and Bedaquiline Resistance among Isolates from Patients with Multidrug-Resistant Tuberculosis. *Antimicrob Agents Chemother* **61** (2017). https://doi.org:10.1128/AAC.00239-17

26 Zimenkov, D. V. *et al.* Examination of bedaquiline- and linezolid-resistant *Mycobacterium tuberculosis* isolates from the Moscow region. *J Antimicrob Chemother* **72**, 1901-1906 (2017). https://doi.org:10.1093/jac/dkx094

27 de Vos, M. *et al.* Bedaquiline Microheteroresistance after Cessation of Tuberculosis Treatment. *N Engl J Med* **380**, 2178-2180 (2019). https://doi.org:10.1056/NEJMc1815121

28 Ghodousi, A. *et al.* Acquisition of Cross-Resistance to Bedaquiline and Clofazimine following Treatment for Tuberculosis in Pakistan. *Antimicrob Agents Chemother* **63** (2019). https://doi.org:10.1128/AAC.00915-19

29 Polsfuss, S. *et al.* Emergence of Low-level Delamanid and Bedaquiline Resistance During Extremely Drug-resistant Tuberculosis Treatment. *Clin Infect Dis* **69**, 1229-1231 (2019). https://doi.org:10.1093/cid/ciz074

30 Mokrousov, I., Akhmedova, G., Polev, D., Molchanov, V. & Vyazovaya, A. Acquisition of bedaquiline resistance by extensively drug-resistant *Mycobacterium tuberculosis* strain of Central Asian Outbreak clade. *Clin Microbiol Infect* **25**, 1295-1297 (2019). https://doi.org:10.1016/j.cmi.2019.06.014

31 Kadura, S. *et al.* Systematic review of mutations associated with resistance to the new and repurposed *Mycobacterium tuberculosis* drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid. *J Antimicrob Chemother* **75**, 2031-2043 (2020). https://doi.org:10.1093/jac/dkaa136

32 Roberts, L. W. *et al.* Repeated evolution of bedaquiline resistance in *Mycobacterium tuberculosis* is driven by truncation of *mmpR5*. *bioRxiv*, 2022.2012.2008.519610 (2022). https://doi.org:10.1101/2022.12.08.519610

33 Sonnenkalb, L. *et al.* Bedaquiline and clofazimine resistance in *Mycobacterium tuberculosis:* an in-vitro and in-silico data analysis. *Lancet Microbe* **4**, e358-e368 (2023). https://doi.org:10.1016/S2666-5247(23)00002-2

34 Ismail, N. *et al.* Genetic variants and their association with phenotypic resistance to bedaquiline in *Mycobacterium tuberculosis*: a systematic review and individual isolate data analysis. *Lancet Microbe* **2**, E604-E616 (2021). https://doi.org:10.1016/S2666-5247(21)00175-0

35 Organization, W. H. Technical report on critical concentrations for TB drug susceptibility testing of medicines used in the treatment of drug-resistant TB., (2018).

36 Nimmo, C. *et al.* Bedaquiline resistance in drug-resistant tuberculosis HIV co-infected patients. *Eur Respir J* **55** (2020). https://doi.org:10.1183/13993003.02383-2019

37 Martinez, E. *et al.* Mutations associated with in vitro resistance to bedaquiline in *Mycobacterium tuberculosis* isolates in Australia. *Tuberculosis (Edinb)* **111**, 31-34 (2018). https://doi.org:10.1016/j.tube.2018.04.007

38 Villellas, C. *et al.* Unexpected high prevalence of resistance-associated *Rv0678* variants in MDR-TB patients without documented prior use of clofazimine or bedaquiline. *J Antimicrob Chemother* **72**, 684-690 (2017). https://doi.org:10.1093/jac/dkw502

39 Merker, M. *et al.* Phylogenetically informative mutations in genes implicated in antibiotic resistance in *Mycobacterium tuberculosis* complex. *Genome Med* **12**, 27 (2020). https://doi.org:10.1186/s13073-020-00726-5

40 Kay, G. L. *et al.* Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* **6**, 6717 (2015). https://doi.org:10.1038/ncomms7717

41 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* **1**, e165-e174 (2020). https://doi.org:10.1016/S2666-5247(20)30031-8

42 Nimmo, C. *et al.* Dynamics of within-host *Mycobacterium tuberculosis* diversity and heteroresistance during treatment. *EBioMedicine* **55**, 102747 (2020). https://doi.org:10.1016/j.ebiom.2020.102747

43 O'Neill, M. B. *et al.* Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Mol Ecol* **28**, 3241-3256 (2019). https://doi.org:10.1111/mec.15120

44 Brynildsrud, O. B. *et al.* Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. *Sci Adv* **4**, eaat5869 (2018). https://doi.org:10.1126/sciadv.aat5869

45 Rutaihwa, L. K. *et al.* Multiple Introductions of *Mycobacterium tuberculosis* Lineage 2-Beijing Into Africa Over Centuries. *Front Ecol Evol* **7** (2019). https://doi.org:ARTN 112
10.3389/fevo.2019.00112

46 Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat Commun* **6**, 10063 (2015). https://doi.org:10.1038/ncomms10063

47 Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med* **11**, 41 (2019). https://doi.org:10.1186/s13073-019-0650-x

48 Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019). https://doi.org:10.1371/journal.pcbi.1006650

49 Menardo, F., Duchene, S., Brites, D. & Gagneux, S. The molecular clock of *Mycobacterium tuberculosis*. *PLoS Pathog* **15**, e1008067 (2019). https://doi.org:10.1371/journal.ppat.1008067

50 Ismail, N., Peters, R. P. H., Ismail, N. A. & Omar, S. V. Clofazimine Exposure In Vitro Selects Efflux Pump Mutants and Bedaquiline Resistance. *Antimicrob Agents Chemother* **63** (2019). https://doi.org:10.1128/AAC.02141-18

51 Andres, S. *et al.* Bedaquiline-Resistant Tuberculosis: Dark Clouds on the Horizon. *Am J Respir Crit Care Med* **201**, 1564-1568 (2020). https://doi.org:10.1164/rccm.201909-1819LE

52 The, C. C. A data compendium associating the genomes of 12,289 *Mycobacterium tuberculosis* isolates with quantitative resistance phenotypes to 13 antibiotics. *PLOS Biology* **20**, e3001721 (2022). https://doi.org:10.1371/journal.pbio.3001721

53 The, C. C. Epidemiological cutoff values for a 96-well broth microdilution plate for high-throughput research antibiotic susceptibility testing of *M. tuberculosis*. *European Respiratory Journal*, 2200239 (2022). https://doi.org:10.1183/13993003.00239-2022

54 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 Med* **12**, 104 (2020). https://doi.org:10.1186/s13073-020-00793-8

55 Hariguchi, N. *et al.* OPC-167832, a Novel Carbostyril Derivative with Potent Antituberculosis Activity as a *DprE1* Inhibitor. *Antimicrob Agents Chemother* **64** (2020). https://doi.org:10.1128/AAC.02020-19

56 Loiseau, C. *et al.* An African origin for *Mycobacterium bovis*. *Evol Med Public Health* **2020**, 49-59 (2020). https://doi.org:10.1093/emph/eoaa005

57 Bateson, A. *et al.* Ancient and recent differences in the intrinsic susceptibility of *Mycobacterium tuberculosis* complex to pretomanid. *J Antimicrob Chemother* **77**, 1685-1693 (2022). https://doi.org:10.1093/jac/dkac070

58 D'Costa, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477**, 457-461 (2011). https://doi.org:10.1038/nature10388

59 Rifat, D. *et al.* Mutations in fbiD (Rv2983) as a Novel Determinant of Resistance to Pretomanid and Delamanid in *Mycobacterium tuberculosis*. *Antimicrob Agents Ch* **65** (2021). https://doi.org:ARTN e01948-20
10.1128/AAC.01948-20

60 Liu, Y. *et al.* Reduced Susceptibility of *Mycobacterium tuberculosis* to Bedaquiline During Antituberculosis Treatment and Its Correlation With Clinical Outcomes in China. *Clin Infect Dis* **73**, e3391-e3397 (2021). https://doi.org:10.1093/cid/ciaa1002

61 Pym, A. S. *et al.* Bedaquiline in the treatment of multidrug- and extensively drug-resistant tuberculosis. *Eur Respir J* **47**, 564-574 (2016). https://doi.org:10.1183/13993003.00724-2015

62 Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* **5**, 4812 (2014). https://doi.org:10.1038/ncomms5812

63 Sobkowiak, B. *et al.* Identifying mixed *Mycobacterium tuberculosis* infections from whole genome sequence data. *BMC Genomics* **19**, 613 (2018). https://doi.org:10.1186/s12864-018-4988-z

64 Bradley, P., den Bakker, H. C., Rocha, E. P. C., McVean, G. & Iqbal, Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat Biotechnol* **37**, 152-159 (2019). https://doi.org:10.1038/s41587-018-0010-1

65 Merker, M. *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet* **47**, 242-249 (2015). https://doi.org:10.1038/ng.3195

66 Luo, T. *et al.* Southern East Asian origin and coexpansion of *Mycobacterium tuberculosis* Beijing family with Han Chinese. *Proc Natl Acad Sci U S A* **112**, 8136-8141 (2015). https://doi.org:10.1073/pnas.1424063112

67 Norheim, G. *et al.* Tuberculosis Outbreak in an Educational Institution in Norway. *J Clin Microbiol* **55**, 1327-1333 (2017). https://doi.org:10.1128/JCM.01152-16

68 Nimmo, C. *et al.* Whole genome sequencing *Mycobacterium tuberculosis* directly from sputum identifies more genetic diversity than sequencing from culture. *BMC Genomics* **20**, 389 (2019). https://doi.org:10.1186/s12864-019-5782-2

69 Dheda, K. *et al.* Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. *Lancet Respir Med* **5**, 269-281 (2017). https://doi.org:10.1016/S2213-2600(16)30433-7

70 Streicher, E. M. *et al.* Molecular Epidemiological Interpretation of the Epidemic of Extensively Drug-Resistant Tuberculosis in South Africa. *J Clin Microbiol* **53**, 3650-3653 (2015). https://doi.org:10.1128/JCM.01414-15

71 Guerra-Assuncao, J. A. *et al.* Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* **4** (2015). https://doi.org:10.7554/eLife.05166

72 Grandjean, L. *et al.* Transmission of Multidrug-Resistant and Drug-Susceptible Tuberculosis within Households: A Prospective Cohort Study. *PLoS Med* **12**, e1001843; discussion e1001843 (2015). https://doi.org:10.1371/journal.pmed.1001843

73 Grandjean, L. *et al.* Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru. *PLoS One* **12**, e0189838 (2017). https://doi.org:10.1371/journal.pone.0189838

74 Ismail, N., Omar, S. V., Ismail, N. A. & Peters, R. P. H. Collated data of mutation frequencies and associated genetic variants of bedaquiline, clofazimine and linezolid resistance in *Mycobacterium tuberculosis*. *Data Brief* **20**, 1975-1983 (2018). https://doi.org:10.1016/j.dib.2018.09.057

75 Ghajavand, H. *et al.* High Prevalence of Bedaquiline Resistance in Treatment-Naive Tuberculosis Patients and Verapamil Effectiveness. *Antimicrob Agents Chemother* **63** (2019). https://doi.org:10.1128/AAC.02530-18

76 pygsi v1.0.0 (2018).

77 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013). https://doi.org:arXiv:1303.3997

78 Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-11 10 33 (2013). https://doi.org:10.1002/0471250953.bi1110s43

79 Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292-294 (2016). https://doi.org:10.1093/bioinformatics/btv566

80 Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* **7**, 51 (2015). https://doi.org:10.1186/s13073-015-0164-0

81 Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056 (2016). https://doi.org:10.1099/mgen.0.000056

82 Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453-4455 (2019). https://doi.org:10.1093/bioinformatics/btz305

83 Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res* **46**, e134 (2018). https://doi.org:10.1093/nar/gky783

84 Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018). https://doi.org:10.1186/s12859-018-2164-8

85 Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol Biol* **17**, 42 (2017). https://doi.org:10.1186/s12862-017-0890-6

86 Baele, G. *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* **29**, 2157-2167 (2012). https://doi.org:10.1093/molbev/mss084

87 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2019). https://doi.org:10.1093/bioinformatics/bty633

88 Yu, G. C., Smith, D. K., Zhu, H. C., Guan, Y. & Lam, T. T. Y. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**, 28-36 (2017). https://doi.org:10.1111/2041-210x.12628