bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Detection of a bedaquiline / clofazimine resistance reservoir in *Mycobacterium tuberculosis* predating the antibiotic era

Running title: Emergence of bedaquiline resistance in tuberculosis

Lucy van Dorp[1]*, 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]

*These authors contributed equally.

Correspondence: Camus Nimmo (c.nimmo@ucl.ac.uk) and François Balloux (f.balloux@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

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 resistance-associated variants (RAVs) in the *Rv0678* gene which can also confer cross-resistance to clofazimine, another TB drug. We compiled a dataset of 3,682 *Mtb* genomes, including 223 carrying *Rv0678* bedaquiline RAVs. We identified at least 15 cases where RAVs were present in the genomes of strains collected prior to the use of bedaquiline in TB treatment regimes. Phylogenetic analyses point to multiple emergence events and in some cases widespread circulation of RAVs in *Rv0678*, often prior to the introduction of bedaquiline or clofazimine. Strikingly, this included three cases predating the antibiotic era. The presence of a pre-existing reservoir of bedaquiline-resistant *Mtb* strains necessitates the urgent implementation of rapid drug susceptibility testing and individualised regimen selection to safeguard the use of bedaquiline in TB care and control.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Drug-resistant tuberculosis (DR-TB) currently accounts for 500,000 of the 10 million new tuberculosis (TB) cases reported annually [1], with incidence expected to rise substantially due to the ongoing Covid-19 pandemic [2]. 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 [1,3]. 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], bedaquiline received in 2012 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, defined by 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]. In addition, bedaquiline is positioned as a key drug in multiple phase III clinical trials for drug-susceptible TB (SimpliciTB, ClinicalTrials.gov NCT03338621), MDR-TB (STREAM2, ClinicalTrials.gov NCT02409290) and XDR-TB (ZENIX-TB, ClinicalTrials.gov NCT03086486).

Unlike other major drug-resistant bacteria, *Mtb* reproduces strictly clonally and systematically acquires resistance by chromosomal mutations rather than via horizontal gene transfer or recombination [12]. Phylogenetic reconstructions based on whole genome sequencing can therefore accurately infer the time of emergence and subsequent spread of *Mtb* resistance-associated variants (RAVs). Phylogenetic studies have demonstrated that there are often multiple *Mtb* lineages introduced into distinct geographical regions, with repeated independent drug resistance emergence events occurring locally [13–16].

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Mtb has demonstrated the ability to acquire resistance to every drug used against it until now. Resistance has been reported to occur soon after the introduction of a novel TB drug [17,18]. For example, mutations conferring resistance to isoniazid – one of the first antimycobacterials – tend to have emerged 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 decades prior to their identification, leading to the largest reported outbreak of extensively drug-resistant TB (XDR-TB) [18].

Mutations conferring resistance to bedaquiline were first selected *in vitro*, and were located in the *atpE* gene encoding the target F1F0 ATP synthase, the target of bedaquiline [19]. Subsequently, resistance-conferring mutations have been found in *pepQ* in a murine model and potentially in a small number of patients [20]. However, the vast majority of resistance observed in clinical isolates has been identified in the context of resistance-associated variants (RAVs) in the *Rv0678* gene, a negative repressor of expression of the MmpL5 efflux pump. Loss of function of *Rv0678* leads to pump overexpression [21] and increased minimum inhibitory concentrations (MIC) to bedaquiline, as well as to the recently repurposed antimycobacterial clofazimine and the azole class of antifungal drugs (which also have antimycobacterial activity) [22].

A diverse range of single nucleotide variants (SNVs) and frameshift *Rv0678* mutations have been associated with resistance to bedaquiline, and are often present as heteroresistant alleles in patients [23–30]. In contrast to most other RAVs in *Mtb,* which often cause many-fold increases in MIC and clear-cut resistance, *Rv0678* variants may be associated with normal MICs or subtle increases in bedaquiline MIC, although they may still be clinically important. These increases may not cross the current WHO critical concentrations used to classify resistant versus susceptible strains (0.25µg/mL on Middlebrook 7H11 agar, or 1µg/mL in Mycobacteria Growth Indicator Tube [MGIT] liquid media). Bedaquiline has a long terminal half-life of up to 5.5 months [6], leading to the possibility of subtherapeutic

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

concentrations, where adherence is suboptimal or treatment is interrupted, which could act as a further driver of resistance.

Bedaquiline and clofazimine cross-resistance has now been reported across three continents following the rapid expansion in usage of both drugs [24,29,31,32], and is associated with poor adherence to therapy and inadequate regimens. However, baseline isolates in 8/347 (2.3%) patients from phase IIb bedaquiline trials demonstrated *Rv0678* RAVs and high bedaquiline MICs in the absence of prior documented use of bedaquiline or clofazimine [33], suggesting that bedaquiline RAVs could pre-exist in many settings where bedaquiline will be used. While there are isolated clinical reports from multiple geographical regions, the global situation regarding bedaquiline resistance emergence and spread has not yet been investigated.

In this study, we characterised and dated the emergence of bedaquiline RAVs in the two global *Mtb* lineage 2 (L2) and lineage 4 (L4) lineages, which include the majority of drug resistance strains [17]. Phylogenetic analyses of two datasets comprising 1,514 *Mtb* L2 and 2,168 L4 whole genome sequences revealed the emergence and spread of multiple *Rv0678* RAVs prior to the use of bedaquiline or clofazimine, with some mutations having been in circulation already before the antibiotic era. This pre-existing reservoir of bedaquiline/clofazimine-resistant *Mtb* strains suggests *Rv0678* RAVs exert a relatively low fitness cost which could be rapidly selected for as bedaquiline and clofazimine are more widely used in the treatment of TB.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

The global diversity of *Mtb* lineage L2 and L4

To investigate the global distribution of *Mtb* isolates with variants in *Rv0678*, we curated two large datasets of whole genomes from the two dominant global lineages L2 and L4. Both datasets were selectively enriched for samples with variants in *Rv0678* (see **Methods**) and those with accompanying full metadata for geolocation and time of sampling (**Figure 1, Supplementary Table S1-S2, Supplementary Figure S1)**. The final L2 dataset included 1,514 isolates collected over 24.5 years (between 1994 and 2019) yielding 29,205 SNPs. The L4 dataset comprised 2,168 sequences collected over 232 years, including three samples from 18th century Hungarian mummies [34], encompassing 67,585 SNPs. Both datasets included recently generated data from South Africa (155 L2, 243 L4) [16,35] and new whole genome sequencing data from Peru (9 L2, 154 L4).

Consistent with previous studies [15,36,37], both datasets are highly diverse and exhibit strong geographic structure (**Figure 2**). As a nonrecombining clonal organism, identification of mutations in *Mtb* can provide a mechanism to predict phenotypic resistance from a known panel of genotypes [38,39]. Based on genotypic profiling [39], within the L2 dataset, 911 strains were classified as MDR-TB (60%) and 295 (20%) as XDR-TB. Within the L4 dataset, 911 isolates were classified as MDR-TB (42%) and 115 as XDR-TB (5%). The full phylogenetic distribution of resistance profiles is provided in **Supplementary Figure S2**. As is commonplace with genomic datasets, these percentages of drug-resistant strains exceed their actual prevalence, due to the overrepresentation of drug-resistant isolates in public repositories.

Both the L2 and L4 phylogenetic trees displayed a significant temporal signal following date randomisation (**Supplementary Figure S3**), making them suitable for time-calibrated phylogenetic inference. We estimated the time to the Most Recent Common Ancestor (tMRCA) of both datasets using a Bayesian tip-dating analysis (BEAST2) run on a representative subset of genomes from each dataset (see **Methods**, **Supplementary Table 3**, **Supplementary Figure S4**). For the final temporal

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

calibration of the L2 dataset we applied an estimated clock rate of 7.7×10^{-8} (4.9×10^{-8} - 1.03×10^{-7}) substitutions per site per year, obtained from the subsampled BEAST2 [40] analysis, to the global maximum likelihood phylogenetic tree resulting in an estimated tMRCA of 1332CE (945CE-1503CE). Using the same approach for the L4 dataset we estimated a clock rate of 7.1×10^{-8} (6.2×10^{-8} - 7.9×10^{-8}) substitutions per site per year resulting in an estimated tMRCA of 853CE (685CE – 967CE) (**Figure 2**). We observed a slightly higher, yet statistically not significant, clock rate in L2 compared to L4 (**Supplementary Table S3**), with all estimated substitution rates falling largely in line with previously published estimates [41].

Identification of *Rv0678* variants

Since *atpE* and *pepQ* bedaquiline RAVs are found at very low prevalence, we focused on characterising the full mutational spectrum of *Rv068* across both lineages. In total we identified the presence of non-synonymous and promoter *Rv0678* variants in 438 sequences (194 L2, 244 L4). We classified all identified non-synonymous and promoter mutations in *Rv0678*, based on the literature, into six phenotypic categories for bedaquiline susceptibility: wild type, hypersusceptible, susceptible, intermediate, resistant and unknown (full references available in **Supplementary Table S4, Supplementary Figures S5-S7)**. Across both lineages, 240 sequences were considered as bedaquiline resistant (i.e. classified as intermediate or resistant based on their genotype). The most commonly observed variants are listed in **Table 1**. Notably we identified several sequenced isolates carrying nonsynonymous variants in *Rv0678* uploaded with collection dates prior to the first clinical trials for bedaquiline in 2007. For L2 we identified ten cases collected before 2007, of which eight comprised variants previously associated to phenotypic bedaquiline resistance (RAVs). For L4 we identified 15 sequences with *Rv0678* variants predating 2007, of which seven have previously been associated with phenotypic bedaquiline resistance (RAVs) and six classified as conferring an intermediate resistance phenotype (**Figure 1c-d, Supplementary Table S5**).

Of the 198 L2 isolates identified as carrying variants in *Rv0678*, 18 samples had more than one variant in the same gene (10%). In L4, 14 samples (6%) were observed with more than one variant co-occurring

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

in the *Rv0678* gene. We identified a significant relationship between the presence of *Rv0678* variants and drug resistance status in both the L2 and L4 datasets (**Supplementary Figure S8-S9**), though in both cases we identified otherwise fully phenotypically susceptible isolates carrying *Rv0678* RAVs (12 L2, 25 L4).

We identified one L2 isolate (ERR2677436 sampled in Germany in 2016) which already had two *Rv0678* RAVs at low allele frequency – Val7fs (11%) and Val20Phe (20%) – and also contained two low frequency *atpE* RAVs: Glu61Asp (3.2%) and Alal63Pro (3.7%). We also identified three isolates obtained in 2007-08 from separate but neighbouring Chinese provinces carrying the *Rv1979c* Val52Gly RAV, which has been reported to be associated with clofazimine resistance in a study from China [24] but was associated with a normal MIC in another [42]. Furthermore, several frameshift and premature stop mutations in *pepQ* have been previously associated with bedaquiline and clofazimine resistance. In this dataset, we identified 18 frameshift mutations in *pepQ* across 11 patients, one of which also had a *Rv0678* frameshift mutation. In one isolate the *pepQ* frameshift occurred at the Arg271 position previously reported to be associated with bedaquiline resistance [20].

Prediction of phenotype based on *Rv0678* variants

Across our datasets we identified 62 genomes with nonsynonymous *Rv0678* variants of unknown phenotypic effect (12 L2, 50 L4), corresponding to 23 unique mutations or combinations of mutations. A gradient-boosted tree classifier was trained and optimised to determine if the amino acid properties of *Rv0678* mutations associated to known bedaquiline resistance phenotypes can be used to predict the resistance status of mutations with no available phenotypic information. The optimised model provided an area under the precision-recall curve (AUPRC) of 0.805 (**Supplementary Table S6**), suggesting that the physiochemical properties of mutations can be used to successfully differentiate between resistance and susceptibility phenotypes (**Supplementary Table S7**). The features of the models were then interpreted using SHAP values (see Methods). Via this approach, we found that 5' end mutations, mutations in the DNA binding domain and polar and positively charged residues are associated with resistance. Conversely, mutations in the dimerisation domains, transitions from negatively to positively

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

charged residues, and mutations involving hydrophobic wild type or variant residues are associated with susceptibility (**Supplementary Figure S10**).

The time to emergence of *Rv0678* variants

To estimate the age of the emergence of different *Rv0678* non-synonymous variants, we identified all nodes in the global time calibrated phylogenies delineating clades of isolates carrying a particular *Rv0678* variant (**Figure 3, Supplementary Table S8**). For the L2 dataset we identified 58 unique phylogenetic nodes where *Rv0678* RAVs emerged, of which 40 were represented by a single genome. The point estimates for these nodes ranged from March 1845 to November 2018. Eight variants, including four bedaquiline RAVs, were estimated to have emergence dates (point estimates) predating the first bedaquiline clinical trial in 2007 (**Supplementary Figure S11**).

For the L4 dataset we identify 85 unique nodes where *Rv0678* RAVs emerged, of which 59 were represented by a single isolate in the dataset. The point estimates for these nodes ranged from September 1701 to January 2019 (**Figure 3, Supplementary Figure S12**). Sixteen *Rv0678* mutations, including six bedaquiline RAVs and two predicted to have an intermediate phenotype, were estimated to have emerged prior to 2007. We also identified one large clade of 65 samples, predominantly collected in Peru, which all carry the Ile67fs *Rv0687* RAV [31,43,44]. While it is not inconceivable that multiple independent emergences of Ile67fs occurred in this clade, the by far more parsimonious scenario is a single ancestral emergence. We estimate the time of this emergence to 1702 (1657-1732). This significantly predates the first use of azoles, clofazimine or indeed bedaquiline (**Supplementary Figure S12-S13**). While we identified no nodes with secondary emergence of *Rv0678* nonsynonymous mutations across the L4 dataset, eight nodes were identified in the L2 dataset where a clade already carrying a nonsynonymous variant in *Rv0678* subsequently acquired a second nonsynonymous mutation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 1: Number of sequences with resistance-associated variants (i.e. classified as resistant or intermediate) in *Rv0678* detected for all variants occurring ≥ 5 times.

***RV0678* RAV**	**L 2**	**L 4**	**TOTAL**
ILE67FS	5	83	88
MET146THR	2	20	22
ARG90CYS	0	9	9
GLU49FS	0	8	8
ALA59VAL	7	0	7
VAL1ALA	6	0	6
ASP141FS	0	6	6
ASN98ASP	0	6	6
GLY121ARG	5	0	5
ARG109LEU & ARG156RFS	0	5	5

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Discussion

Our work establishes that the emergence of variants in *Rv0678*, including RAVs, is not solely driven by the use of bedaquiline and clofazimine or azoles (which have also been proposed as a further potential selective force) [22]. In particular we identified 12 cases of emergence of bedaquiline RAVs prior to the first clinical trials of bedaquiline in 2007. Phylogenetic inference estimated the oldest bedaquiline-resistant clade, composed mostly of samples from Peru carrying the Ile67fs RAV, to have emerged around 1702 (1657-1732), suggesting bedaquiline RAVs have been in circulation for as long as 300 years. Our phylogenetic inference, pointing to multiple emergences of *Rv0678* nonsynonymous variants predating the use of bedaquiline, is also confirmed by the observation of 15 *Mtb* genomes carrying *Rv0678* RAVs sampled prior to 2007. The long-term circulation of bedaquiline RAVs predating the use of the drug is of concern as it suggests that non-synonymous mutations in *Rv0678* exert little fitness cost. It also points to a pre-existent reservoir of bedaquiline resistant *Mtb*, including in some otherwise fully susceptible strains, which are likely to rapidly expand under drug pressure with the increasing use of bedaquiline and clofazimine in TB treatment.

We identified a large number of different *Rv0678* nonsynonymous variants across both of our *Mtb* lineage cohorts; 45 in L2 (including 10 unique RAV combinations) and 67 in L4 (nine unique RAV combinations). Any mutation leading to the loss of function of the *Rv0678* protein is expected to translate into raised bedaquiline MICs, through the overexpression of the MmpL5 efflux pump, although there have been some exceptions reported [33]. As such, the mutational target leading to bedaquiline resistance is wider than for most other current TB drugs and raises concerns about the ease with which bedaquiline resistance can emerge during treatment. It is further concerning that resistance to the new class of nitroimidazole drugs, such as pretomanid and delamanid, is also conferred by loss of function mutations in any of at least five genes, suggesting that they may also have a low barrier to resistance.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

While we identified many non-synonymous variants in *Rv0678*, we acknowledge that several of our detected variants have no associated MIC values available in the literature and are thus currently not phenotypically validated. In the absence of phenotypic data, machine learning approaches offer some potential to predict the resistance status of given variants, and our small-scale analysis suggests the potential of such an approach. However, even determining the phenotypic consequences of *Rv0678* variants that have previously been described is challenging as there are often only limited reports correlating MICs to genotypes. Moreover, at least four different methods are used to determine MICs, some of which do not have associated critical concentrations. Even where critical concentrations have been set, there is an overlap in MICs of isolates that are genetically wild type and those that have mutations likely to cause resistance [45], making a correlation between genotype, phenotype and clinical impact challenging.

Prediction of phenotypic bedaquiline resistance from genomic data is further complicated by the existence of hyper-susceptibility variants. For example, the C-11A variant located in the promoter of *Rv0678*, which appears to increase susceptibly to bedaquiline [33], was observed to be fixed throughout a large clade within L2. The early emergence of this variant and its geographical concentration in South Africa and eSwatini may further suggest the role of non-pharmacological influences on *Rv0678* which regulates multiple MmpL efflux systems [21]. While large-scale genotype/phenotype analyses will likely support the development of rapid molecular diagnostics, targeted or whole genome sequencing, at reasonable depths, may provide the only opportunity to detect all possible *Rv0678* RAVs in clinical settings.

Bedaquiline resistance can also be conferred by other RAVs including in *pepQ* (bedaquiline and clofazimine), *atpE* (bedaquiline only) [44] and *Rv1979c* (clofazimine only). We only found *atpE* RAVs at low allele frequency in one patient who also had *Rv0678* variants (sample accession ERR2677436), which is in line with other evidence suggesting they rarely occur in clinical isolates, likely due to a high fitness cost. Likewise, we only identified *Rv1979c* RAVs in three patients in China, although there were other variants in *Rv1979c* for which ability to cause phenotypic resistance has not been previously

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

assessed. Frameshift *pepQ* mutations that are potentially causative of resistance were identified in 11 patients, in keeping with its possible role as an additional rare resistance mechanism.

Our findings are of high clinical relevance as the presence of *Rv0678* variants during therapy in clinical strains has been associated with substantially worse outcomes in patients treated with drug regimens including bedaquiline [31]. Although it is uncertain what the impact of *Rv0678* RAVs are on outcomes when present prior to treatment [46,47], it is imperative to monitor and prevent the wider transmission of bedaquiline resistant clones, particularly in high MDR/XDR-TB settings. The large and disparate set of mutations in *Rv0678* we identified, with differing phenotypes and some being already in circulation before the pre-antibiotic era, adds further urgency to the development of rapid drug susceptibility testing for bedaquiline to inform effective treatment choices and mitigate the further spread of DR-TB.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Materials and methods

Sample collection

In this study we curated large representative datasets of *Mtb* whole genome sequences encompassing the global genetic and geographic distribution of lineages 2 (L2) and L4 (**Figure 1, Supplementary Tables S1-S2**). The dataset was enriched to include all available sequenced isolates with *Rv0678* variants, which in some cases included isolates with no, or limited, published metadata. In all other cases samples for which metadata on the geographic location and date of collection was available were retained. To ensure high quality consensus alignments we required that all samples mapped with a minimum percentage cover of 96% and a mean coverage of 30x to the H37Rv reference genome (NC_000962.3). We excluded any samples with evidence of mixed strain infection as identified by the presence of lineage-specific SNPs to more than one sublineage [48] or the presence of a high proportion of heterozygous alleles [49]. The total number of samples included in these datasets, and their source is shown in **Supplementary Table S2.** An index of all samples is available in **Supplementary Table S1**.

A large global dataset of 1,669 L4 *Mtb* sequences has recently been constructed, which we used as the basis for curating our L4 dataset [13]. We refer to this as the 'base dataset' for L4. For L2, we constructed a 'base dataset' by screening the Sequence Read Archive (SRA) and European Nucleotide Archive (ENA) using BIGSI [50] for the *rpsA* gene sequence containing the L2 defining variant *rpsA* a636c [48] with a 100% match. This search returned 6,307 *Mtb* genomes, of which 1,272 represented unique samples that had the minimum required metadata. Metadata from three studies were also added manually as they were not included in their respective SRA submissions but were available within published studies [14,51,52].

For isolates with only information on the year of sample collection, we set the date to be equal to the middle of the year. For those with information on the month but not the date of collection we set the date of collection to the first of the month. For sequenced samples which were missing associated

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

metadata (32 L2 genomes and 19 L4 genomes) we attempted to estimate an average time of sample collection in order to impute a sampling date. To do so we computed the average time between date of collection and sequence upload date for all samples with associated dates separately in each of the L2 and L4 datasets (**Supplementary Figure S1**). For L2 we estimated a mean lag time of 4.7 years (0.5–12.6 years 95% CI). For L4, having excluded three sequences obtained from 18th Century mummies from Hungary [34], we estimated a mean lag time of 6.9 years (0.6-19.1 years 95% CI).

To enrich the datasets for isolates with *Rv0678* variants, we included further sequences from our own published studies in KwaZulu-Natal, South Africa [16,35], other studies of drug-resistant TB in southern Africa [18,37,53–56], and Peru [57,58]. We additionally supplement the Peruvian data with 163 previously unpublished isolates. In these cases, and to facilitate the most accurate possible estimation of the date of resistance emergence, we included samples with *Rv0678* variants as well as genetically related sequences without *Rv0678* variants.

To identify further published raw sequencing data with *Rv0678* variants from studies where bedaquiline/clofazimine resistance may have been previously unidentified, we screened the NCBI Sequencing Read Archive (SRA) for sequence data containing 85 previously published *Rv0678* variants [16,27–29,35,59,60] with BIGSI [50]. BIGSI was employed against a publicly available indexed database of complete SRA/ENA bacterial and viral whole genome sequences current to December 2016 (available here: http://ftp.ebi.ac.uk/pub/software/bigsi/nat_biotech_2018/all-microbial-index-v03/, last accessed 30/07/2020), and also employed locally against an updated in-house database which additionally indexed SRA samples from January 2017 until January 2019. Samples added using this approach are flagged 'BIGSI' in **Supplementary Table S1**. We also used the PYGSI tool (DOI:10.5281/zenodo.1407085) to interrogate BIGSI with the *Rv0678* sequence adjusted to include every possible single nucleotide substitution. In each instance we included 30 bases upstream and downstream of the gene as annotated on the H37Rv *Mtb* reference genome. For the purpose of this study we only considered coding region, non-synonymous substitutions and insertions and deletions. Samples added following the PYGSI screen are flagged 'PYGSI' in **Supplementary Table S1**. In the

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

final L2 dataset 194/1514 (12.8%) samples had *Rv0678* variants, and in L4 this proportion was 244/2168 (11.3%). A breakdown of the different datasets used is provided in **Supplementary Table S2**.

Reference mapping and variant calling

Original fastq files for all included sequences were downloaded and paired reads mapped to the H37Rv reference genome with bwa mem v0.7.17 [61]. Mapped reads were sorted and de-duplicated using Picard Tools v2.20 followed by indel realignment with GATK v3.8 [62]. Alignment quality and coverage was recorded with Qualimap v2.21 [63]. Variant calling was performed using bcftools v1.9, based on reads mapping with a minimum mapping quality of 20, base quality of 20, no evidence of strand or position bias, a minimum coverage depth of 10 reads, and a minimum of four reads supporting the alternate allele, with at least two of them on each strand. Moreover, SNPs that were less than 2bp apart of an indel were excluded from the analysis. Similarly, only indels 3bp apart of other indels were kept.

All sites with insufficient coverage to identify a site as variant or reference were excluded (marked as 'N'), as were those in or within 100 bases of PE/PPE genes, or in insertion sequences or phages. SNPs present in the alignment with at least 90% frequency were used to generate a pseudoalignment of equal length to the H37Rv reference using a custom Python script for use in phylogenetic analysis. Samples with more than 10% of the alignment represented by ambiguous bases were excluded. Those positions with more than 10% of ambiguous bases across all the samples were also removed. In order to avoid bias on the tree structure, positions known to be associated with drug resistance were not included.

A more permissive variant calling pipeline was used to identify *Rv0678* variants, as they are often present at <100% frequency with a high incidence of frameshift mutations. Here we instead employed FreeBayes v1.2 [64] to call all variants present in the *Rv0678* gene (or up to 100 bases upstream) that were present at ≥5% frequency (alternate allele fraction *-F* 0.05) and supported by at least four reads including one on each strand.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Classification of resistance variants

All raw fastq files were screened using the rapid resistance profiling tool TBProfiler [39,65] against a curated whole genome drug resistance mutations library. This allowed rapid assignment of polymorphisms associated with resistance to different antimycobacterial drugs and categorisation of MDR and XDR *Mtb* status (**Supplementary Figure S2, Supplementary Figures S5-S9**).

Classification of *Rv0678* variants

The diverse range of *Rv0678* variants and paucity of widespread MIC testing means that there are limited data from which to infer the phenotypic consequences of identified *Rv0678* variants. The approach we used was to assign whether nonsynonymous variants confer a normal or raised MIC based on published phenotypic tests for strains carrying that variant. A full list of the literature reports used for each mutation is provided in **Supplementary Table S4**. We also introduced an intermediate category to describe isolates with MICs at the critical concentration (e.g. 0.25µg/mL on Middlebrook 7H11 agar), where there is an overlap of the MIC distributions of *Rv0678* mutated and wild type isolates with uncertain clinical implications [45]. We assumed that all other disruptive frameshift and stop mutations would confer resistance in light of the role of *Rv0678* as a negative repressor, where loss of function should lead to efflux pump overexpression. All other promoter and previously unreported missense mutations were categorised as unknown (**Supplementary Table S4**). We were able to categorise 29/85 (34.1%) of the different non-synonymous and promoter mutations identified.

Prediction of phenotypic effect of *Rv0678* variants

A gradient-boosted tree classifier was developed using the XGBoost API (v1.0.2) [66] to determine whether the phenotypes of genomes with variants in *Rv0678* of unknown effect could be predicted based on the change in amino acid properties for known resistant and susceptible variants. Fifteen features were engineered based on the wild type and mutant residues of each mutation as follows. Two features represent the amino-acid residue of the wild type and of the mutant. Two features encode whether they are non-polar, polar, positively charged or negatively charged. Two features represent whether the wild type and the mutant residues are hydrophobic or hydrophilic, based on the

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

hydrophobicity scale proposed by Janin [67]. Two features encode the molecular weight of wild type and mutant AA. Two features represent the change in charge or molecular weight from the wild type to mutant, where non-polar and polar residues are assumed to contribute a charge of zero. One feature represents the change in hydrophobicity, where a hydrophobic→hydrophilic residue change is coded as +1 and the reverse as -1. Three features represent mutations in the DNA-binding domain, in the dimerisation domain, and to the residues in contact with 2-stearoylglycerol. The last feature represented the 5'- 3' position of amino acid mutations.

Only variants in *Rv0678* which have a demonstrated association to a bedaquiline-resistant or susceptible phenotype were used, resulting in 59 resistance and 32 susceptibility mutations. Model parameters were optimised to maximise the F1 score and model performance was estimated using a nested, stratified, 10 x 10 cross-validation procedure. AUPRC was used for model evaluation due to class imbalance in the dataset [68]. The trained model was then interpreted using TreeExplainer as part of the shap API (v0.35.0) (64). Each feature is assigned a SHAP value which represents the change in predicted probability score in each prediction when a feature is included or excluded from the model. Visualisation of the SHAP values in tandem with the feature values (**Supplementary Figure S10**) allows inference of how each feature contributes to each prediction. All scripts used for the analysis are hosted on GitHub (https://github.com/cednotsed/TB-Bedaquiline-Resistance-Modelling.git). Predictions were also made using the Protein Variation Effect Analyzer (PROVEAN) via the online interface [69]. The final predicted probability of resistance and associated PROVEAN scores are provided in **Supplementary Table S7**.

Global phylogenetic inference

The alignments for phylogenetic inference were masked for the *Rv0678* region using bedtools v2.25.0. All variant positions were extracted from the resulting global phylogenetic alignments using snp-sites v2.4.1 [70], including a L4 outgroup for the L2 alignment (NC_000962.3) and a lineage 3 (L3) outgroup for the L4 alignment (SRR1188186). This resulted in a 67,585 SNP alignment for the L4 dataset and 29,205 SNP alignment for the L2 dataset. A maximum likelihood phylogenetic tree was constructed for both SNP alignments using RAxML-NG v0.9.0 [71] specifying a GTR+G substitution model, correcting

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

for the number of invariant sites using the ascertainment flag (ASC_STAM) and specifying a minimum branch length of $1x10^{-9}$ reporting 12 decimal places (--precision 12).

Estimating the age of emergence of *Rv0678* variants

To test whether the resulting phylogenies can be time-calibrated we first dropped the outgroups from the phylogeny and rescaled the trees so that branches were measured in unit of substitutions per genome. We then computed a linear regression between root-to-tip distance and the time of sample collection using BactDating [72], which additionally assesses the significance of the regression based on 10,000 date randomisations. We obtained a significant temporal correlation for both the L2 and L4 phylogenies, both with and without imputation of dates for samples with missing metadata, (**Supplementary Figures 3**).

We employed the Bayesian method BactDating v1.01 [72], run without updating the root (updateRoot=F), a mixed relaxed gamma clock model and otherwise default parameters to both global datasets. The MCMC chain was run for $1x10^7$ iterations and $3x10^7$ iterations. BactDating results were considered only when MCMC chains converged with an Effective Sample Space (ESS) of at least 100. The analysis was applied to the imputed and non-imputed collection dates for genomes with missing metadata (**Supplementary Table 3**).

To independently infer the evolutionary rates associated with each of our datasets, we sub-sampled both the L4 and L2 datasets to 200 isolates, selected so as to retain the maximal diversity of the tree using Treemmer v0.3 [73]. This resulted in a dataset for L4 comprising 25,104 SNPs and spanning 232 years of sample collection dates and for L2 comprising 8,221 SNPs and spanning 24 years of sample collection dates. In both cases the L3 sample SRR1188186 was used as an out-group given this has an associated collection date. Maximum likelihood trees were constructed using RaXML-NG v0.9.0 [71], as previously described, and a significant temporal regression was obtained for both sub-sampled datasets (**Supplementary Figure S4**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

BEAST2 v2.6.0 [74] was run on both subsampled SNP alignments allowing for model averaging over possible choices of substitution models [75]. All models were run with either a relaxed or a strict prior on the evolutionary clock rate for three possible coalescent demographic models: exponential, constant and skyline. To speed up the convergence, the prior on the evolutionary clock rate was given as a uniform distribution (limits 0 to 10) with a starting value set to 10^{-7}. In each case, the MCMC chain was run for 500,000,000 iterations, with the first 10% discarded as burn-in and sampling trees every 10,000 chains. The convergence of the chain was inspected in Tracer 1.7 and through consideration of the ESS for all parameters (ESS>200). The best-fit model to the data for these runs was assessed through a path sampling analysis [76] specifying 100 steps, 4 million generations per step, alpha = 0.3, pre-burn-in = 1 million generations, burn-in for each step = 40%. For both datasets, the best supported strict clock model was a coalescent Bayesian skyline analysis. The rates (mean and 95% HPD) estimated under these subsampled analyses (L2 7.7×10^{-8} [4.9×10^{-8} - 1.03×10^{-7}] substitutions per site per year; L4 7.1×10^{-8} [6.2×10^{-8} - 7.9×10^{-8}] substitutions per site per year) were used to rescale the maximum likelihood phylogenetic trees generated across the entire L2 and L4 datasets. This resulted in an estimated tMRCA of 1332CE (945CE-1503CE) for L2 and 853CE (685CE – 967CE) for L4 (**Figure 2**).

The resulting phylogenetic trees were visualised and annotated for place of geographic sampling and *Rv0678* variant status using ggtree v1.14.6 [77]. All nonsynonymous mutations in *Rv0678* were considered, with the phenotypic status assigned in **Supplementary Table S4**. For the purpose of this analysis, and to be conservative, 'unknown' variants classified using XGBoost were still considered 'unknown'. Clades carrying shared variants in *Rv0678* were identified and the age of the node (point estimates and 95% HPDs) extracted from the time-stamped phylogeny using the R package Ape v5.3 [78]. For isolated samples (single emergences) exhibiting variants in *Rv0678*, the time of sample collection was extracted together with the date associated with the upper bound on the age of the next closest node of the tree (**Figure 3, Supplementary Figures S11-S12**). For the oldest bedaquiline resistance clade, which comprised Ile67fs carriers predominately from Peru, Bayesian skyline analysis was implemented through the skylineplot analysis functionality available in Ape v5.3 [78].

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Data availability

Raw sequence data and full metadata for all newly generated isolates are available on NCBI Sequencing Read Archive under BioProject ID: XXXXX.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Footnotes

Author Contributions

LvD, CN and FB conceived and designed the study. JM, NP, AG, MO, AP, OBB, VE and LG provided sequence data. ATO, JP, MA, CCST and XD performed and advised on computational analyses. LvD, CN and FB wrote the manuscript with input from all co-authors. All authors read and approved the final manuscript.

Acknowledgments

CN and JM are supported by the Wellcome Trust (203583/Z/16/Z and 203919/Z/16/Z, respectively). LvD and FB acknowledge financial support from the Newton Trust UK-China NSFC initiative (MRC grant MR/P007597/1) and a Wellcome Institutional Strategic Support Fund (ISSF3) – AI in Healthcare (19RX03). 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. World Health Organization. *Global Tuberculosis Report*. (2019).
2. Adepoju, P. Tuberculosis and HIV responses threatened by COVID-19. *lancet. HIV* **7**, e319–e320 (2020).
3. Cegielski, J. P. *et al.* Multidrug-Resistant Tuberculosis Treatment Outcomes in Relation to Treatment and Initial Versus Acquired Second-Line Drug Resistance. *Clin. Infect. Dis.* **62**, 418–430 (2015).
4. Andries, K. *et al.* A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. *Science (80-.).* **307**, 223–227 (2005).
5. Diacon, A. H. *et al.* Multidrug-Resistant Tuberculosis and Culture Conversion with Bedaquiline. *N. Engl. J. Med.* **371**, 723–732 (2014).
6. *Sirturo (bedaquiline) product insert*.
7. Borisov, S. E. *et al.* Effectiveness and safety of bedaquilinecontaining regimens in the treatment of MDR- and XDR-TB: A multicentre study. *Eur. Respir. J.* **49**, (2017).
8. Guglielmetti, L. *et al.* Long-term outcome and safety of prolonged bedaquiline treatment for multidrug-resistant tuberculosis. *Eur. Respir. J.* **49**, (2017).
9. Olayanju, O. *et al.* Long-term bedaquiline-related treatment outcomes in patients with extensively drug-resistant tuberculosis from South Africa. *Eur. Respir. J.* **51**, 1800544 (2018).
10. Ndjeka, N. *et al.* High treatment success rate for multidrug-resistant and extensively drug-resistant tuberculosis using a bedaquiline-containing treatment regimen. *Eur. Respir. J.* **52**, (2018).
11. World Health Organization. *WHO consolidated guidelines on drug-resistant tuberculosis treatment*. (2019).
12. V, E. & F, B. Antimicrobial Resistance in Mycobacterium Tuberculosis: The Odd One Out. *Trends Microbiol.* **24**, (2016).
13. Brynildsrud, O. B. *et al.* Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. *Sci. Adv.* **4**, (2018).
14. Merker, M. *et al.* Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nat. Genet.* **47**, 242–249 (2015).
15. Rutaihwa, L. K. *et al.* Multiple introductions of Mycobacterium tuberculosis Lineage 2-Beijing into Africa over centuries. *Front. Ecol. Evol.* **7**, (2019).
16. 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. *The Lancet Microbe* **1**, e165–e174 (2020).
17. Manson, A. L. *et al.* Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nat. Genet.* **49**, 395–402 (2017).
18. Cohen, K. A. *et al.* Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLoS Med.* **12**, (2015).
19. 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).
20. 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).
21. Andries, K. *et al.* Acquired resistance of Mycobacterium tuberculosis to bedaquiline. *PLoS One* **9**, (2014).
22. 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).
23. Bloemberg, G. V., Gagneux, S. & Böttger, E. C. Acquired resistance to bedaquiline and delamanid in therapy for tuberculosis: To the editor. *N. Engl. J. Med.* **373**, 1986–1988 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

24. Xu, J. *et al.* Primary Clofazimine and Bedaquiline Resistance among Isolates from Patients with Multidrug-Resistant Tuberculosis. *Antimicrob. Agents Chemother.* **61**, e00239-17 (2017).
25. 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).
26. de Vos, M. *et al.* Bedaquiline Microheteroresistance after Cessation of Tuberculosis Treatment. *N. Engl. J. Med.* **380**, 2178–2180 (2019).
27. Ghodousi, A. *et al.* Acquisition of Cross-Resistance to Bedaquiline and Clofazimine following Treatment for Tuberculosis in Pakistan. *Antimicrob. Agents Chemother.* **63**, (2019).
28. 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).
29. 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.* (2019). doi:10.1016/j.cmi.2019.06.014
30. 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.* (2020). doi:10.1093/jac/dkaa136
31. Nimmo, C. *et al.* Bedaquiline resistance in drug-resistant tuberculosis HIV co-infected patients. *Eur. Respir. J.* (2020). doi:10.1183/13993003.02383-2019
32. Martinez, E. *et al.* Mutations associated with in vitro resistance to bedaquiline in Mycobacterium tuberculosis isolates in Australia. *Tuberculosis (Edinb).* **111**, 31–34 (2018).
33. 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).
34. 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).
35. Nimmo, C. *et al.* Dynamics of within-host Mycobacterium tuberculosis diversity and heteroresistance during treatment. *EBioMedicine* **55**, (2020).
36. O'Neill, M. B. *et al.* Lineage specific histories of *Mycobacterium tuberculosis* dispersal in Africa and Eurasia. *Mol. Ecol.* **28**, mec.15120 (2019).
37. Brynildsrud, O. B. *et al.* Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. *Sci. Adv.* **4**, eaat5869 (2018).
38. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6**, 10063 (2015).
39. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* **11**, (2019).
40. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).
41. Menardo, F., Duchêne, S., Brites, D. & Gagneux, S. The molecular clock of Mycobacterium tuberculosis. *PLOS Pathog.* **15**, e1008067 (2019).
42. Merker, M. *et al.* Phylogenetically informative mutations in genes implicated in antibiotic resistance in Mycobacterium tuberculosis complex. *Genome Med.* **12**, 27 (2020).
43. 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).
44. Andres, S. *et al.* Bedaquiline-resistant Tuberculosis: Dark Clouds on the Horizon. *Am. J. Respir. Crit. Care Med.* rccm.201909-1819LE (2020). doi:10.1164/rccm.201909-1819LE
45. World Health Organization. *Technical report on critical concentrations for TB drug susceptibility testing of medicines used in the treatment of drug-resistant TB*. (2018).
46. 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.* (2020). doi:10.1093/cid/ciaa1002
47. Pym, A. S. *et al.* Bedaquiline in the treatment of multidrug- and extensively drug-resistant tuberculosis. *Eur. Respir. J.* **47**, 564–74 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

48. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, (2014).
49. Sobkowiak, B. *et al.* Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. *BMC Genomics* **19**, (2018).
50. 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).
51. 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).
52. Norheim, G. *et al.* Tuberculosis outbreak in an educational institution in Norway. *J. Clin. Microbiol.* **55**, 1327–1333 (2017).
53. Nimmo, C. *et al.* Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. *BMC Genomics* **20**, (2019).
54. 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).
55. 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).
56. Guerra-Assunção, J. A. *et al.* Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *Elife* **2015**, (2015).
57. Grandjean, L. *et al.* Transmission of Multidrug-Resistant and Drug-Susceptible Tuberculosis within Households: A Prospective Cohort Study. *PLoS Med.* **12**, (2015).
58. Grandjean, L. *et al.* Convergent evolution and topologically disruptive polymorphisms among multidrug-resistant tuberculosis in Peru. *PLoS One* **12**, e0189838 (2017).
59. 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 Br.* **20**, 1975–1983 (2018).
60. Ghajavand, H. *et al.* High prevalence of bedaquiline resistance in treatment-naive tuberculosis patients and verapamil effectiveness. *Antimicrob. Agents Chemother.* **63**, e02530-18 (2019).
61. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
62. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013). doi:10.1002/0471250953.bi1110s43
63. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, btv566 (2015).
64. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).
65. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7**, 51 (2015).
66. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-August-2016**, 785–794 (Association for Computing Machinery, 2016).
67. JANIN, J. Surface and inside volumes in globular proteins. *Nature* **277**, 491–492 (1979).
68. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One* **10**, e0118432 (2015).
69. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
70. Keane, J. A. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2**, (2016).
71. 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* (2019). doi:10.1093/bioinformatics/btz305
72. 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–e134 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

73. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).
74. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 1–8 (2007).
75. Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* **17**, 42 (2017).
76. 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).
77. Yu, G., Smith, D. K., Zhu, H., 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).
78. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

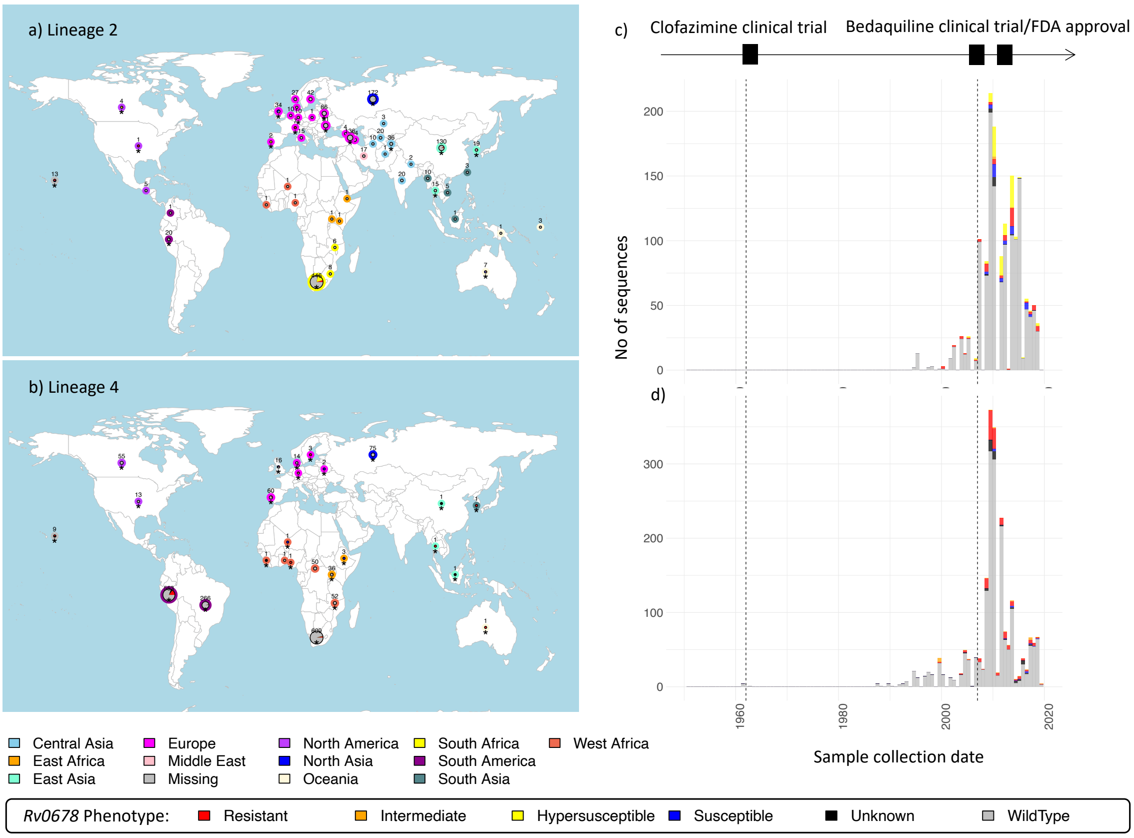

Figure 1: Compiled global *Mtb* genomic datasets.

Panels a) and b) provide the geographic location of isolates included in the lineage 2 and lineage 4 datasets respectively. Pies are scaled by the number of samples (given above each pie) with the colour of the inner pie providing the fraction of samples with any variants in *Rv0678* (coloured as per the legend at bottom), as highlighted with an asterisk. Samples without associated metadata on the geographic location of sampling are shown in the Pacific Ocean with a grey outer ring. Coloured outer rings provide the geographic location as given in the legend at bottom. c) and d) provide the collection dates associated with each sample in the lineage 2 and lineage 4 datasets respectively highlighting those with any variants in *Rv0678* (colour). Lineage 4 *Mtb* obtained from 18th century mummies are excluded from this plot but included in all analyses. The timeline at top indicates the dates of the first clofazimine and bedaquiline and clinical trials and FDA approval.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

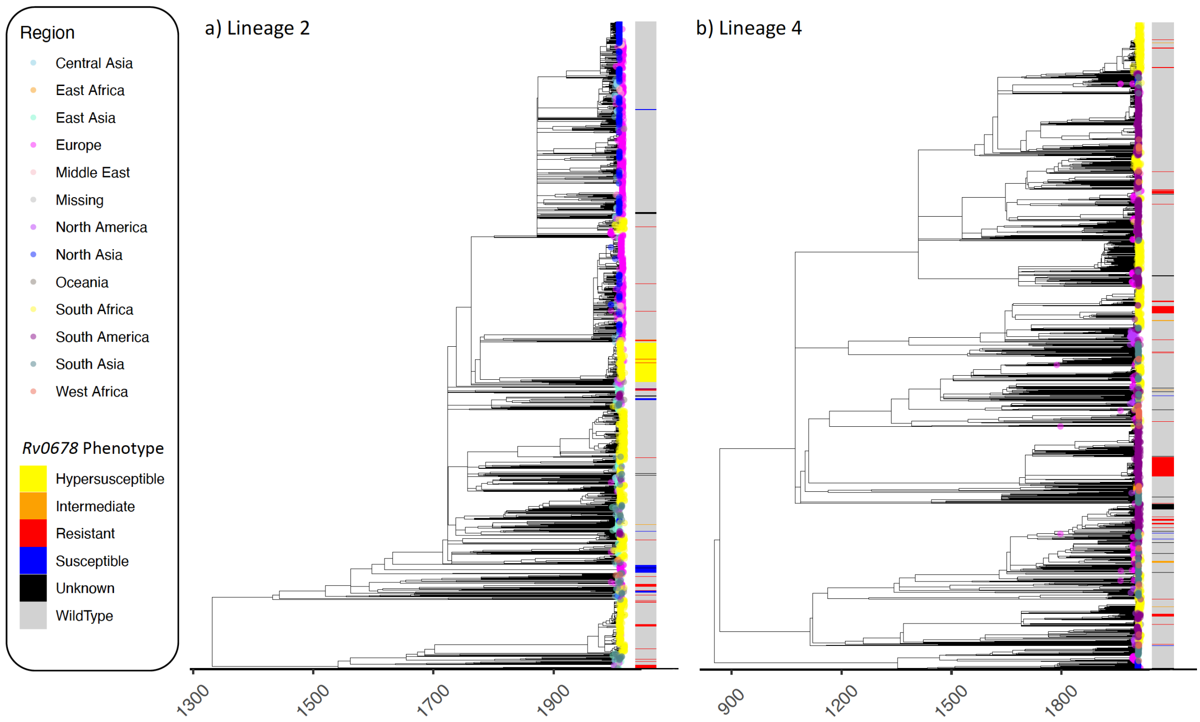

Figure 2: Global time calibrated *Mtb* phylogenies.

Inferred dated phylogenies (x-axis) for the a) lineage 2 and b) lineage 4 datasets. Tips are coloured by the geographic region of sampling as given in the legend. The bar provides the *Rv0678* phenotype (colour) based on assignment of nonsynonymous variants in *Rv0678*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.06.328799; this version posted October 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

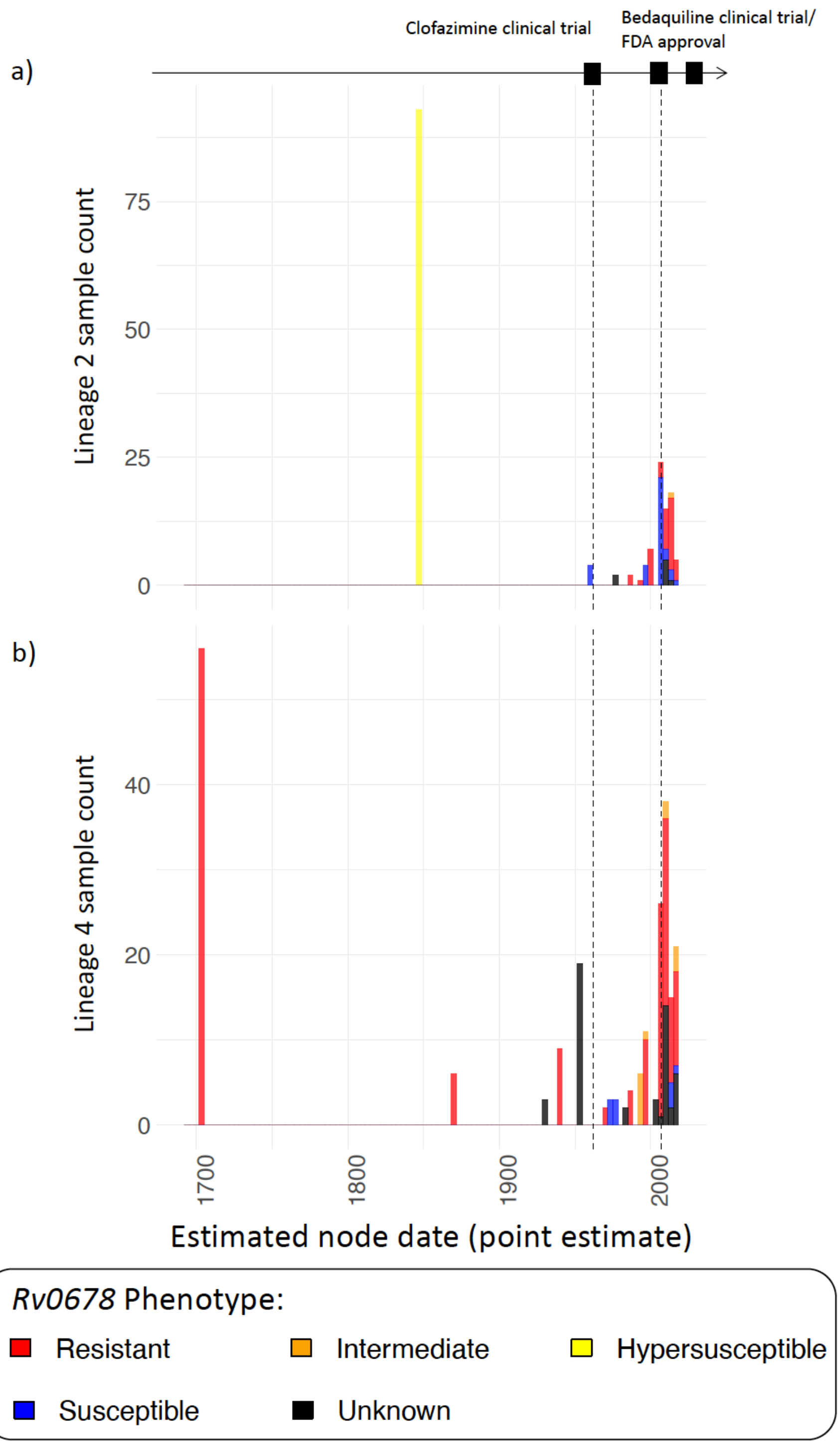

Figure 3: Estimated age of emergence of *Rv0678* nonsynonymous variants.

Inferred point estimates for the dates of clades with *Rv0678* variants for the lineage 2 (a) and lineage 4 (b) datasets. Predicted *Rv0678* phenotype is given by the colour as defined in the legend at bottom. The full mutation timelines are provided in **Supplementary Figures 11-12** and **Supplementary Table S8**.