Deciphering Bedaquiline and Clofazimine Resistance in Tuberculosis: An Evolutionary Medicine Approach

Lindsay Sonnenkalb[+,1], Joshua Carter[+,2], Andrea Spitaleri[5,6], Zamin Iqbal[7], Martin Hunt[7,8], Kerri Malone[7], Christian Utpatel[1], Daniela Maria Cirillo[5], Camilla Rodrigues[9], Kayzad S. Nilgiriwala[10], the CRyPTIC Consortium[11], Philip W. Fowler[&, 8], Matthias Merker[&,1], Stefan Niemann[&,1,3,4]

Affiliations

[1]Molecular and Experimental Mycobacteriology, Research Center Borstel Leibniz Lung Center, Borstel Germany

[2]Medical Scientist Training Program, Stanford University, Stanford, USA

[3]National Reference Center, Research Center Borstel Leibniz Lung Center, Borstel Germany

[4]German Centre for Infection Research (DZIF), Partner site Hamburg-Lübeck-Borstel, Borstel, Germany

[5]Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy

[6]Center for Omics Sciences, IRCCS San Raffaele Scientific Institute, Milan, Italy

[7]European Bioinformatics Institute, Cambridge, UK.

[8]Experimental Medicine, Nuffield Department of Medicine, University of Oxford, UK

[9]Department of Microbiology, P. D. Hinduja National Hospital and Medical Research Centre, Mumbai, India

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

[11]Comprehensive Resistance Prediction for Tuberculosis: an International Consortium. Author list in supplement.

[+]equally contributing first authors

[&]equally contributing last authors

*Corresponding author, email: sniemann@fz-borstel.de

Abstract

Bedaquiline (BDQ) and clofazimine (CFZ) are core drugs for treatment of multidrug resistant tuberculosis (MDR-TB), however, our understanding of the resistance mechanisms for these drugs is sparse which is hampering rapid molecular diagnostics. To address this, we employed a unique approach using experimental evolution, protein modelling, genome sequencing, and minimum inhibitory concentration data (MIC) combined with genomes from a global strain collection of over 14,151 *Mycobacterium tuberculosis* complex isolates. Overall, 189 genomic variants causing elevated BDQ and/or CFZ MICs could be discerned, with 175 (95.1%) variants affecting the transcriptional repressor (Rv0678) of an efflux system (MmpS5-MmpL5). Structural modelling of Rv0678 suggests four major mechanisms that confer resistance: impairment of DNA binding, reduction in protein stability, disruption of protein dimerization, and reduction in affinity for its fatty acid ligand. These modelling and experimental techniques will improve personalized medicine in an impending drug resistant era.

40 **Introduction**

41 Multidrug-resistant (MDR: resistance to at least isoniazid [INH] and rifampicin [RMP]) Mycobacterium
42 tuberculosis complex (Mtbc) strains represent a serious challenge for global tuberculosis (TB) control [1,2].
43 The World Health Organization (WHO) estimates that close to half a million people worldwide were
44 infected in 2019 with an RMP-resistant (RR) Mtbc strain, of whom 78 % were MDR [1].

45 Compared to patients with sensitive TB, treating MDR-TB treatment takes longer (years), the drugs are
46 less effective and more toxic, and cure rates are low (about 60 % globally) [1,3]. As a consequence, ineffective
47 MDR-TB regimens in many high incidence settings have led to an expansion of drug resistant Mtbc strains.
48 The WHO has recommended new treatment regimens for MDR-TB, including new and repurposed drugs
49 such as bedaquiline (BDQ) and clofazimine (CFZ) [4–7]. Both drugs are now central to MDR-TB therapy, and
50 are also part of the recently WHO-endorsed shorter, all-oral MDR-TB regimen [7,8].

51 While initial studies indicated that resistance to BDQ was mediated by mutations in *atpE*, *pepQ*, and
52 *Rv1979c* [9–11], clinical studies investigating patients with failing BDQ-containing regimens, revealed a
53 different mechanism involving the gene Rv0678 [12–17]. Rv0678 is a MarR-like transcriptional regulator of
54 the MmpS5-MmpL5 efflux pump with a canonical N-terminal winged helix-turn-helix (wHTH) DNA binding
55 domain and a C-terminal helical dimerization domain. The Rv0678 protein forms a homodimer which
56 binds to the promotor region of the MmpS5-MmpL5 operon [18–21]. This repression is reversed upon the
57 conformational change of the Rv0678 molecule in the presence of a fatty acid ligand [20]. Importantly, Mtbc
58 strains with mutations in Rv0678 have been shown to confer resistance against both BDQ and CFZ [22,23],
59 thus undermining the recently WHO-endorsed treatment guidelines [8]. Moreover, mutations in Rv0678 are
60 the most common resistance mechanism to CFZ and BDQ found in patient isolates today [16,17,24–29].

61 So far, mutations scattered across the full gene length of *Rv0678* (498bp) have been found to occur *in*
62 *vitro* and in vivo leading to variable shifts in BDQ and CFZ minimum inhibitory concentrations (MICs) [13,29,30].
63 However, a comprehensive understanding of the associations between Rv0678 mutations and their
64 resulting phenotype, as well as their structural effects on the transcriptional repressor Rv0678, is lacking.

65 In this paper we present the results of an experimental model that evolves Mtbc strains under sub-lethal
66 drug concentrations. Such a weak selection pressure has been shown to select for diverse pheno- and
67 genotypes in other bacterial species and likely also reflects the physiological conditions in the granulomas

68 of TB patients [31–33]. Single resistant clones and total bacterial populations from consecutive cultivation
69 passages were selected at clinically relevant drug concentrations, analyzed by whole genome sequencing
70 (WGS), and compared with mutations identified in 14,151 patient isolates collected by the CRyPTIC
71 consortium. Selected mutations were further investigated for their putative effects on the Rv0678 protein
72 structure and function. The data generated provide a comprehensive resistance mutation catalogue for
73 genotypic BDQ/CFZ drug susceptibility testing and will improve MDR-TB treatment design and likely
74 outcomes for patients receiving a BDQ and/or CFZ containing therapies.

75 **Results**

76 **Experimental evolution under sub-lethal drug concentrations**

77 To better understand the resistance mechanisms of Mtbc strains against BDQ (which often includes
78 correlated resistance to CFZ), we evolved Mtbc strains under sub-lethal drug concentrations. The
79 experimental set-up was developed to allow for a high number of bacterial generations under a moderate
80 drug selection pressure, thus increasing the mutation diversity compared to classical mutant selection
81 experiments, i.e. exposing a culture at a late exponential phase to the critical drug concentrations (Figure
82 S1). The MICs of BDQ and CFZ for the laboratory reference strain H37Rv (ATCC 27294) were 0.25 mg/L
83 and 0.5 mg/L, respectively – we shall refer to these as the wild-type (WT) MICs. During five passages
84 spread over up to 20 days, we exposed H37Rv to either BDQ or CFZ at one of four concentrations ranging
85 from one-half to one-sixteenth of the WT MIC (Figure S2), and selected colonies on solid media (7H11
86 plates) supplemented with 0.12-0.25 mg/L for BDQ and 0.25 mg/L for CFZ. At passages 1, 3, and 5, single
87 colonies as well as the whole bacterial population were genotypically characterized, and single selected
88 mutations were phenotypically defined (elevated MIC) (Tables 1, S1).

89 Strikingly, exposure to BDQ or CFZ concentrations as low as one-eighth of the WT MIC over 20 days (12 to
90 19 generations) led to the selection and enrichment of significantly more resistant cells as compared to
91 growth in the absence of the drugs ($P=2.8 \times 10^{-10}$, and $P=1.7 \times 10^{-3}$, respectively, Figure S2). As expected,
92 there was less enrichment of resistant populations at early timepoints, e.g. after 4 days.

93 **Phenotypic and genotypic characterization of BDQ resistant mutants**

94 In total, we randomly selected 270 single colonies from BDQ- and CFZ-supplemented agar plates that had
95 been inoculated with samples from passages 1 (4 days), 3 (12 days), and 5 (20 days) (Figure S1). In total,
96 197/270 mutants were successfully pheno- and genotyped (75 were excluded due to DNA library
97 preparation issues or inadequate growth). The single colonies were sub-cultured and the MICs for BDQ
98 and CFZ were determined for each clone using broth microtiter dilution. All clones exhibited an elevated
99 MIC at least 2-fold higher than the WT drug susceptible ancestor and we identified in total 60 unique
100 clones with on mutation in the *Rv0678* gene per clones. We additionally selected four unique clones with
101 a mutation in the *atpE* (*Rv1305*) gene (Table 1).

102 **Table 1: *In vitro* selected mutations conferring BDQ resistance.** Genotype of each variant was determined
103 by whole genome sequencing, and substitution annotated by coding position, "alternative allele" denotes
104 base pair change. Minimum inhibitory concentration (MIC) was determined by resazurin microtiter plate
105 assay and compared to wild type (WT) susceptible ancestor to describe MIC fold increase. The number of
106 times a mutation was independently selected was included under "# of clones". Finally, from which
107 experiment (BDQ or CFZ) the mutant was isolated from included as "isolate from".

Gene	Position	Substitu-tion	Type	Alt. allele	# of clones	BDQ MIC (mg/L)	BDQ MIC fold increase	CFZ MIC (mg/L)	CFZ MIC fold increase	Isolated from
	WT ancestor					0.25-0.5		0.5-1		
Rv0678	778980	promotor fs	indel	-9 ins g	1	2	4-8	1	0-2	BDQ
Rv0678	779005	6fs	indel	16 del g	1	1	2-4	2	2-4	BDQ
Rv0678	779015	9fs	indel	26 del ag	1	2	4-8	NA	NA	BDQ
Rv0678	779029	14fs	indel	40 del c	1	2	4-8	2	2-4	BDQ
Rv0678	779038	17fs	indel	49 del a	1	2	4-8	2	2-4	CFZ
Rv0678	779051	21fs	indel	62 del a	1	2	4-8	1	0-2	BDQ
Rv0678	779060	G24V	SNP	T	3	2	4-8	2	2-4	CFZ
Rv0678	779062	G25S	SNP	A	1	2	4-8	NA	NA	BDQ
Rv0678	779078	30fs	indel	89 del g	1	2	4-8	2	2-4	CFZ

Rv0678	779101	R38!	SNP	T	4	2	4-8	2	2-4	CFZ
Rv0678	779102	R38P	SNP	C	1	1	2-4	2	2-4	CFZ
Rv0678	779108	L40S	SNP	C	1	1	2-4	2	2-4	CFZ
Rv0678	779109	40fs	indel	120 del g	1	1	2-4	2	2-4	CFZ
Rv0678	779117	L43P	SNP	C	5	1-2	2-8	2	2-4	BDQ/CFZ
Rv0678	779117	L43R	SNP	G	1	1	2-4	2	2-4	BDQ
Rv0678	779120	L44P	SNP	C	4	1	2-4	2	2-4	CFZ
Rv0678	779137	R50W	SNP	T	12	1	2-4	2	2-4	CFZ
Rv0678	779141	Q51R	SNP	G	2	2	4-8	4	4-8	CFZ
Rv0678	779149	E54!	SNP	T	1	2	4-8	2	2-4	CFZ
Rv0678	779152	E55!	SNP	T	1	1	2-4	2	2-4	CFZ
Rv0678	779159	A57E	SNP	A	1	2	4-8	2	2-4	CFZ
Rv0678	779174	A62D	SNP	A	1	2	4-8	2	2-4	CFZ
Rv0678	779176	S63G	SNP	G	5	2	4-8	2	2-4	CFZ
Rv0678	779181	64fs	indel	192 ins g	20	1-4	2-16	2-4	2-8	BDQ/CFZ
Rv0678	779182	65fs	indel	193 del g	12	1-2	2-8	2-4	2-8	BDQ/CFZ
Rv0678	779186	G66E	SNP	A	1	2	4-8	1	0-2	BDQ
Rv0678	779191	S68G	SNP	G	9	1-2	2-8	2	2-4	BDQ/CFZ
Rv0678	779203	R72W	SNP	T	2	1-2	2-8	2	2-4	CFZ
Rv0678	779204	R72L	SNP	T	2	2-4	4-16	4-6	4-12	BDQ
Rv0678	779210	L74P	SNP	C	8	1-2	2-8	2-4	2-8	BDQ/CFZ
Rv0678	779217	76fs	indel	228 ins t	1	2	4-8	NA	NA	BDQ
Rv0678	779224	F79V	SNP	G	1	1	2-4	2	2-4	CFZ
Rv0678	779225	F79C	SNP	G	2	1	2-4	2	2-4	BDQ
Rv0678	779240	A84V	SNP	T	2	1-2	2-8	1-2	0-4	BDQ
Rv0678	779252	88fs	indel	263 ins t	1	2	4-8	1	0-2	BDQ
Rv0678	779263	92fs	indel	274 ins a	3	1	2-4	2	2-4	CFZ
Rv0678	779270	94fs	indel	281 del g	1	2	4-8	2	2-4	CFZ
Rv0678	779280	97fs	indel	291 ins a	2	2	4-8	2	2-4	CFZ
Rv0678	779281	98fs	indel	292 del g	1	2	4-8	2	2-4	BDQ
Rv0678	779285	A99V	SNP	T	16	1-2	2-8	1-2	0-4	BDQ/CFZ
Rv0678	779293	A102T	SNP	A	4	1-4	2-16	2-4	2-8	BDQ
Rv0678	779296	G103R	SNP	C	1	1	2-4	2	2-4	BDQ
Rv0678	779330	L114P	SNP	C	1	2	4-8	2	2-4	CFZ
Rv0678	779332	Q115!	SNP	T	1	2	4-8	2	2-4	BDQ

Rv0678	779342	A118D	SNP	C	1	1	2-4	2	2-4	CFZ
Rv0678	779349	120fs	indel	360 del g	7	1-4	2-16	2-4	2-8	BDQ/CFZ
Rv0678	779351	G121E	SNP	A	3	2	4-8	2-4	2-8	CFZ
Rv0678	779354	L122P	SNP	C	1	2	4-8	2	2-4	CFZ
Rv0678	779363	125fs	indel	374 del t	1	2	4-8	2	2-4	BDQ
Rv0678	779363	L125P	SNP	C	1	2	4-8	2	2-4	CFZ
Rv0678	779384	R132P	SNP	C	1	2	4-8	2	2-4	CFZ
Rv0678	779392	R135W	SNP	T	1	2	4-8	2	2-4	BDQ
Rv0678	779396	L136P	SNP	C	2	2	4-8	2-4	2-8	BDQ
Rv0678	779406	139fs	indel	417 ins CGGGATCTGTTGGCATATAT	2	1-2	2-8	2	2-4	BDQ
Rv0678	779414	142fs	indel	425 del t	1	2	4-8	2	2-4	CFZ
Rv0678	779415	142fs	indel	426 ins TTGGCATA	1	2	4-8	NA	NA	BDQ
Rv0678	779424	Y145!	SNP	G	6	1-2	2-8	2	2-4	BDQ/CFZ
Rv0678	779440	S151P	SNP	C	1	2	4-8	2	2-4	BDQ
Rv0678	779450	L154P	SNP	C	2	2	4-8	2	2-4	CFZ
Rv0678	779454	155fs	indel	465 ins c	2	1-2	2-8	2	2-4	CFZ
Rv0678	779455	R156!	SNP	T	3	2	4-8	2	2-4	BDQ/CFZ
Rv0678	-	-	inversion	-	3	2	4-8	2	2-4	BDQ
Rv1305	1461127	D28G	SNP	G	1	≥10	20	NA	NA	BDQ
Rv1305	1461127	D28V	SNP	T	3	≥8	16	NA	NA	BDQ
Rv1305	1461227	E61D	SNP	T	5	≥10	20	1-2	0-4	BDQ
Rv1305	1461231	A63P	SNP	C	7	≥10	20	1-2	0-4	BDQ

Alt- alternative, BDQ- bedaquiline, CFZ- clofazimine, del- deletion, fs- frameshift, ins- insertion, MIC- minimum inhibitory concentration, NA- not available, SNP- single nucleotide polymorphism, !-stop codon

Clones which harbored a mutation in *Rv0678* had a 2-4-fold MIC increase for BDQ, whereas mutations in *atpE*, exhibited 4 to over 10 times the MIC of the WT strain (Table 1). For 12 randomly selected clones with seven different variants we verified the results obtained in broth microtiter dilution plates using a Mycobacterium growth indicator tube (MGIT) assay. These experiments showed comparable MICs of the mutant clones with those obtained from the previous method (Table S2).

115 The 60 unique mutations in *Rv0678* affected 53 different codons (Table 1), out of which 33% (21) were
116 frameshift (fs) mutations, 55% (33) had one non-synonymous SNP, and 10% (6) a premature inserted stop
117 codon. The mutations were scattered over the entire sequence of *Rv0678* and no dominant cluster could
118 be identified (Figure 1a). However, the four mutations which were most frequently, selected were: 192
119 ins g (64fs), A99V, and 193 del g (65fs), and R50W; these mutations were independently selected 20, 16,
120 and 12 times, respectively (Table 1).

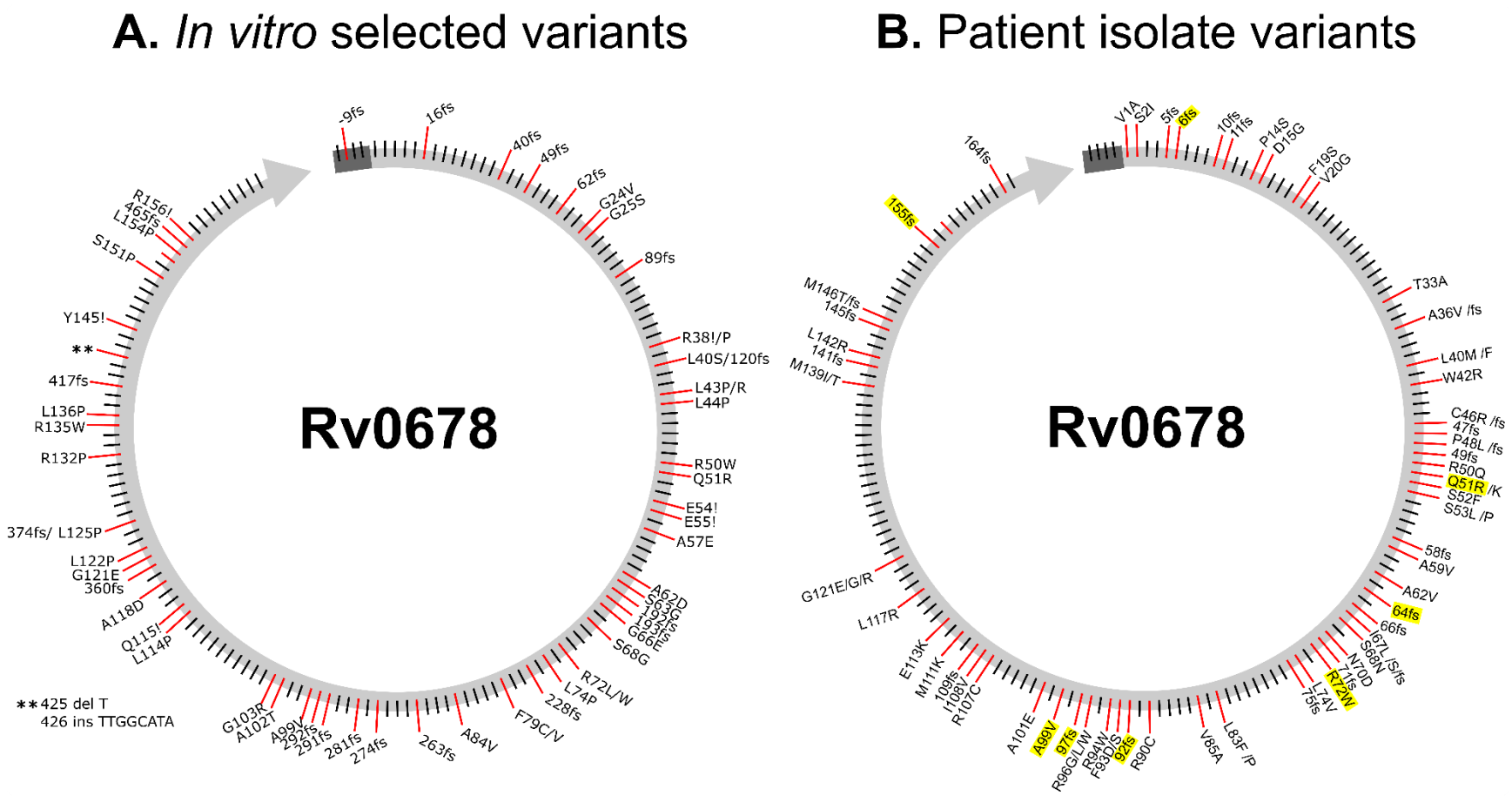


Figure 1: BDQ resistance associated variants mapped in *Rv0678* gene. Resistance associated mutations detected within patient dataset and *in vitro* selection were designated along Rv0678 gene at corresponding coding positions by non-synonymous single nucleotide polymorphisms, and by a protein disrupting mutation such as a stop codon insertion (!) or frameshift (fs) mutations. Data sets divided by (A) mutations which were selected in *in vitro* experiments with elevated minimum inhibitory concentrations verified to bedaquiline, and (B) mutations which were detected in patient isolates either collected by CRyPTIC or described in published literature. Overlapping mutations in observed in both patients and *in vitro* are highlighted in yellow.

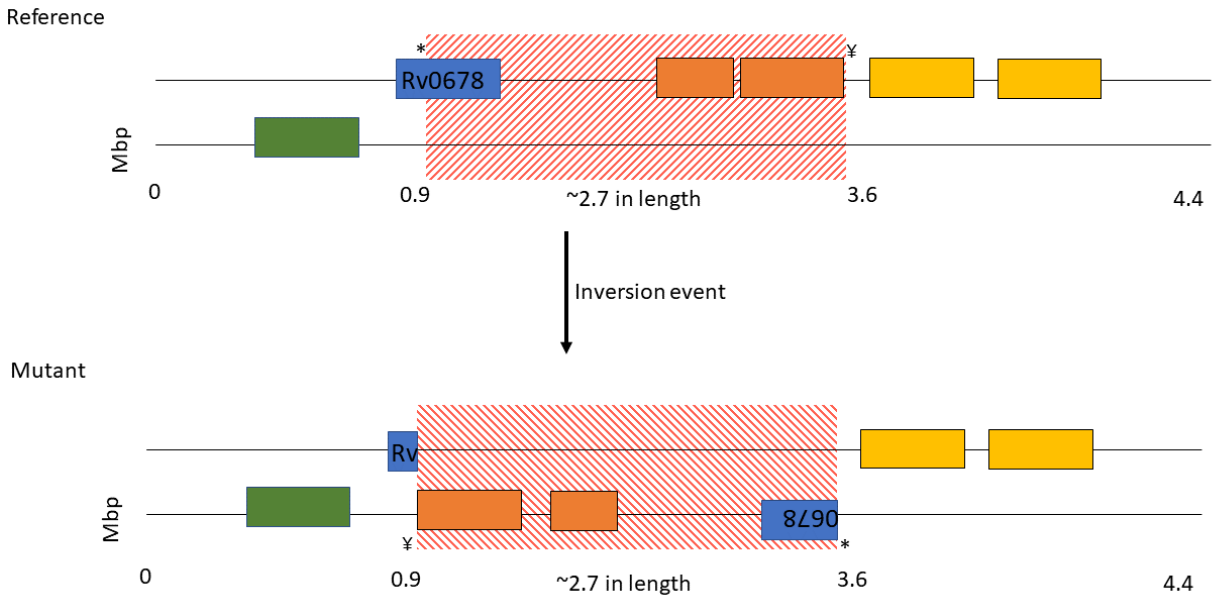
128 As mentioned before, no mutant clone had more than one resistance mediating variant in *Rv0678* or *atpE*,
129 however, 38 mutants harbored a second mutation in one of 14 genes that have not been previously
130 proposed to be associated with resistance against BDQ and/or CFZ (Table S3). The most common mutation
131 found in addition to *Rv0678* mutations was R119H in *Rv1890c* – this gene codes for a conserved
132 hypothetical protein [34], and the mutation was selected 11 times along with six different resistance
133 mutations. The second most common co-selected mutation was Stop130R in *Rv1871c* (a conserved
134 hypothetical membrane protein [35]) that was selected seven times along with four different resistance
135 mutations. Other secondary mutations occurred in genes which were involved in cell wall synthesis,
136 information pathways, metabolism/respiration, protein regulation, or lipid metabolism (Table S3).

137 In addition to single colony sequencing, we also employed a more unbiased approach based on population
138 sequencing of all the mutant colonies on a given plate, in order to elucidate all possible variants. On
139 average the genome wide coverage for a total of 81 samples was 308x (SEM +-50). We then reported
140 mutations if the position was covered by at least one read in both forward and reverse orientations and
141 there were two reads with a phred score over 20. This identified 45 additional mutations in *Rv0678* and
142 one in *atpE* (A63T), as well as four mutations in *pepQ* (Rv2535c; F97V, G96G, V92G, A87G) (Table S1). As
143 in the single colony analysis, we found the following mutations dominating in different independent
144 evolutionary experiments and selected across different drug concentrations: *Rv0678* 192 ins g (64fs),
145 A99V, R50W, and L43P.

146 **Resistance caused by huge inversion interruption in Rv0678**

147 Three mutants with elevated BDQ (2 mg/L) and CFZ (2 mg/L) MICs but lacking mutations in the BDQ/CFZ
148 resistance associated genes were further subjected to long read sequencing with the PacBio Sequel II
149 system. A *de novo* assembly employing the PacBio smrtlink software resulted in one closed Mtbc genome,
150 and two assemblies with 1 and 4 four contiguous sequences (contigs), respectively. All assemblies covered
151 more than 99.9% of the H37Rv reference genome (NC_000962.3). All three assemblies showed a large-
152 scale sequence re-arrangement with the borders at position 779,073 (Rv0678 coding sequence) and
153 3,552,584 (intergenic region), and flanked by a transposase open reading frame (IS*6110*). A 2.5Mb
154 fragment was inverted and thus split the *Rv0678* gene into halves (Figure 2), which was not detected by a
155 classical reference mapping approach, and revealed a further interesting resistance mechanism. We
156 scanned all BDQ resistant samples in the CRyPTIC strain collection for similar events, using local assembly

157 with ARIBA [36]. In one isolate we assembled one contig which clearly represented a similar inversion
158 disrupting *Rv0678*, but since the depth of support was just 1 read (mean, across the contig) and this was
159 Illumina (short read) data, we considered this only circumstantial evidence.



160

161 **Figure 2: Large scale gene rearrangement in Rv0678 gene.** *De novo* assembly using PacBio smrtlink
162 software indicated a 2.7 Mb inversion (red) impacting *Rv0678* (blue), at positions 779,073 coding
163 sequence (amino acid 34 in Rv0678) and 3,552,584 intergenic region; flanked by transposase related
164 genes (orange and yellow).

165 **Bedaquiline and clofazimine mutations associated with resistance in clinical isolates**

166 Next, we aimed to comprehensively describe the BDQ and CFZ resistant variants occurring in patients, to
167 develop a catalogue of known resistance variants. We searched the literature for patient reports
168 describing BDQ resistance during MDR-TB therapy, and analyzed pheno- and genotypes of 14,151 clinical
169 Mtbc strains collected by the CRyPTIC consortium (Table 2). We identified 81 mutations throughout
170 *Rv0678* from Mtbc strains reported with high MICs for BDQ and CFZ, 35 of which were novel to the
171 CRyPTIC collection. Of these 35 CRyPTIC-derived mutations, for which lineage information about the Mtbc
172 strains was available, seven codons in *Rv0678* harbored mutations across multiple lineages, suggesting
173 homoplasy (Table S4). Eleven mutations from the patient-derived set were also shared with our *in vitro*
174 derived datasets (selected variants and population sequencing) (Figure 3A).

175 **Table 2: *Rv0678* mutations detected in patients (published literature and CRyPTIC collection).** Reference
176 catalog of *Rv0678* mutations which have been described in previously published patient isolates or
177 collected by the Comprehensive Resistance Prediction for Tuberculosis an International Consortium
178 (CRyPTIC) partners. Additionally, alternative (Alt.) allele information included, describes the variants at
179 the base pair position. Variants included in list if minimum inhibitory concentration (MIC) was at or above
180 0.25 mg/L to bedaquiline (BDQ). Clofazimine (CFZ) MIC. Wild-type ancestor MIC was included when
181 available and were derived from serial isolates from the same patient. MIC determined by different drug
182 susceptibility testing (DST) techniques depending on the study, abbreviated as: [A] BD-MGIT™, [B]
183 UKMYC5, [C] UKMYC6, [D] Alamar blue/resazurin assay, [E] 7H10 plates, [F] 7H11 plates

Gene	Substitu -tion	Alt Allele	Type	BDQ MIC (mg/L)	BDQ MIC (mg/L) Susceptible ancestor	CFZ MIC (mg/L)	CFZ MIC (mg/L) Susceptible ancestor	DST Method	Author
Rv0678	V1A	NA	SNP	6.4		>4		[A]	[24]
Rv0678	S2I	NA	SNP	0.73		4		[D]	[28]
Rv0678	5fs	15 del c	indel	0.5		0.25		[C]	CRyPTIC
Rv0678	6fs	18 del gg	indel	0.5		NA		[F]	[37]
Rv0678	6fs	16 del gg	indel	0.5		NA			[38]
Rv0678	10fs	29 del t	indel	0.25		0.12		[B]	CRyPTIC
Rv0678	11fs	32 del g	indel	0.06-0.25		NA		[F]	[39]
Rv0678	P14S	NA	SNP	0.25		0.5		[B]	CRyPTIC
Rv0678	D15G	NA	SNP	0.25		NA		[B]	CRyPTIC
Rv0678	F19S	NA	SNP	0.5		NA		[B]	CRyPTIC
Rv0678	V20G	NA	SNP	0.25-2	0.06-0.25	2	0.5	[F]-BDQ only, [A]	[40], [41], [29]
Rv0678	T33A	NA	SNP	0.5		2		[D]	[23]
Rv0678	A36V	NA	SNP	0.5		4		[D]	[23]
Rv0678	36fs	107 ins g	indel	0.5		0.25		[B]	CRyPTIC
Rv0678	L40M	NA	SNP	0.5		NA		[B]	CRyPTIC
Rv0678	L40F	NA	SNP	0.5		NA		[C]	CRyPTIC
Rv0678	W42R	NA	SNP	0.12-0.25		2		[B]/[E]/[D]	CRyPTIC, [14], [23]
Rv0678	46fs	136 ins g	indel	1	0.03	NA		[F]	[41]
Rv0678	46fs	137 ins g	indel	0.25		1		[C]	CRyPTIC
Rv0678	46fs	138 ins g	indel	0.5,2	0.03-0.75	4	0.5	[F]-BDQ only, [A]	CRyPTIC, [40], [41]

Rv0678	C46R	NA	SNP	0.25		NA		[B]	CRyPTIC
Rv0678	C46fs	NA	indel	1		2		[F]	[29]
Rv0678	47fs	139 ins g	indel	0.25		NA		[F]	[39]
Rv0678	47fs	141 ins c	indel	0.25-4	0.03.0.5	0.12-4	0.5	[C], [F], [A],[D]	CRyPTIC, [40], [23], [41]
Rv0678	47fs	140 ins g	indel	0.25		NA		[F]	[37]
Rv0678	D47fs	NA	indel	0.5		2		[F]	[29]
Rv0678	48fs	144 ins c	indel	0.25		NA		[F]	[39]
Rv0678	P48L	NA	SNP	0.25-0.5		NA		[B]	CRyPTIC
Rv0678	E49fs	NA	indel	0.03-0.25	0.03	0.25-2		[F]	[29]
Rv0678	R50Q	NA	SNP	0.25		NA		[B]	CRyPTIC
Rv0678	Q51K	NA	SNP	0.25		1		[B]	CRyPTIC
Rv0678	Q51R	NA	SNP	0.3		1		[B]	CRyPTIC
Rv0678	S52F	NA	SNP	0.48		NA		[E]	[14]
Rv0678	S53L	NA	SNP	0.25-0.4		2		[D]	[28]
Rv0678	S53P	NA	SNP	0.5		2-4		[D]	[28]
Rv0678	58fs	172 ins IS6110	indel	0.5		1		[D]	[23]
Rv0678	A59V	NA	SNP	0.48		NA		[E]	[14]
Rv0678	A62V	NA	SNP	0.48		NA		[E]	[14]
Rv0678	64fs	192 ins g	indel	0.25-4	0.03-0.5	0.12-4	1	[B], [F]-BDQ only, [A]	CRyPTIC, [40]
Rv0678	66fs	198 ins g	indel	0.24-1		NA		[E]	[14]
Rv0678	I67L	NA	SNP	0.5		0.12		[B]	CRyPTIC
Rv0678	I67S	NA	SNP	0.5		0.25		[C]	CRyPTIC
Rv0678	I67fs	NA	indel	0.25-0.5	0.06	2	0.5	[F]	[29]
Rv0678	S68N	NA	SNP	0.5		2		[B]	CRyPTIC
Rv0678	N70D	NA	SNP	0.25		NA		[C]	CRyPTIC
Rv0678	71fs	212 del c	indel	0.5		1		[D]	[23]
Rv0678	R72W	NA	SNP	0.25		>4		[B]	CRyPTIC
Rv0678	L74V	NA	SNP	0.25		0.12		[B]	CRyPTIC
Rv0678	75fs	224 ins a	indel	0.24		NA		[E]	[14]
Rv0678	L83F	NA	SNP	0.5		NA		[C]	CRyPTIC
Rv0678	L83P	NA	SNP	0.25		1		[C]	CRyPTIC
Rv0678	V85A	NA	SNP	0.12-0.25		0.25		[B], [F]	CRyPTIC, [39]

Rv0678	R90C	NA	SNP	0.3		0.16		[B]	CRyPTIC
Rv0678	92fs	274 ins a	indel	0.25-1		0.5		[B], [E]	CRyPTIC, [14]
Rv0678	F93D	NA	SNP	0.25		NA		[F]	[41]
Rv0678	F93S	NA	SNP	0.25		1		[F]	[29]
Rv0678	R94W	NA	SNP	2	0.25	4	0.5	[A]	[17]
Rv0678	R96W	NA	SNP	0.25		NA			[38]
Rv0678	R96G	NA	SNP	0.25		0.25		[B]	CRyPTIC
Rv0678	R96L	NA	SNP	0.25		NA		[C]	CRyPTIC
Rv0678	97fs	291 del c	indel	0.25		NA		[F]	[39]
Rv0678	A99V	NA	SNP	0.25		NA		[F]	[39]
Rv0678	A101E	NA	SNP	0.25		0.12		[B]	CRyPTIC
Rv0678	R107C	NA	SNP	0.25		NA		[C]	CRyPTIC
Rv0678	I108T	NA	SNP	0.25		0.25		[B]	CRyPTIC
Rv0678	109fs	325 ins g	indel	1		NA		[C]	CRyPTIC
Rv0678	M111K	NA	SNP	0.25		NA			[38]
Rv0678	E113K	NA	SNP	0.25		0.25		[B]	CRyPTIC
Rv0678	L117R	NA	SNP	0.25-1.5		0.12-4		[B], [D]	CRyPTIC, [28]
Rv0678	G121E	NA	SNP	0.25		NA		[F]	[39]
Rv0678	G121R	NA	SNP	0.5		NA		[B]	CRyPTIC
Rv0678	G121G	NA	SNP	0.25		0.12		[C]	CRyPTIC
Rv0678	M139T	NA	SNP	0.25	0.015	NA		[F]	[37]
Rv0678	M139I	NA	SNP	0.25		0.25		[B]	CRyPTIC
Rv0678	141fs	421 ins g	indel	0.25		NA		[F]	[39]
Rv0678	L142R	NA	SNP	0.25-1.0		NA		[F]	[39]
Rv0678	145fs	435 del t	indel	0.25		NA		[F]	[39]
Rv0678	M146T	NA	SNP	0.25-0.8		1.2		[C], [D]	CRyPTIC, [28]
Rv0678	146fs	438 ins t	indel	0.25	0.03	0.5	0.25	[F]	[41], [29]
Rv0678	155fs	464 ins gc	indel	0.5		1		[B]	CRyPTIC
Rv0678	164fs	492 ins ga	indel	0.25		NA		[B]	CRyPTIC
Rv1305	D28N	NA	SNP	0.12		NA		[F]	[39]
Rv1305	A63P	NA	SNP	>8		2		[A]	[17]
Rv1305	A63V	NA	SNP	1		NA		[F]	[39]

Alt- alternative, BDQ- bedaquiline, CFZ- clofazimine, DST- drug susceptibility testing, del- deletion, fs- frameshift, ins- insertion, MIC- minimum inhibitory concentration, NA-not available, SNP- single nucleotide polymorphism, !-stop codon

187 Fifty-five further mutations in Rv0678 were found to be associated with MICs less than or equal to 0.25
188 mg/L for BDQ. These mutations may be categorized as benign variants since their MIC is below the WHO-
189 recommended critical concentration and proposed epidemiological cut-off (Table S5) [42].

190 Plotting all 175 resistance-associated *Rv0678* mutations, including *in vitro* and patient data (Figure 1B) on
191 the gene sequence, reveals that they occur throughout the entire length of the gene with no single hot
192 spot. However, there are several regions where mutations are more likely to occur, i.e. codons 46-53 and
193 62-68 (Figure 3B).

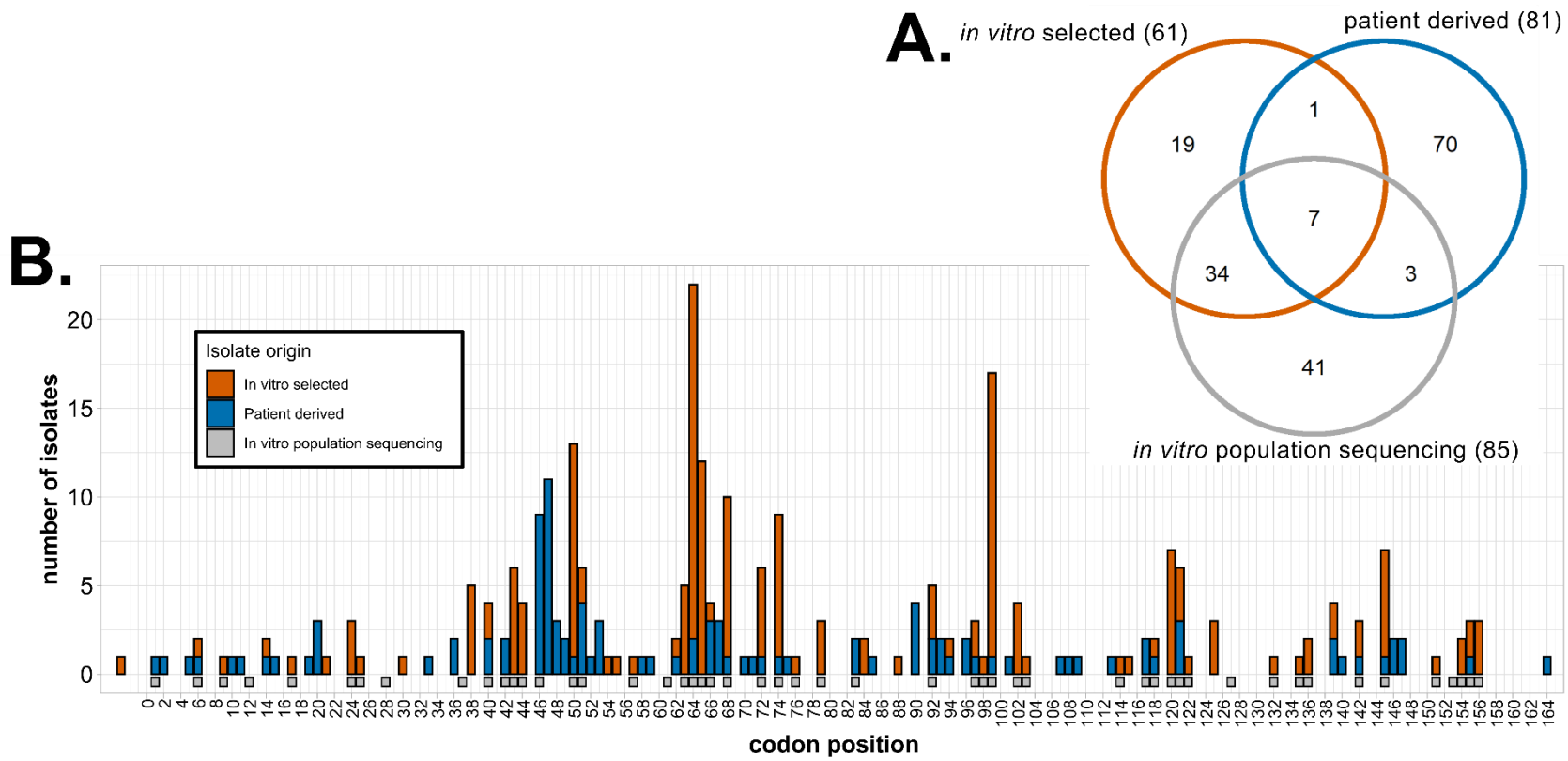


Figure 3: Mutations and coding position overlap throughout *Rv0678, in vitro* and patient datasets. Overlapping mutations which conferred BDQ resistance were compared between three datasets: *in vitro* selected mutants (Table 1), *in vitro* population sequencing (Table S1), and patient derived variants (Table 2). (A) Venn diagram describes number overlapping mutations in the three datasets, all mutations must have the same non-synonymous SNP or a deletion/insertion at same coding position to overlap. (B) Data represents number of mutants selected at given coding position for *in vitro* selected (orange) and patient derived variants (blue), indicated with gray box if detected in *in vitro* population sequencing.

200 **Structural variation in Rv0678**

201 In order to gain a mechanistic understanding on how *Rv0678* mutations result in BDQ/CFZ resistance, we
202 mapped the 175 mutations we identified to the previously determined Rv0678 crystal structure (PDB ID
203 4NB5) [20]. As mutations in different parts of the structure are likely to have different effects, we
204 hypothesized four different mechanisms for conferring resistance: impairing DNA binding, reducing folded
205 protein/protein stability, altering protein dimerization, and reducing affinity for its fatty acid ligand.

206 Direct interference with DNA binding appears to be the most common resistance mechanism for Rv0678
207 mutations. The cluster of mutations at amino acids 62-68 correlates with the recognition helix of the DNA
208 binding domain, which has previously been shown to be critical for binding affinity in other MarR-family
209 regulators [43,44]. To confirm the ability of these mutations to impact DNA binding, we structurally aligned
210 the wHTH DNA binding domain of Rv0678 to several other Mtbc MarR-family transcriptional regulators
211 whose DNA-bound structures have been determined (Figure 4B, S3; average RMSD 0.81 Å)[44]. Our modeled
212 DNA-bound Rv0678 structure confirms that mutations at positions 62-68 are poised to disrupt interactions
213 with the DNA and subsequent repression of the *mmpL5-mmpS5* operon (Figure 4B). Furthermore, the
214 structure also reveals the importance of residues 88-92, which constitute the wing of the DNA binding
215 domain that inserts its conserved DxR motif into the minor groove of the DNA (Figure 4B) [43,44].

216 Another mechanism of resistance appears to be loss of the functional folded protein as evidenced by the
217 numerous introductions of premature stop codons and frameshift mutations as reported here and
218 elsewhere[22,39,45]. Consistent with this, if we compare the missense *Rv0678* mutations in resistant isolates
219 with those in susceptible isolates, we find the former are more commonly predicted to destabilize Rv0678
220 (destabilization defined as $\Delta\Delta G <= -1$ kcal/mol as measured by mCSM protein stability, 40/49 R, 13/26 S,
221 OR 4.3, Fisher exact p=0.007), suggesting that disruption of protein stability is also resistance mechanism.
222 These findings are also compatible with the observation that more missense mutations from resistant
223 isolates are predicted to have a deleterious effect by SNAP2 score compared with those from susceptible
224 isolates (Wilcoxon p=0.002) [46].

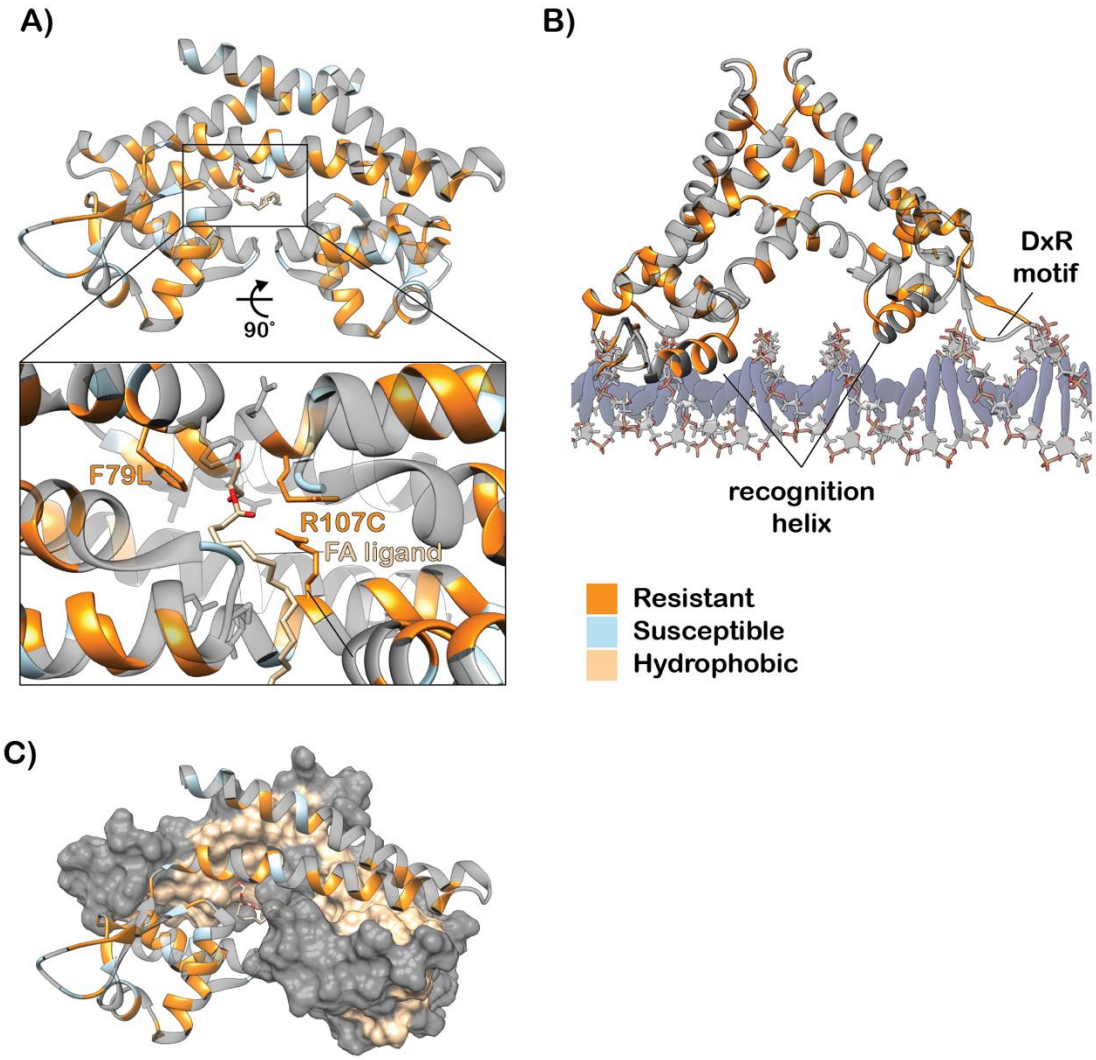


Figure 4: Structural mechanisms of BDQ resistance. (A) Rv0678 dimer with resistant (orange) and susceptible (blue) mutations shown. Pullout highlights the resistant mutations in the ligand binding pocket. (B) Rv0678 dimer modeled onto DNA. Conserved MarR-family DNA binding elements are noted. (C) Resistant mutations occur across the hydrophobic dimerization domain.

Another possibility is that mutations alter the stability of the ligand or DNA-bound conformations of the Rv0678 dimer. The dimer interface of ligand-bound Rv0678 is 2625 $Å^2$ and is primarily mediated through hydrophobic contacts (Figure 4C)[20]. Mutations associated with resistance occur all along the helices of the dimer interface, suggesting they may impact dimer formation and/or stability (Figure 4C). Indeed, a higher proportion of dimer interface mutations from resistant isolates are predicted to have a significantly destabilizing effect (mCSM PPI $\Delta\Delta G <= -1$ kcal/mol) on the ligand-bound Rv0678 conformation than those

236 from susceptible isolates (14/15 R, 8/16 S, OR 12.8, Fisher exact p=0.016). Interestingly, this is not the
237 case for the DNA-bound conformation, although precise prediction of the protein-protein interactions in
238 this region is complicated by the significant conformational change the C-terminal dimerization domain
239 undergoes to facilitate DNA binding (Figure 4B & S3). All isolates where an amino acid that contacts the
240 fatty acid ligand 2-stearoylglycerol was mutated were resistant, suggesting that disrupting the binding of
241 the ligand confers resistance, possibly through perturbing non-DNA bound homodimer formation (Figure
242 4A).

243 **Molecular dynamics simulations**

244 As Rv0678 must undergo a conformational rearrangement to bind DNA and this is not captured by the
245 structure above, we employed molecular dynamics simulations to understand how particular mutations
246 in the "hinge" region connecting the DNA binding and dimerization domains might disrupt this motion
247 and lead to resistance. To do this, we performed 100ns simulations on the WT Rv0678 homodimer as well
248 as Rv0678 homodimers containing either L40F and A101E mutations (which are associated with
249 resistance) or the L40V mutation (which is not). These simulations revealed that mutations from resistant
250 isolates (L40F and A101E) resulted in changes in the conformational dynamics and pocket crosstalk of the
251 individual Rv0678 monomers distinct from WT or the L40V mutation (Figure 5). In particular, A101E
252 formed a new stable salt-bridge interaction in the "hinge" domain, which could disrupt the conformational
253 flexibility required by this region to transition between the DNA- and ligand-bound homodimer states
254 (Figure S4).

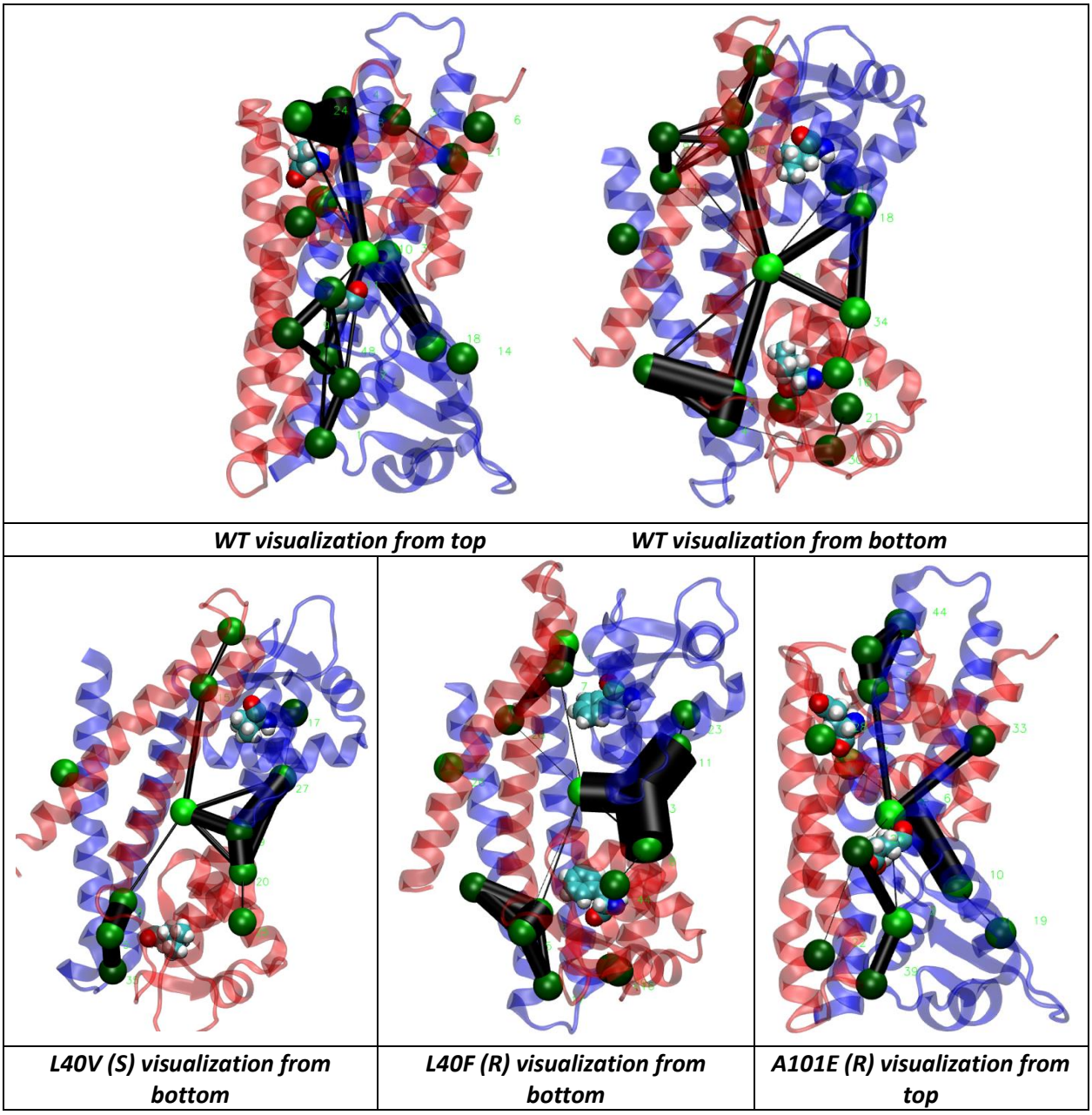


Figure 5: Networks of the most persistent pockets found in the wild-type (WT) and mutated Rv0678. The pocket crosstalk analysis on simulated systems identifies allosteric signaling. All systems shared a large common pocket located at the interface of the two monomers (Figure S5). This large pocket communicates with other smaller pockets at different zones of the homodimer by the network edges (black lines). Comparative analysis showed that the resistance-conferring mutations, L40F and A101E displayed different network edges during the 100ns of simulations with respect to the WT. R- resistant, S-susceptible

262 **Discussion**

263 We have combined *in vitro* evolution experiments not just with WGS and MIC data but also an analysis of
264 14,151 clinical Mtbc strains provided by the CRyPTIC consortium (http://www.crypticproject.org/). The
265 results have allowed us to define a comprehensive set of genomic variants associated with resistance to
266 BDQ (& CFZ), which are central to treating MDR, pre-XDR and XDR TB [8]. This project also demonstrates
267 the potential of *in vitro* evolution experiments employing selection under sub-lethal drug concentrations
268 to generate clones with a wide spectrum of mutations, including clinically relevant mutations.
269 Importantly, we find that sub-MIC drug concentrations effectively select for resistance, which, combined
270 with the discovery of previously existing variation conferring resistance in clinical isolates, suggests that
271 careful consideration of drug pharmacokinetics and strain genetic background should be used to inform
272 BDQ/CFZ dosing in MDR-TB regimens [28]. Collectively, this work helps extend the use of genomics in
273 providing personalized medicine to patients with MDR, pre-XDR, and XDR-TB, through bolstering
274 genotypic catalogues with phenotypically verified variant lists.

275 Although BDQ/CFZ are rolled out worldwide as core drugs in treatment regimens for MDR-TB patients [8],
276 antibiotic susceptibility testing for these compounds is rarely available and patients are often treated
277 empirically. Indeed, recent studies report treatment failure and/or poor outcome of MDR-TB patients as
278 mutations in *Rv0678* emerge that are associated with BDQ/CFZ resistance [12,15,17,26,41]. Additionally,
279 misdiagnosis can promote the selection and subsequent transmission of specific MDR Mtbc strains. For
280 example, it was recently reported in Eswatini and South Africa, that patients infected with an MDR
281 outbreak strain harboring the RMP mutation I491F, which is not detectable by conventional phenotypic
282 or genotypic testing, probably continued to receive RMP although it was ineffective [30]. These strains now
283 represent more than 60% of MDR strains in Eswatini and additionally more than half of these developed
284 enhanced BDQ/CFZ MICs via acquired *Rv0678* mutations [1,29,30,47]. This scenario underlines the urgent need
285 for approaches to rapidly detect resistance, including to BDQ/CFZ, with baseline diagnostics and during
286 therapy (e.g. via WGS or targeted genome sequencing) to avoid treatment failure, further resistance
287 development and ongoing transmission of specific MDR strains.

288 Genome sequencing techniques have been demonstrated to accurately identify genomic variants in all
289 target regions involved in resistance development in a single analysis [3,48]. However, one major challenge
290 is linking genotype and phenotype, and distinguishing resistance-mediating from benign mutations [49]. This

291 work catalogues the phenotype of 239 unique genomic variants. Indeed, 55 of these mutations in *Rv0678*
292 are not associated with resistance to BDQ/CFZ (Table S5), which facilitates the prediction of BDQ/CFZ
293 susceptibility. Importantly, *in vitro* laboratory evolution experiments identified 94 resistance-conferring
294 mutations that have not yet appeared in clinical strains, thus allowing us to characterize and identify a
295 significant proportion of resistance-conferring mutations before they have even occurred in a patient.
296 These results highlight the value of *in vitro* evolution experiments as a complementary method for
297 prospective identification of resistance, as compared to traditional surveillance of patient isolates. In
298 addition, by using PacBio sequencing to establish the full genome sequencing of mutant clones without
299 any mutation in the relevant targets, we were able to identify the largest inversion known in Mtbc strains
300 that disrupts *Rv0678* as novel BDQ/CFZ resistance mechanism.

301 Another method for prospectively identifying mutations that confer resistance is structure-based
302 modeling, which has recently been combined with machine learning methods to design algorithms for
303 predicting resistance to rifampicin, isoniazid, and pyrazinamide [50–54]. These approaches rely on the fact
304 that resistance-conferring mutations are often located in drug binding pockets or active sites and have
305 distinct biophysical consequences as compared to their susceptible counterparts [55]. This study identifies
306 several resistance mechanisms in *Rv0678* with quantifiable structural effects (protein stability, dimer
307 interactions, SNAP2 scores, and interaction with the DNA), suggesting that a structure-based machine
308 learning approach could also be successful for predicting BDQ/CFZ resistance.

309 Importantly, any approach to predict BDQ resistance must consider not only mutations present in *Rv0678*,
310 but also potential resistance determinants in other putative resistance-conferring genes e.g. *atpE*, *pepQ*,
311 *Rv1979c,* and *mmpL5*. It has been attempted to predict BDQ resistance from *atpE* mutations [56], so one
312 feasible option would be a meta-approach that flags if any individual gene-level algorithms predict
313 resistance. In addition, CRyPTIC and other studies are producing large MIC datasets, which will enable
314 much more nuanced modeling of resistance, which is particularly important for drugs with efflux-
315 mediated resistance that often exhibit small sub-threshold changes in MIC.

316 While genome sequencing of clinical Mtbc strains has been deployed in several countries, its
317 implementation is constrained not just by technical and data analysis considerations, but also because
318 working directly from clinical samples (e.g. sputum) is difficult [3]. We recently demonstrated that targeted
319 genome sequencing using the commercially available Deeplex®-MycTB has a high sensitivity and

320 specificity for first- and second-line drug resistance prediction, including BDQ/CFZ resistance mutations
321 that have been found in patients [47,57,58]. In combination with our comprehensive set of mutations that
322 confer resistance to BDQ/CFZ, Deeplex®-MycTB and other targeted sequencing approaches allow for rapid
323 detection of BDQ/CFZ resistance from clinical samples in few days using small portable genome
324 sequencers such as the MinION [59].

325 As indicated by the overlap of *in vitro* selected mutations with 8 resistance conferring mutations found in
326 patients, sub-MIC drug concentration can select for clinically relevant resistance variants. This modeling
327 might also immolate the intra-patient situation, since the complex pathology of TB disease prevents
328 certain drugs from reaching all bacteria at a lethal level (for example structures like pulmonary cavities
329 and granuloma lesions) [33,61–63]. BDQ and CFZ in particular, do not penetrate the central caseous lesion (or
330 necrotic center) of the granuloma effectively [64].

331 In addition to clear-cut resistance-associated mutations, additional "secondary" mutations were co-
332 selected with resistance-associated variants. These secondary mutations affected a number of genes in
333 the bacteria, mainly associated with cellular processes such as cell wall synthesis, metabolic processes,
334 and protein regulation; with mutations also in genes of unknown function (Table S3). It is possible that
335 these mutations bring about an additional phenotypic affect, or they may be compensatory mutations [5,65].
336 However, until further experiments are conducted, such as competitive fitness assays (or even
337 transcriptomic analysis), it cannot be ruled out that these are merely hitchhiking mutations or culture-
338 selected variants.

339 In conclusion, our work advances our understanding of how resistance can arise to BDQ and CFZ. This
340 information provides a comprehensive catalogue comprised of mutations and the large structural
341 variation associated with BDQ/CFZ resistance with additional variants considered to be benign. Employing
342 this information in DNA sequencing-based diagnostic approaches will allow for the first time a rational
343 design of BDQ and CFZ -containing MDR-TB therapies.

344 ## Materials and Methods

345 ### *In vitro* evolution experiments

346 *In vitro* experiments were carried out with H37Rv strain (ATCC 27294). Bedaquiline was purchased from
347 Janssen-Cilag GmbH and clofazimine was ordered from Sigma (C8895-1G) both were reconstituted from
348 powder in DMSO and stored at -20°C.

349 Evolutionary experiments were conducted with the *Mycobacterium tuberculosis* complex (Mtbc) lab strain
350 H37Rv. The bacteria were cultured from frozen stocks and (pre)cultured at 37° in 7H9 medium,
351 supplemented with 0.2% glycerol, 10% OADC, and 0.05% Tween[80] (termed culture medium). At
352 exponential growth phase (between 0.3 and 0.6 optical density (OD) of 600nm), bacteria were transferred
353 to new culture medium with a final OD of 0.05 in 50 mL, additionally supplemented with either
354 bedaquiline (BDQ) or clofazimine (CFZ). Four concentrations below the minimum inhibitory concentration
355 (MIC) were included for each drug at 1:2, 1:4, 1:8, and 1:16 below the strain MIC (and an antibiotic-free
356 control). The MIC of the susceptible WT ancestor was determined to be 0.12 mg/L BDQ and 0.5 mg/L CFZ
357 in this system.

358 These experiments were carried out over 20 days with five culture passages total. Culture passages were
359 conducted every four days with a final culture OD of 0.05 transferred. After passages 1, 3, and 5; bacteria
360 were plated on selective agar plates of 7H11, supplemented with 0.5% glycerol, 10% OADC and either
361 0.12 or 0.25 mg/L of BDQ, or 0.5 mg/L CFZ. After 21-28 days of growth on selective agar plates, single
362 colonies were transferred to culture medium, grown for 2-4 weeks, then two 1 mL aliquots were frozen
363 at -80, and samples were taken for whole genome sequencing (WGS). All remaining colonies (collected by
364 experimental condition) were collected for WGS, i.e. total population sequencing.

365 ### Whole genome sequencing

366 Isolates from *in vitro* evolutionary experiments underwent DNA isolation by CTAB method [66]. Paired-end
367 DNA library preparations and sequencing was performed with Illumina technology (Nextera-XT and
368 NextSeq500) according to the manufacturer's instructions and with a minimum average genome coverage
369 of 50x. Fastq files (raw sequencing data) were mapped to the *M. tuberculosis* H37Rv reference genome
370 (GenBank ID: NC_000962.3) using the MTBseq pipeline [67].

371 Briefly, for the analysis of single mutants we considered single nucleotide polymorphisms (SNPs) which
372 were covered by a minimum of four reads in both forward and reverse orientation, four reads calling the
373 allele with at least a phred score of 20, and an allele frequency of 75%. For the detection of insertions and
374 deletion (InDels) we allowed a frequency over 50%.

375 Deep sequencing was conducted for population diversity analysis with an average genome wide coverage
376 of 200 – 420; SNPs and InDels were called with at least one forward and one reverse read and a phred
377 score >20 for at least 2 reads.

378 Three isolates underwent sequencing using the PacBio Sequel II System (Pacific Biosciences). Libraries
379 were prepared with the SMRTbell® Express Template Prep Kit 2.0 according to the manufacturer's
380 instructions. Barcoded overhang adapters for multiplexing were ordered at IDT (Integrated DNA
381 Technologies). During demultiplexing barcodes were filtered for a minimum quality of 50t, which yielded
382 long read sequencing data of an average of 4.4 GB and mean subread length of 9 KB. Long read sequences
383 were *de novo* assembled using PacBio SMRT Link software version 9 and the "Microbial Assembly"
384 workflow with a set genome size of 4.5 GB with default parameters.

385 **Screening of Rv0678 mutations in clinical samples via the CRyPTIC strain collection**

386 Strains were selected from the CRyPTIC database if they had matched genotype and phenotype data, a
387 HIGH quality BDQ phenotype, and a mutation in Rv0678. This led to the curation of 179 strains in total.

388 The full analysis pipeline for CRyPTIC is documented in [paper in preparation], but we outline the key steps
389 here. Sequencing reads were deposited at the European Bioinformatics Institute and run through a
390 bespoke bioinformatics pipeline (publicly available here: https://github.com/iqbal-lab-org/clockwork). In
391 short, reads were filtered against human and other microbial species reads before being mapped to the
392 H37Rv reference genome. Two parallel variant callers (SamTools and Cortex) were used [68,69], one of which
393 makes high sensitivity SNP calls (SamTools) and one which makes high specificity SNP and indel calls
394 (Cortex). A graph-based adjudication tool (Minos, https://github.com/iqbal-lab-org/minos) was then used
395 to combine these results and create a final set of variants for downstream bioinformatics analysis. All
396 strains were re-genotyped at positions that were variant in at least one CRyPTIC sample, creating a final
397 variant call file with a call for each variant position in the *M. tuberculosis* H37Rv genome.

398 **Phenotyping**

399 *In vitro* single selected mutants were further analyzed for phenotypic drug susceptibility to both BDQ and
400 CFZ by broth microtiter dilution as a resazurin assay. All mutants were grown in antibiotic free culture
401 media to exponential growth phase (optical density of 0.3-0.8), then diluted to the McFarland standard of
402 1, with an additional 1:10 dilution, 100 µl of which was seeded in 96-well flat bottom plates (about $1x10^5$
403 CFU per well). Next 100 µL of antibiotic was added to each well with final concentrations of BDQ as follows:
404 8, 4, 2, 1, 0.5, 0.25, 0.12, and 0 mg/L; and CFZ: 16, 8, 4, 2, 1, 0.5, 0.25, 0.12, and 0 mg/L. After nine days
405 incubation at 37°, 30µl of resazurin was added to each well. After overnight incubation, fluorescence and
406 absorbance were measure in Biotek plate reader Synergy 2. MIC was determined as the highest
407 concentration which no bacterial growth was detected, either visually or by fluorescence measurement.

408 CRyPTIC strains were phenotyped using either the UKMYC5 plate or the updated UKMYC6 plate [70]. Plates
409 were read at 14 days. In addition to manual plate readings, all plate images underwent an automated
410 reading using AMyGDA software [71]. Plates without essential agreement for a drug MIC were marked as
411 low quality and sent to a citizen science project (BashTheBug, http://bashthebug.net) for additional
412 verification. Plates with exact agreement between at least two phenotyping methods were marked as
413 HIGH quality. Low (no method agrees) and medium (at least two methods with essential agreement)
414 quality phenotypes were excluded from this analysis.

415 ***Rv0678* variant literature search**

416 An extensive search was performed to include and summarize previously published BDQ and/or CFZ
417 resistant associated mutations from patient derived isolates. We used PubMed, Google, and Google
418 Scholar to search literature published from 2014 to September 2020. Search criteria included the key "TB",
419 "Mycobacterium tuberculosis", "MTB", "bedaquiline", "clofazimine", "treatment", "clinical report",
420 "patient", "MDR-TB", "XDR-TB", "diarylquinoline", and "drug resistance".

421 All single mutations which were described in patients with a defined MIC to at least BDQ were collected
422 and organized into one dataset with CRyPTIC strains (Table 2).

423 **Structural modelling**

424 The crystal structure of Rv0678 (PDB ID: 4NB5) was visualized using UCSF Chimera [72]. Structural alignment
425 of the wHTH domain was performed using the MatchMaker tool in Chimera with a Needleman-Wunsch
426 algorithm using a BLOSUM-62 matrix, and was iteratively pruned until no long atom-pair was > 2 Å
427 resulting in a final average RMSD of 0.81 Å over the 5 guide structures (PDB IDs: 5HSO, 5HSM, 4FX0, 4FX4,
428 4YIF). All protein stability, protein-protein and protein-DNA interactions were modeled using established
429 mCSM methods with either ligand-bound or DNA-bound Rv0678 [73,74]. mCSM tools are available at
430 http://biosig.unimelb.edu.au/biosig.

431 **Molecular dynamics simulations**

432 All the systems simulated in the present work, including Rv0678-WT, Rv0678-A101E, Rv0678-L40V and
433 Rv0678-L40F, were prepared using BiKi Life Sciences Software Suite version 1.3.5 of BiKi Technologies s.r.l
434 [75]. Each simulated system consisted of Rv0678 homodimer unit X-ray structure (PDB ID 4NB5) and the
435 mutations were generated using UCSF Chimera software [72]. The Amber14 force field was used in all
436 molecular dynamic simulations. TIP3P waters were added to make an orthorhombic box. Adding a suitable
437 number of counter-ions neutralized the overall system. Then, the energy of the whole system was
438 minimized. Four consecutive equilibration steps were then performed: 1) 100 ps in the NVT ensemble at
439 100K with the protein backbone restrained (k=1000 kJ/mol nm^2), 2) 100 ps in the NVT ensemble at 200K
440 with the protein backbone similarly restrained, 3) 100 ps in the NVT ensemble at 300K with the protein
441 backbone restrained, and 4) 1000 ps in NPT ensemble at 300K with no restraints. For atoms less than
442 1.1nm apart, electrostatic forces were calculated directly; for atoms further apart electrostatics were
443 calculated using the Particle Mesh Ewald. Van der Waals forces were only calculated for atoms within 1.1
444 nm of one another. The temperature was held constant using the velocity rescale thermostat, which is a
445 modification of the Berendsen's coupling algorithm. Finally, simulations 100 ns long in the NPT ensemble
446 at 300K were performed for each system. To detect allosteric signal transmission networks across the
447 protein surface, defined as interconnected pocket motions, we carried out the allosteric communication
448 network analysis using the Pocketron module in BiKi Life Sciences Suite version 1.3.5 [75].

449

450

451 **CRyPTIC Ethics statements as at 22 Feb 2021.** Approval for CRyPTIC study was obtained by Taiwan Centers
452 for Disease Control IRB No. 106209, University of KwaZulu Natal Biomedical Research Ethics Committee
453 (UKZN BREC) (reference BE022/13) and University of Liverpool Central University Research Ethics
454 Committees (reference 2286), Institutional Research Ethics Committee (IREC) of The Foundation for
455 Medical Research, Mumbai (Ref nos. FMR/IEC/TB/01a/2015 and FMR/IEC/TB/01b/2015), Institutional
456 Review Board of P.D. Hinduja Hospital and Medical Research Centre, Mumbai (Ref no. 915-15-CR [MRC]),
457 scientific committee of the Adolfo Lutz Institute (CTC-IAL 47-J / 2017) and in the Ethics Committee
458 (CAAE: 81452517.1.0000.0059) and Ethics Committee review by Universidad Peruana Cayetano
459 Heredia (Lima, Peru) and LSHTM (London, UK).

460 **Authors' contributions.** LS, JC, AS, ZI, MH, KM, CU, DS, CR, KN, PF, MM and SN conceived the idea and
461 designed the study, and analyzed and interpreted the data. All authors contributed to obtaining and
462 assembling the data. LS, JC, AS, ZI, MM, PF and SN wrote the initial draft of the paper. All authors
463 contributed to data interpretation, final draft of the paper and approved the final version of the
464 manuscript.

465 **Competing interest.** Authors declare no competing interests.

466 **Acknowledgments.** We thank Dr. Musco Giovanna and Dr. Ballabio Federico (San Raffale Hospital) for
467 their help in the molecular dynamics set up. We would like to acknowledge the technical support at
468 Research Center Borstel, especially Vanessa Mohr, Tanja Niemann, Carina Hahn, Silvia Mass, and Doreen
469 Beyer for their aid in the laboratory.

470 **Funding.** All *in vitro* experimental work was financially supported though EvoLUNG: Evolutionary Medicine
471 of the Lung, a Leibniz science campus for evolutionary medicine research (http://evolung.fz-
472 borstel.de/). JC is funded by the Rhodes Trust and the Stanford University Medical Scientist Training
473 program. PWF is funded by the NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS
474 Foundation Trust, John Radcliffe Hospital, Oxford, UK. The views expressed are those of the author(s) and
475 not necessarily those of the NHS, the NIHR or the Department of Health. The work of the CRyPTIC
476 consortium was supported by the Bill & Melinda Gates Foundation [OPP1133541] and the Welcome Trust
477 [200205/Z/15/Z].

478 **References**

479 1. *WHO | Global tuberculosis report 2019*. http://www.who.int/tb/publications/global_report/en/
480 (2019).

481 2. Dheda, K. *et al.* The Lancet Respiratory Medicine Commission: 2019 update: epidemiology,
482 pathogenesis, transmission, diagnosis, and management of multidrug-resistant and incurable
483 tuberculosis. *The Lancet Respiratory Medicine* **7**, 820–826 (2019).

484 3. Dheda, K. *et al.* The epidemiology, pathogenesis, transmission, diagnosis, and management of
485 multidrug-resistant, extensively drug-resistant, and incurable tuberculosis. *The Lancet Respiratory*
486 *Medicine* **5**, 291–360 (2017).

487 4. Kendall, E. A., Fofana, M. O. & Dowdy, D. W. Burden of transmitted multidrug resistance in epidemics
488 of tuberculosis: a transmission modelling analysis. *The Lancet Respiratory Medicine* **3**, 963–972
489 (2015).

490 5. Merker, M. *et al.* Compensatory evolution drives multidrug-resistant tuberculosis in Central Asia. *eLife*
491 **7**, e38200 (2018).

492 6. Shah, N. S. *et al.* Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *New England*
493 *Journal of Medicine* **376**, 243–253 (2017).

494 7. WHO Consolidated Guidelines on Tuberculosis, Module 4: Treatment - Drug-Resistant Tuberculosis
495 Treatment. https://www.who.int/publications-detail-redirect/9789240007048.

496 8. WHO | Rapid Communication: Key changes to the treatment of drug-resistant tuberculosis. *WHO*
497 http://www.who.int/tb/publications/2019/rapid_communications_MDR/en/.

498 9. Huitric, E. *et al.* Rates and mechanisms of resistance development in Mycobacterium tuberculosis to
499 a novel diarylquinoline ATP synthase inhibitor. *Antimicrobial Agents and Chemotherapy* (2010)
500 doi:10.1128/AAC.01611-09.

501 10. Almeida, D. *et al.* Mutations in pepQ confer low-level resistance to bedaquiline and clofazimine in
502 Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy* (2016)
503 doi:10.1128/AAC.00753-16.

504 11. Zhang, S. *et al.* Identification of novel mutations associated with clofazimine resistance in
505 Mycobacterium tuberculosis. *Journal of Antimicrobial Chemotherapy* (2015) doi:10.1093/jac/dkv150.

506 12. de Vos, M. *et al.* Bedaquiline Microheteroresistance after Cessation of Tuberculosis Treatment. *New*
507 *England Journal of Medicine* (2019) doi:10.1056/nejmc1815121.

508 13. Ismail, N. A. *et al.* Defining Bedaquiline Susceptibility, Resistance, Cross-Resistance and Associated
509 Genetic Determinants: A Retrospective Cohort Study. *EBioMedicine* (2018)
510 doi:10.1016/j.ebiom.2018.01.005.

511 14. Villellas, C. *et al.* Unexpected high prevalence of resistance-associated Rv0678 variants in MDR-TB
512 patients without documented prior use of clofazimine or bedaquiline. *The Journal of antimicrobial*
513 *chemotherapy* (2017) doi:10.1093/jac/dkw502.

514 15. Kranzer, K. *et al.* New World Health Organization treatment recommendations for multidrug-resistant
515 tuberculosis: Are we well enough prepared? *American Journal of Respiratory and Critical Care*
516 *Medicine* (2019) doi:10.1164/rccm.201902-0260LE.

517 16. Polsfuss, S. *et al.* Emergence of Low-level Delamanid and Bedaquiline Resistance During Extremely
518 Drug-resistant Tuberculosis Treatment. *Clinical Infectious Diseases* **69**, 1229–1231 (2019).

519 17. Andres, S. *et al.* Bedaquiline-Resistant Tuberculosis: Dark Clouds on the Horizon. *Am J Respir Crit Care*
520 *Med* **201**, 1564–1568 (2020).

521 18. Islam, M. M. *et al.* Drug resistance mechanisms and novel drug targets for tuberculosis therapy.
522 *Journal of Genetics and Genomics* **44**, 21–37 (2017).

523 19. Milano, A. *et al.* Azole resistance in Mycobacterium tuberculosis is mediated by the MmpS5–MmpL5
524 efflux system. *Tuberculosis* **89**, 84–90 (2009).

525 20. Radhakrishnan, A. *et al.* Crystal Structure of the Transcriptional Regulator Rv0678 of Mycobacterium
526 tuberculosis*. *Journal of Biological Chemistry* **289**, 16526–16540 (2014).

527 21. Gao, Y.-R. *et al.* Structural analysis of the regulatory mechanism of MarR protein Rv2887 in M.
528 tuberculosis. *Sci Rep* **7**, 6471 (2017).

529 22. Hartkoorn, R. C., Uplekar, S. & Cole, S. T. Cross-Resistance between Clofazimine and Bedaquiline
530 through Upregulation of MmpL5 in Mycobacterium tuberculosis. *Antimicrobial Agents and*
531 *Chemotherapy* **58**, 2979–2981 (2014).

532 23. Andries, K. *et al.* Acquired Resistance of Mycobacterium tuberculosis to Bedaquiline. *PLoS ONE* **9**,
533 e102135 (2014).

534 24. Bloemberg, G. V. *et al.* Acquired Resistance to Bedaquiline and Delamanid in Therapy for Tuberculosis.
535 *N Engl J Med* **373**, 1986–1988 (2015).

536 25. Hoffmann, H. *et al.* Delamanid and Bedaquiline Resistance in Mycobacterium tuberculosis Ancestral
537 Beijing Genotype Causing Extensively Drug-Resistant Tuberculosis in a Tibetan Refugee. *Am J Respir*
538 *Crit Care Med* **193**, 337–340 (2016).

539 26. Liu, Y. *et al.* Reduced Susceptibility of Mycobacterium tuberculosis to Bedaquiline During
540 Antituberculosis Treatment and Its Correlation With Clinical Outcomes in China. *Clinical Infectious*
541 *Diseases* (2020) doi:10.1093/cid/ciaa1002.

542 27. Nimmo, C. *et al.* Dynamics of within-host Mycobacterium tuberculosis diversity and heteroresistance
543 during treatment. *EBioMedicine* **55**, (2020).

544 28. Xu, J. *et al.* Primary Clofazimine and Bedaquiline Resistance among Isolates from Patients with
545 Multidrug-Resistant Tuberculosis. *Antimicrob Agents Chemother* **61**, e00239-17, e00239-17 (2017).

546 29. Nimmo, C. *et al.* Population-level emergence of bedaquiline and clofazimine resistance-associated
547 variants among patients with drug-resistant tuberculosis in southern Africa: a phenotypic and
548 phylogenetic analysis. *The Lancet Microbe* **1**, e165–e174 (2020).

549 30. Beckert, P. *et al.* MDR M. tuberculosis outbreak clone in Eswatini missed by Xpert has elevated
550 bedaquiline resistance dated to the pre-treatment era. *Genome Medicine* **12**, 104 (2020).

551 31. Liu, A. *et al.* Selective Advantage of Resistant Strains at Trace Levels of Antibiotics: a Simple and
552 Ultrasensitive Color Test for Detection of Antibiotics and Genotoxic Agents. *Antimicrobial Agents and*
553 *Chemotherapy* **55**, 1204–1210 (2011).

554 32. Gullberg, E. *et al.* Selection of Resistant Bacteria at Very Low Antibiotic Concentrations. *PLoS*
555 *Pathogens* **7**, e1002158 (2011).

556 33. Dartois, V. The path of anti-tuberculosis drugs: from blood to lesions to mycobacterial cells. *Nat Rev*
557 *Micro* **12**, 159–167 (2014).

558 34. Song, H., Sandie, R., Wang, Y., Andrade-Navarro, M. A. & Niederweis, M. Identification of outer
559 membrane proteins of Mycobacterium tuberculosis. *Tuberculosis* **88**, 526–544 (2008).

560 35. Gu, S. *et al.* Comprehensive Proteomic Profiling of the Membrane Constituents of a Mycobacterium
561 tuberculosis Strain*S. *Molecular & Cellular Proteomics* **2**, 1284–1296 (2003).

562 36. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads.
563 *Microbial Genomics,* **3**, e000131 (2017).

564 37. Veziris, N. *et al.* Rapid emergence of Mycobacterium tuberculosis bedaquiline resistance: lessons to
565 avoid repeating past errors. *Eur Respir J* **49**, 1601719 (2017).

566 38. Battaglia, S. *et al. Characterization of genomic variants associated with resistance to bedaquiline and*
567 *delamanid in naïve Mycobacterium tuberculosis clinical strains*.
568 http://biorxiv.org/lookup/doi/10.1101/2020.05.27.120451 (2020) doi:10.1101/2020.05.27.120451.

569 39. Zimenkov, D. V. *et al.* Examination of bedaquiline- and linezolid-resistant Mycobacterium tuberculosis
570 isolates from the Moscow region. *Journal of Antimicrobial Chemotherapy* **72**, 1901–1906 (2017).

571 40. Ghodousi, A. *et al.* Acquisition of Cross-Resistance to Bedaquiline and Clofazimine following
572 Treatment for Tuberculosis in Pakistan. *Antimicrob Agents Chemother* **63**, e00915-19,
573 /aac/63/9/AAC.00915-19.atom (2019).

574 41. Nimmo, C. *et al.* Bedaquiline resistance in drug-resistant tuberculosis HIV co-infected patients.
575 *European Respiratory Journal* (2020) doi:10.1183/13993003.02383-2019.

576 42. Consortium, T. Cr. & Fowler, P. W. Epidemiological cutoff values for a 96-well broth microdilution
577 plate for high-throughput research antibiotic susceptibility testing of M. tuberculosis. *medRxiv*
578 2021.02.24.21252386 (2021) doi:10.1101/2021.02.24.21252386.

579 43. Hong, M., Fuangthong, M., Helmann, J. D. & Brennan, R. G. Structure of an OhrR-ohrA operator
580 complex reveals the DNA binding mechanism of the MarR family. *Molecular Cell* (2005)
581 doi:10.1016/j.molcel.2005.09.013.

582 44. Kumarevel, T. The MarR Family of Transcriptional Regulators - A Structural Perspective. *Antibiotic*
583 *Resistant Bacteria - A Continuous Challenge in the New Millennium* (2012) doi:10.5772/28565.

584 45. Kadura, S. *et al.* Systematic review of mutations associated with resistance to the new and repurposed
585 Mycobacterium tuberculosis drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid.
586 *Journal of Antimicrobial Chemotherapy* **75**, 2031–2043 (2020).

587 46. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC*
588 *Genomics* **16**, S1 (2015).

589 47. Makhado, N. A. *et al.* Outbreak of multidrug-resistant tuberculosis in South Africa undetected by
590 WHO-endorsed commercial tests: an observational study. *The Lancet Infectious Diseases* **18**, 1350–
591 1359 (2018).

592 48. Gröschel, M. I. *et al.* Pathogen-based precision medicine for drug-resistant tuberculosis. *PLoS Pathog*
593 **14**, e1007297 (2018).

594 49. Merker, M. *et al.* Phylogenetically informative mutations in genes implicated in antibiotic resistance
595 in Mycobacterium tuberculosis complex. *Genome Med* **12**, 27 (2020).

596 50. Kumar, S. & Jena, L. Understanding Rifampicin Resistance in Tuberculosis through a Computational
597 Approach. *Genomics Inform* **12**, 276–282 (2014).

598 51. Zhang, Q. *et al.* Uncovering the Resistance Mechanism of Mycobacterium tuberculosis to Rifampicin
599 Due to RNA Polymerase H451D/Y/R Mutations From Computational Perspective. *Front. Chem.* **7**,
600 (2019).

601 52. Marney, M. W., Metzger, R. P., Hecht, D. & Valafar, F. Modeling the structural origins of drug
602 resistance to isoniazid via key mutations in Mycobacterium tuberculosis catalase-peroxidase, KatG.
603 *Tuberculosis* **108**, 155–162 (2018).

604 53. Mehmood, A. *et al.* Structural Dynamics Behind Clinical Mutants of PncA-Asp12Ala, Pro54Leu, and
605 His57Pro of Mycobacterium tuberculosis Associated With Pyrazinamide Resistance. *Front. Bioeng.*
606 *Biotechnol.* **7**, (2019).

607 54. Carter, J. J. *et al.* Prediction of pyrazinamide resistance in Mycobacterium tuberculosis using
608 structure-based machine learning approaches. *bioRxiv* 518142 (2019) doi:10.1101/518142.

609 55. Tunstall, T. *et al.* Combining structure and genomics to understand antimicrobial resistance.
610 *Computational and Structural Biotechnology Journal* **18**, 3377–3394 (2020).

611 56. Karmakar, M. *et al.* Empirical ways to identify novel Bedaquiline resistance mutations in AtpE. *PLOS*
612 *ONE* **14**, e0217169 (2019).

613 57. Feuerriegel, S. *et al.* Rapid genomic first- and second-line drug resistance prediction from clinical
614 Mycobacterium tuberculosis specimens using Deeplex®-MycTB. *European Respiratory Journal* (2020)
615 doi:10.1183/13993003.01796-2020.

616 58. Jouet, A. *et al.* Deep amplicon sequencing for culture-free prediction of susceptibility or resistance to
617 13 anti-tuberculous drugs. *European Respiratory Journal* (2020) doi:10.1183/13993003.02338-2020.

618 59. Cabibbe, A. M. *et al.* Application of Targeted Next-Generation Sequencing Assay on a Portable
619 Sequencing Platform for Culture-Free Detection of Drug-Resistant Tuberculosis from Clinical Samples.
620 *Journal of Clinical Microbiology* **58**, (2020).

621 60. Dartois, V. The path of anti-tuberculosis drugs: From blood to lesions to mycobacterial cells. *Nature*
622 *Reviews Microbiology* (2014) doi:10.1038/nrmicro3200.

623 61. Prideaux, B. *et al.* The association between sterilizing activity and drug distribution into tuberculosis
624 lesions. *Nature Medicine* (2015) doi:10.1038/nm.3937.

625 62. Dheda, K. *et al.* Drug-penetration gradients associated with acquired drug resistance in patients with
626 tuberculosis. *American Journal of Respiratory and Critical Care Medicine* (2018)
627 doi:10.1164/rccm.201711-2333OC.

628 63. Strydom, N. *et al.* Tuberculosis drugs' distribution and emergence of resistance in patient's lung
629 lesions: A mechanistic model and tool for regimen and dose optimization. *PLoS Medicine* (2019)
630 doi:10.1371/journal.pmed.1002773.

631 64. Sarathy, J. P. *et al.* Prediction of Drug Penetration in Tuberculosis Lesions. *ACS Infect Dis* **2**, 552–563
632 (2016).

633 65. Müller, B., Borrell, S., Rose, G. & Gagneux, S. The heterogeneous evolution of multidrug-resistant
634 Mycobacterium tuberculosis. *Trends in Genetics* **29**, 160–169 (2013).

635 66. van Soolingen, D., Hermans, P. W., de Haas, P. E., Soll, D. R. & van Embden, J. D. Occurrence and
636 stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an
637 insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis.
638 *Journal of Clinical Microbiology* **29**, 2578–2586 (1991).

639 67. Kohl, T. A. *et al.* MTBseq: a comprehensive pipeline for whole genome sequence analysis of
640 Mycobacterium tuberculosis complex isolates. *PeerJ* **6**, e5895 (2018).

641 68. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

642 69. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants
643 using colored de Bruijn graphs. *Nature Genetics* **44**, 226–232 (2012).

644 70. Rancoita, P. M. V. *et al.* Validating a 14-Drug Microtiter Plate Containing Bedaquiline and Delamanid
645 for Large-Scale Research Susceptibility Testing of Mycobacterium tuberculosis. *Antimicrob Agents*
646 *Chemother* **62**, e00344-18, /aac/62/9/e00344-18.atom (2018).

647 71. Fowler, P. W. *et al.* Automated detection of bacterial growth on 96-well plates for high-throughput
648 drug susceptibility testing of mycobacterium tuberculosis. *Microbiology (United Kingdom)* (2018)
649 doi:10.1099/mic.0.000733.

650 72. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis.
651 *Journal of Computational Chemistry* **25**, 1605–1612 (2004).

652 73. Pires, D. E. V., Blundell, T. L. & Ascher, D. B. MCSM-lig: Quantifying the effects of mutations on protein-
653 small molecule affinity in genetic disease and emergence of drug resistance. *Scientific Reports* **6**, 1–8
654 (2016).

655 74. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. MCSM: Predicting the effects of mutations in proteins
656 using graph-based signatures. *Bioinformatics* **30**, 335–342 (2014).

657 75. Decherchi, S., Bottegoni, G., Spitaleri, A., Rocchia, W. & Cavalli, A. BiKi Life Sciences: A New Suite for
658 Molecular Dynamics and Related Methods in Drug Discovery. *J. Chem. Inf. Model.* **58**, 219–224 (2018).