

1 **Insertion and deletion evolution reflects antibiotics selection pressure in a**  
2 ***Mycobacterium tuberculosis* outbreak**

3

4 Maxime Godfroid<sup>1,\*</sup>, Tal Dagan<sup>1</sup>, Matthias Merker<sup>2,3</sup>, Thomas A. Kohl<sup>2,3</sup>, Roland Diel<sup>4,5</sup>,  
5 Florian P. Maurer<sup>6,7</sup>, Stefan Niemann<sup>2,3</sup>, Anne Kupczok<sup>1,\*</sup>

6 <sup>1</sup> Institute of General Microbiology, Kiel University, Kiel, Germany

7 <sup>2</sup> Molecular and Experimental Mycobacteriology, Research Center Borstel, Borstel,  
8 Germany.

9 <sup>3</sup> German Center for Infection Research, Partner Site Hamburg-Lübeck-Borstel-Riems,  
10 Borstel, Germany

11 <sup>4</sup> Institute for Epidemiology, University Medical Hospital Schleswig-Holstein, Kiel, Germany

12 <sup>5</sup> Lungenclinic Grosshansdorf, Airway Research Center North (ARCN), Member of the  
13 German Center for Lung Research (DZL), Großhansdorf, Germany

14 <sup>6</sup> National and WHO Supranational Reference Center for Mycobacteria, Research Center  
15 Borstel, Borstel, Germany

16 <sup>7</sup> Institute of Medical Microbiology, Virology and Hygiene, University Medical Center  
17 Hamburg-Eppendorf, Hamburg, Germany

18

19 \* corresponding author

20 E-mail: [mgodfroid@ifam.uni-kiel.de](mailto:mgodfroid@ifam.uni-kiel.de), [akupczok@ifam.uni-kiel.de](mailto:akupczok@ifam.uni-kiel.de)

## 21 **Abstract**

22 In genome evolution, genetic variants are the source of diversity, which natural selection acts upon.  
23 Treatment of human tuberculosis (TB) induces a strong selection pressure for the emergence of  
24 antibiotic resistance in the infecting *Mycobacterium tuberculosis* (MTB) strains. MTB evolution in  
25 response to treatment has been intensively studied and mainly attributed to point substitutions.  
26 However, the contribution of insertions and deletions (indels) to MTB genome evolution remains  
27 poorly understood. Here, we analyzed a multi-drug resistant MTB outbreak for the presence of high-  
28 quality indels and substitutions. We find that indels are significantly enriched in genes conferring  
29 antibiotic resistance. Furthermore, we show that indels are inherited during the outbreak and follow  
30 a molecular clock with an evolutionary rate of 5.37e-9 indels/site/year, which is 23x lower compared  
31 to the substitution rate. Inherited indels may co-occur with substitutions in genes along related  
32 biological pathways; examples are iron storage and resistance to second-line antibiotics. This suggests  
33 that epistatic interactions between indels and substitutions affect antibiotic resistance and  
34 compensatory evolution in MTB.

## 35 **Author summary**

36 *Mycobacterium tuberculosis* (MTB) is a human pathogen causing millions of deaths every year. Its  
37 genome evolution has been intensively characterized through point substitutions, i.e., nucleotide  
38 exchanges that are inherited. Additional mutations are short or long insertions and deletions of  
39 nucleotides, termed indels. Short indels in genes might change the reading frame and disrupt the gene  
40 product. Here we show that antibiotic treatment has a strong impact on indel evolution in an MTB  
41 outbreak. Namely, indels occur frequently in genes causing antibiotic resistance upon disruption.  
42 Furthermore, we show that the molecular clock, i.e., the temporal emergence of variants over time,  
43 holds for short indels in MTB genomes. Finally, we observe that indels may co-occur with substitutions  
44 in genes along related biological pathways. These results support the notion that indels are important  
45 contributors to MTB evolution. We anticipate that including indels in the analyses of MTB outbreaks  
46 will improve our understanding of antibiotic resistance evolution.

## 47 **Introduction**

48 *Mycobacterium tuberculosis* complex (MTBC) strains, the causative agents of tuberculosis (TB), are  
49 strict host-associated pathogens (1). With estimated numbers of ten million new infections and 1.2  
50 million deaths in 2018 (2), TB is a major cause of human disease and mortality. In addition,  
51 *Mycobacterium tuberculosis sensu stricto* (MTB), the human-adapted member of the MTBC, has a high  
52 level of intrinsic and evolved antibiotic resistance (ABR), including multi-drug resistance (3). MTB

53 genomes have a low genetic diversity and furthermore, comparative genomics of MTB genomes  
54 showed that genetic variation is only vertically inherited, likely due to the absence of horizontal  
55 transfer mechanisms in MTB (4,5). Consequently, MTB antibiotic resistance is considered to evolve *de*  
56 *novo* via point and segmental mutations and not by horizontal transfer of genetic material (6).  
57 Antibiotic resistance may induce high fitness costs that are frequently ameliorated by compensatory  
58 mutations (7). For example in MTB, mutations in *rpoB*, encoding the beta-subunit of the RNA  
59 polymerase, can lead to rifampicin resistance (8) and mutations in *rpoC* often compensate ABR-  
60 conferring mutations in *rpoB* (9,10). Notably, in asexual organisms, beneficial alleles are linked to the  
61 genetic background where they appeared. This results in competition between beneficial alleles (also  
62 known as clonal interference) and the hitchhiking of neutral or slightly deleterious alleles with  
63 beneficial ones. Indeed, time series patient sampling revealed that clonal interference and hitchhiking  
64 contribute to antibiotic resistance evolution in MTB (11,12).

65 Genetic variation in MTB strains is generally characterized by the emergence of substitutions  
66 that are observed as single-nucleotide polymorphisms (SNPs). Substitutions are the major source of  
67 variation in MTB genomes followed by insertions and deletions (indels). Short indels (up to 50 bp)  
68 were found to occur primarily in non-coding regions, in the repeat-containing PE-PPE genes and in  
69 ABR-conferring genes (13). Additionally, long insertions in MTB are mainly due to integration of the  
70 mobile element IS6110, a transposase-mediated insertion sequence (14). Importantly, previous  
71 studies analyzing MTB strain genome evolution provided evidence for the role of indels in ABR  
72 evolution (15–17).

73 Similarly to resistance determination, transmission dynamics within MTB outbreaks is  
74 generally inferred by SNP-based phylogeny reconstruction, after detecting SNPs from short-read  
75 sequencing data aligned to the complete and well characterized reference genome H37Rv (18,19).  
76 Outbreak reconstructions have furthermore been used to identify signals of positive selection in MTB  
77 strain evolution, for example, by identifying convergent evolution, i.e., variants that evolved  
78 independently multiple times. Convergent evolution in MTB has been observed in ABR-conferring  
79 genes (20) or in virulence factors (21). Furthermore, time-series sampling of MTB strains showed that  
80 substitutions in MTB genomes evolve at an approximately constant pace, i.e., substitutions follow a  
81 molecular clock (22). Notably, the substitution rate of MTB is on the lower end spectrum of prokaryotic  
82 substitution rates (6). Despite the low evolutionary rate observed for MTB strains, molecular dating  
83 can be used to infer the time of emergence of ABR-conferring substitutions (23) or the introduction  
84 time of strains into specific parts of the world (24).

85 Although previous studies extensively investigated the rate and impact of substitutions in  
86 MTB strain evolution, the contribution of indels has been sparsely analyzed. To address this question,

87 we estimated the evolutionary rate and the phenotypic impact of insertions and deletions in MTB  
88 outbreak strains. In asexual organisms, genetic linkage is strong and might lead to epistatic  
89 interactions between variants. Hence, we investigated phylogenetically co-occurring indels and  
90 substitutions to describe their putative combined effects on MTB phenotype. As a paramount example  
91 for drug resistance evolution, we analyze a previously described multi-drug resistant clade of MTB  
92 lineage 2 (Beijing) strains, i.e., the Central Asian outbreak (CAO) (25). We further compare some  
93 aspects of indel evolution to the drug-susceptible lineage 4 (Euro-American) 'Hamburg outbreak' (26).

## 94 **Results**

95 To study the evolution of point and segmental mutations in *M. tuberculosis*, we analyzed 353 MTB  
96 strains of the lineage 2 CAO that was detected previously to be involved in transmission of multi-drug  
97 resistant TB mainly in central Asia (Table S1a). Genetic variants were inferred by comparing the sample  
98 genomes to a closely related reference genome (strain *M. tuberculosis* 49-02 (25)). To assess the  
99 robustness of the genetic variation inference, we developed a back-genotyping approach, in which  
100 the inference procedure is performed against a simulated reference genome that includes the  
101 detected variants (Fig S1). Variants that are not supported by back-genotyping were considered as  
102 uncertain in this sample and variants that were inferred to be uncertain in many samples are  
103 unreliable and discarded from the analysis (Fig S2). Using our approach, we inferred in total 1806 high-  
104 quality variants in the CAO strains. These variants comprise a majority of SNPs (1598, 88.5%) and 208  
105 insertions and deletions, where the majority of inferred indels are short ( $\leq 50$ bp; Table 1, Fig S3). We  
106 noted a peak in the distribution of insertion length around 1360bp that corresponds to 38 different  
107 insertions of the mobile element IS6110 (Fig S3). IS6110 insertions are known to be found  
108 preferentially in some genomic regions, i.e., insertional hotspots, where IS6110 insertions confer a  
109 growth advantage (14). The distribution of distances between IS6110 insertions in the CAO revealed  
110 seven insertional hotspots in the MTB genome, of which two hotspots have been previously described  
111 (Fig S4) (14).

112 **Table 1. Summary and genomic localization of detected variants.**

	SNPs	Insertions		Deletions		Total
		Short	Long	Short	Long	
<b>In gene</b>	1362 (85.23%)	48 (73.85%)	28 (63.64%)	65 (73.86%)	11 (100%)	<b>1514</b>
<b>Intergenic</b>	202 (12.64%)	13 (20%)	13 (29.54%)	19 (21.59%)	-	<b>247</b>
<b>In pseudogene</b>	34 (2.13%)	4 (6.15%)	3 (6.82%)	4 (4.55%)	-	<b>45</b>
<b>Total</b>	<b>1598</b>	<b>65</b>	<b>44</b>	<b>88</b>	<b>11</b>	<b>1806</b>
<b>Parsimony informative</b>	434 (27.2%)	18 (27.7%)	14 (31.8%)	16 (18.2%)	4 (36.4%)	<b>486</b>
<b>Compatible</b>	394 (90.8%)	11 (61.1%)	10 (71.4%)	13 (81.2%)	2 (50%)	<b>428</b>

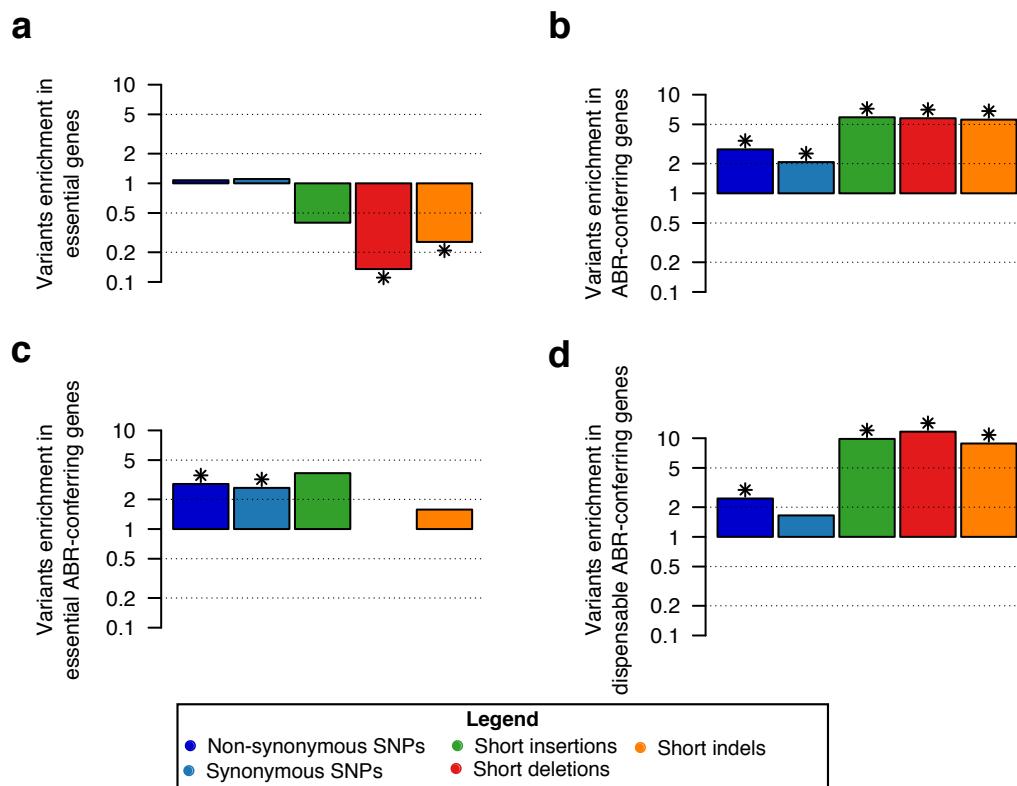
113 Percentages are calculated based on the total number of variants in each variant class, except for the  
114 compatible variants, where the percentage is calculated based on the number of parsimony  
115 informative variants in each variant class. The majority of SNPs and indels are found in coding regions.  
116 The distribution of variants in protein-coding and intergenic regions differs significantly from the  
117 expectation by chance (Fisher's exact test,  $p < 0.01$ ), where 247 (13.7%) variants are inferred in  
118 intergenic regions that span 7.8% of the genome.

119

120 **Short indels contribute significantly to antibiotic resistance evolution.**

121 To infer the putative phenotypic impact of the inferred variants in coding regions we examined their  
122 localization in genes of known function. For this purpose, we retrieved a list of ABR-conferring genes  
123 (Table S2), i.e., genes where mutations were found to confer antibiotic resistance. Additionally, we  
124 classified the MTB genes in two categories of essentiality, according to their requirement for growth  
125 *in vitro* (i.e., essential) or not (i.e., dispensable) (27). In particular, we investigated the distribution of  
126 genetic variants in essential and ABR-conferring genes. Depletion of variants in specific gene  
127 categories indicates purifying selection acting on that category, whereas enrichment serves as an  
128 indication for positive selection.

129 First, we observed a fourfold depletion of short indel frequency in essential genes; the  
130 distribution of SNPs, however, is not significantly different between essential and dispensable genes  
131 (Fig 1a). Furthermore, SNPs and short indels are enriched in ABR-conferring genes compared to the  
132 remaining genes (Fig 1b). When we classify the ABR-conferring genes into essential (27, 29.4% of ABR-  
133 conferring genes) and dispensable (65, 70.6% of ABR-conferring genes), we observed that SNPs are  
134 enriched both in ABR-conferring genes that are essential and in ABR-conferring genes that are  
135 dispensable. In contrast, short indels are significantly enriched in the ABR-conferring genes that are  
136 dispensable but not in the essential ABR-conferring genes (Fig 1c,d). In comparison, the drug-  
137 susceptible Hamburg outbreak did not show variants in ABR-conferring genes (Fig S5).



138

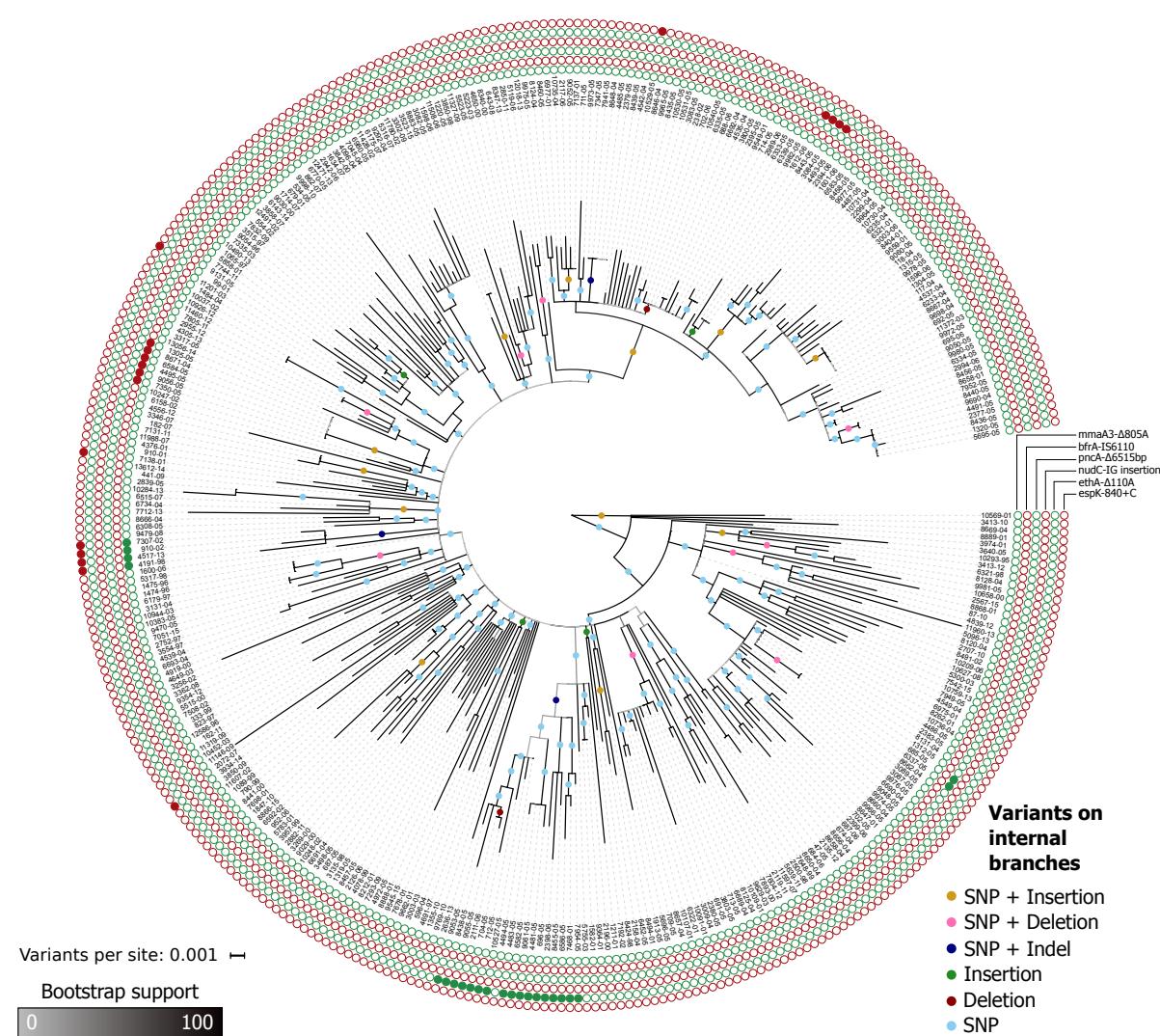
139 **Fig 1. Enrichment analyses of variants in gene categories.** The variants enrichment is calculated as  
140 the ratio of the proportion of genes with variants in a gene category and the proportion of genes with  
141 variants outside the gene category. For example, three essential genes (0.6% of all essential genes)  
142 have short indels and the remaining short indels occur in 94 dispensable genes (3.3% of all dispensable  
143 genes), which results in a variants ratio of short indels in essential genes of 0.25, i.e., a fourfold  
144 depletion. We show the ratio for the gene categories (a) Essential, (b) ABR-conferring, (c) Essential  
145 and ABR-conferring, and (d) Dispensable and ABR-conferring. Significant enrichment or depletion,  
146 marked by a star, is estimated using Fisher's exact test (p-value < 0.05, corrected for FDR, Table S3).

147

148 The enrichment analyses highlight the selection pressures on SNPs and indels in the CAO. The  
149 depletion of short indels in essential genes provides evidence for the presence of strong purifying  
150 selection against indels in essential genes. In addition, the observed enrichment in ABR-conferring  
151 genes likely stems from the strong selection pressure on antibiotic resistance in the multi-drug  
152 resistant CAO. The significant enrichment of short indels in ABR-conferring genes that are dispensable  
153 shows that indels contribute to the evolution of antibiotic resistance in a highly resistant outbreak,  
154 potentially by frameshifts that disrupt the protein sequences.

155 **Insertions and deletions contribute phylogenetic signal in an MTB outbreak.**

156 To study the transmission of indels in an outbreak, we next describe how the genetic variants are  
157 inherited in the CAO. To this end, we reconstructed the outbreak phylogeny from the presence-  
158 absence pattern of the variants in the strain genomes, where uncertain variants in a sample  
159 correspond to gapped positions. This analysis revealed that SNPs are the main contributors to the  
160 phylogenetic signal, where most of parsimony informative SNPs are compatible with the phylogeny  
161 (Table 1). The inclusion of indels in the phylogenetic reconstruction increases the resolution of the  
162 tree topology at multiple places, where six internal branches are supported by a single short indel only  
163 (Fig 2). In comparison, the analysis of 64 samples of the 'Hamburg outbreak' resulted in 112 variants,  
164 of which the majority are SNPs as well (Fig S5a). In the Hamburg outbreak phylogeny, two internal  
165 branches are supported by indels only (Fig S5b).



166  
167 **Fig 2. Phylogenetic tree of the CAO.** 486 (26.9%) of the variants are parsimony informative, and 56  
168 variants (3.1%) are incompatible with the tree topology (Table 1). The root position is the temporal  
169 root estimated by dating the phylogeny with LSD. Circles on branches represent variants that are

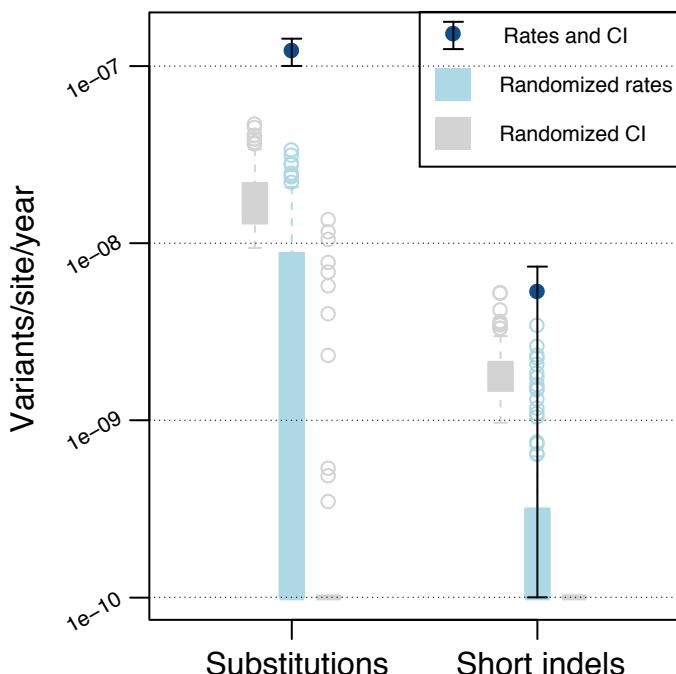
170 compatible with the branch, i.e., they likely have emerged on that branch. We found that six branches  
171 in the tree have only short indels (four branches with insertions, two branches with deletions). The  
172 outer circles show example variants that are highlighted in the text.

173

174 We also compared our approach to the standard SNP-based phylogenetic approach for the  
175 CAO data. The complete tree contains 157 internal branches of length larger than zero, whereas the  
176 SNP phylogeny contains 149 internal branches of length larger than zero. Notably, 138 branches  
177 (92.6% of the branches in the SNP phylogeny) are found in both trees. Thus, our phylogenetic  
178 inference is consistent with the standard SNP-based phylogeny, where the resolution of additional  
179 branches indicates that indels provide additional insights into possible transmission events.

180 **Substitutions and short indels in the CAO follow a molecular clock.**

181 To compare the pace of substitutions and indels in the outbreaks, we examined their evolutionary  
182 rates on the phylogeny. We found that substitutions passed the stringent test for temporal signal,  
183 with an estimated rate of 1.23e-7 substitutions/site/year (Fig 3). For comparison, substitutions in the  
184 Hamburg outbreak passed the intermediate test for temporal signal, with an estimated rate of 7.51e-  
185 8 substitutions/site/year (Fig S5c). This rate is lower compared to the CAO rate; however, the  
186 confidence intervals of the estimates overlap (Fig 3, S5c). The substitution rate estimated here for the  
187 CAO is within previously estimated rates for lineage 2 (22). Furthermore, there is a known difference  
188 in the rate of evolution between MTB lineage 2 and lineage 4 (28), which is consistent with our  
189 estimates for the lineage 2 CAO and the lineage 4 Hamburg outbreak.



190

191 **Fig 3. Evolutionary rates for substitutions and short indels and their associated 95% confidence**  
192 **intervals (CI) estimated with LSD.** Substitution rate is estimated at 1.23e-7 [1.00e-7–1.43e-7]  
193 substitutions/site/year and the short indel rate is estimated at 5.37e-9 [1e-10–7.37e-9]  
194 indels/site/year. The randomized rates are estimated with the date-randomization test (29). For  
195 substitutions, there is no overlap in rate and confidence interval between real data and randomized  
196 data, indicating temporal signal accorded by the stringent test. For short indels, the rate of the real  
197 data does not overlap with the rates of the randomized data; however, the confidence interval of the  
198 real data overlaps with the confidence intervals of the randomized data, which indicates temporal  
199 signal of short indels accorded by the simple test, is weaker than that of substitutions. Due to the  
200 limited number of events, there is not sufficient temporal signal to estimate evolutionary rates for  
201 short insertions and deletions separately or for long indels.

202

203 We then estimated the evolutionary rate of indels in the CAO. Short insertions and deletions  
204 are assumed to emerge by similar point mutation processes, in contrast to long indels that are due to  
205 segmental mutations (e.g., (30)). We thus considered short indels and long indels separately for the  
206 rate estimation, where only short indels have temporal signal (according to the simple test), with a  
207 rate of 5.37e-9 short indels/site/year (Fig 3). Hence, our analysis revealed that substitutions and short  
208 indels follow a molecular clock and that short indels evolve 23 times slower than substitutions in MTB  
209 lineage 2 strains. The difference between substitution and indel rates might be explained by the  
210 extremely efficient and redundant MTB repair mechanisms. Error-prone repair mechanisms, such as

211 the DnaE2 pathway that is involved in trans-lesion synthesis, are known to introduce substitutions  
212 (31). Thus, MTB repair mechanisms might generate a mutational bias towards substitutions, leading  
213 to a stable genome with few structural variations over evolutionary time.

214 **Indels are subject to vertical inheritance and convergent evolution in the CAO.**

215 To describe the function of variants that are transmitted in the outbreak or that evolved multiple times  
216 independently, we explored the phylogenetic distribution and congruence of indel variants in the  
217 reconstructed CAO tree. Parsimony informative variants that are compatible with the phylogeny  
218 (termed compatible variants hereafter) are inferred as vertically inherited and transmitted to multiple  
219 hosts; thus, their effect on MTB fitness is likely not deleterious or even advantageous. In addition,  
220 incompatible variants are convergent events, where the same variant occurs independently in two (or  
221 more) disparate branches of the tree, indicating convergent evolution. Convergent variants might  
222 serve as evidence for positive selection if they emerge in similar genetic backgrounds and have  
223 identical phenotypic impact (32).

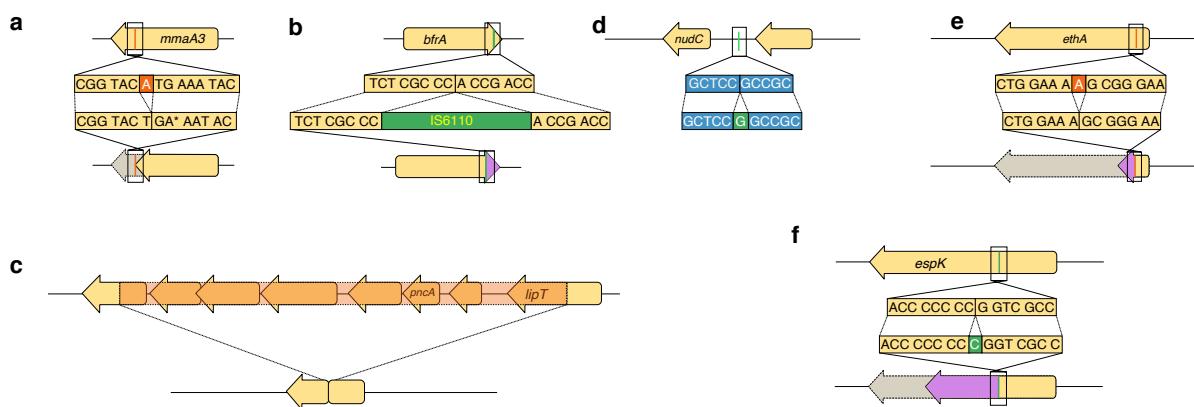
224 We found 36 compatible and 16 incompatible indels (31.8%), where incompatible indels are  
225 enriched among the parsimony informative indels in comparison to incompatible SNPs (40, 9.22% of  
226 all parsimony informative SNPs,  $p\text{-value} < 0.01$ , Fisher's exact test). This difference can be traced back  
227 to convergent short indels in homopolymer regions and also inference bias (Table S4). We found that  
228 30 of the 36 compatible indels are located in coding regions (83.3%; Table S4), out of which 29 are in  
229 dispensable genes and two are in ABR-conferring genes (*mmaA3* and *ethA*; Fig 2). In contrast, nine of  
230 the 16 incompatible indels are found in coding regions (56.2%; Table S4). Thereof, eight indels are  
231 found in dispensable genes and two in ABR-conferring genes, where the only incompatible indels in  
232 ABR-conferring genes are long deletions completely or partially deleting *pncA*. In contrast,  
233 incompatible SNPs are mainly found in genes conferring antibiotic resistance (21 of 29 incompatible  
234 SNPs in coding regions, 72.4%, Table S5). This is in agreement with previous results where convergent  
235 SNPs have been observed in MTB genes that confer antibiotic resistance and compensatory  
236 mechanisms (17,20).

237 We next considered genes in which multiple indels were inferred, i.e., genes affected by  
238 convergent evolution due to indels. We found 15 indels affecting four ABR-conferring genes; in three  
239 of these genes (*rpoB*, *tlyA*, *ethA*) additional SNPs were inferred in different samples, whereas no SNPs  
240 were inferred in *pncA* (Table S6). In addition, twelve genes where multiple indels have been found do  
241 not confer antibiotic resistance (e.g., *espB* and four PE genes; Table S6); these genes do not contain  
242 SNPs in any sample.

243 PE and PPE are repeat-containing genes that are secreted and they are hypothesized to be  
244 important for MTB interaction with the host immune system (33). An examination of all variants  
245 inferred in PE and PPE genes in the CAO strains revealed 17 short indels. Of these, only three (17.6%)  
246 cause a frameshift, which is much lower than the proportion of frameshift-causing indels in all coding  
247 sequences (77.8%). Furthermore, we found seven parsimony informative short indels, out of which  
248 six are compatible (five in-frame) and one is an incompatible in-frame deletion. The vertical  
249 inheritance and the enrichment of in-frame indels in PE and PPE genes indicate that these proteins  
250 are fast evolving, further supporting the hypothesis that they are involved in host recognition (33).

251 Taken together, we found 52 parsimony informative indels in 34 different genes. It is  
252 remarkable that only two of these genes (*ethA* and a nitronate monooxygenase) likely evolved under  
253 positive selection as inferred by the ratio of nonsynonymous to synonymous substitutions (dN/dS>1,  
254 Table S7). We note that the inference of positive selection can only be performed for three of the 34  
255 genes with parsimony informative indels due to the lack of SNPs in the remaining genes. We further  
256 discovered that four genes with parsimony informative indels (11.8%) are included in a set of 116 (3%)  
257 genes that were found to be under positive selection in a recent survey of dN/dS in MTB (34). Thus,  
258 while indels can be found in genes that are under positive selection as calculated by the dN/dS ratio,  
259 they might also uncover additional genes involved in adaptation. In the following, we study six  
260 example indels in detail that were selected to highlight convergent evolution, inherited antibiotic  
261 resistance, and putative epistatic interactions.

262 **A short deletion shortens the ABR-conferring gene *mmaA3*.** A compatible deletion of one base pair  
263 was observed in the gene *mmaA3* in four related samples (Fig 2). The deletion results in a frameshift  
264 and a premature stop codon yielding a truncated protein sequence (Fig 4a). The protein MmaA3 acts  
265 along the synthesis pathway of mycolic acids, which are essential components of the bacterial  
266 membrane (35). The gene is classified as ABR-conferring, yet it is classified as dispensable *in vitro*. In  
267 addition, we observed five SNPs and one IS6110 insertion that co-occur with the 1bp deletion in the  
268 same four samples (Table 2). Three of the five SNPs are non-synonymous substitutions in genes that  
269 encode proteins involved in membrane biogenesis (Table 2). Our results thus revealed several  
270 substitutions and indels, which emerged and were vertically inherited together, and which likely have  
271 an effect on the function of membrane biosynthesis genes.



273 **Fig 4. Examples of indels subject to vertical inheritance and convergent evolution.** The orientation  
274 of the gene is relative to the reference genome 49-02. At the top is the ancestral (reference) sequence  
275 and at the bottom the evolved sequence. See Fig 2 for phylogenetic locations. Gene names are  
276 displayed according to the Mycobrowser annotations (<https://mycobrowser.epfl.ch/>, last accessed 28  
277 November 2019), when available. Following annotations are shown as (Locus tag in 49-02, homolog  
278 in H37Rv). (a) Single base-pair deletion close to the 5' end of *mmaA3* (MT49\_RS03370, Rv0643c). The  
279 deletion removes the residue at position 805 of the coding sequence, resulting in a stop codon where  
280 the mutated protein (269 amino acids) corresponds to 91.5% of the wild-type protein. Six variants co-  
281 occur on the same internal branch (Table 2). (b) IS6110 integration close to the 3' end of the *bfrA* gene  
282 (MT49\_RS09760, Rv1876). The integration occurs at position 471 of the coding sequence (98.1% of  
283 total CDS length), resulting in a protein of 162 amino acids (2 amino acids longer than the wild-type),  
284 with the last two amino acids of the wild type different in the mutated protein. We find eight co-  
285 occurring variants on the same internal branch (Table 2). (c) 6515 base-pair deletion removing the  
286 *pncA* gene (MT49\_RS10715, Rv2043c) and five of its neighboring genes. The left breakpoint is located  
287 at position 529 in the gene *ugpC* (MT49\_RS10690, Rv2038c). The right breakpoint is located at position  
288 453 in a gene encoding for a carboxylesterase/lipase family protein (MT49\_RS10725, Rv2045c). Four  
289 of the deleted neighboring genes encode for ABC transporters (Table S4). Two variants co-occur with  
290 this long deletion (Table 2). (d) Intergenic single base-pair insertion 39bp upstream of the *nudC* gene  
291 (MT49\_RS16870, Rv3199c). We could not detect the promoter sequence in this intergenic region using  
292 BPROM (36). (e) Single base-pair deletion in the beginning of the *ethA* gene (MT49\_RS20315,  
293 Rv3854c). This deletion occurs at position 110 of the coding sequence (7.5% of the CDS length), which  
294 results in a frameshift where the resulting protein is truncated with a length of 62 amino acids long  
295 (12.6% of the wild type length). This deletion co-occurs with two variants (Table 2). (f) Single base-pair  
296 insertion in the *espK* gene (MT49\_RS20440, Rv3879c). This insertion occurs at position 840 of the  
297 coding sequence (37.7%), resulting in a truncated protein of 465 amino acids (62.7% of the wild type  
298 length).

299 **Table 2. Variants that co-occur with the example indels (Fig 2).**

Genes with focus variant	Co-occurring variant	Locus tag	H37Rv homolog	Gene name	Product	Notes
<i>mmaA3</i>	sSNP	MT49_RS03380	Rv0645c	<i>mmaA1</i>	mycolic acid methyltransferase MmaA1	Involved in membrane biogenesis
<i>mmaA3</i>	Intergenic SNP	MT49_RS08155	Rv1535		hypothetical protein	
<i>mmaA3</i>	IS6110	MT49_RS16445	Rv3126c		hypothetical protein	
<i>mmaA3</i>	nsSNP	MT49_RS17845	Rv3383c	<i>idsB</i>	polypropenyl synthetase family protein	Involved in membrane biogenesis
<i>mmaA3</i>	nsSNP	MT49_RS19685	Rv3740c		wax ester/triacylglycerol synthase family O-acyltransferase	Involved in membrane biogenesis
<i>mmaA3</i>	nsSNP	MT49_RS20030	Rv3800c	<i>pks13</i>	acyltransferase domain-containing protein	Involved in membrane biogenesis
<i>bfrA</i>	nsSNP	MT49_RS01400	Rv0265c		ABC transporter substrate-binding protein	Probable periplasmic iron-transport lipoprotein (mycobrowser)
<i>bfrA</i>	nsSNP	MT49_RS03540	Rv0676c	<i>mmpL5</i>	siderophore RND transporter MmpL5	ABR-conferring
<i>bfrA</i>	sSNP	MT49_RS04920	Rv0935	<i>pstC1</i>	phosphate ABC transporter permease	
<i>bfrA</i>	nsSNP	MT49_RS08525	Rv1625c	<i>cya</i>	adenylate cyclase	
<i>bfrA</i>	nsSNP	MT49_RS09530	Rv1830		MerR family transcriptional regulator	
<i>bfrA</i>	sSNP	MT49_RS09595	Rv1843c	<i>guaB1</i>	GuaB1 family IMP dehydrogenase-related protein	
<i>bfrA</i>	nsSNP	MT49_RS12315	Rv2337c		hypothetical protein	
<i>bfrA</i>	nsSNP	MT49_RS20315	Rv3854c	<i>ethA</i>	FAD-containing monooxygenase EthA	ABR-conferring
<i>pncA</i>	Intergenic SNP	MT49_RS03955	Rv0755c	<i>PPE12</i>	PPE family protein PPE12	
<i>pncA</i>	sSNP	MT49_RS13630	Rv2582	<i>ppiB</i>	peptidyl-prolyl cis-trans isomerase	
<i>ethA</i>	nsSNP	MT49_RS03490	Rv0667	<i>rpoB</i>	DNA-directed RNA polymerase subunit beta	Essential ABR-conferring gene
<i>ethA</i>	Intergenic IS6110	MT49_RS09400	Rv1804c		hypothetical protein	predicted secreted protein

300 Gene names are displayed according to the Mycobrowser annotations (<https://mycobrowser.epfl.ch/>,  
 301 last accessed 28 November 2019), when available. Locus tags and products are taken from the  
 302 annotation of the 49-02 reference genome. Notes show additional information related to protein  
 303 functions.

304

305 **An IS6110 insertion elongates a bacterioferritin gene.** A compatible insertion of an IS6110 element  
 306 was identified at the 3' end of *bfrA* in six related samples (Fig 2; Fig 4b). The deleterious effect of indels  
 307 at the 3' end of genes is considered minimal due to the indel location at the end of the open reading  
 308 frame, thus maintaining the majority of the coding sequence (37). Indeed, the IS element insertion  
 309 yields a protein sequence that is two amino acids longer than the wildtype, where 158 amino acids  
 310 (98.8%) of the wild type protein are retained (Fig 4b). BfrA is an essential component for iron storage  
 311 and distribution depending on iron availability (38) and it is classified as dispensable *in vitro*. We  
 312 observed eight additional SNPs that co-occur with the IS6110 insertion in the same six samples (Table  
 313 2). Two of the SNPs are non-synonymous substitutions in ABR-conferring genes: *mmpL5* and *ethA*.  
 314 *MmpL5* is annotated as a siderophore transporter and is associated with bedaquiline resistance; *EthA*  
 315 is associated with resistance to ethionamide (3). Of the remaining co-occurring SNPs, one non-  
 316 synonymous substitution is observed in a gene encoding for a protein related to iron export and one  
 317 non-synonymous substitution is found in a MerR transporter family that plays a role in responding to  
 318 environmental stresses, such as oxidative stress, heavy metals or antibiotics (39) (Table 2). Thus, the  
 319 IS1660 insertion in *bfrA* might hitchhike with the co-occurring substitutions in *mmpL5* and *ethA* or it  
 320 might even be a compensatory variant for these substitutions that confer antibiotic resistance and  
 321 may be additionally related to iron transport and storage.

322 **A long deletion completely removes the ABR-conferring *pncA* and neighboring genes.** A compatible  
323 6515bp deletion of *pncA* with five neighboring genes was observed in two related samples (Fig 2). This  
324 deletion furthermore disrupts two additional neighboring genes and results in a chimeric coding  
325 sequence (Fig 4c). PncA activates pyrazinamide, a first-line antibiotic of the category “prodrug”, i.e., a  
326 compound that needs to be activated to exhibit toxicity. The disruption of *pncA* renders pyrazinamide  
327 inactive and thus results in antibiotic resistance; hence, indels in *pncA* have been previously observed  
328 to confer resistance to pyrazinamide (40). In our data, in addition to the multi-gene deletion, we  
329 observed a complete *pncA* deletion and nine disruptive indels across the tree (four short insertions,  
330 one long insertion, and four long deletions). Thus, *pncA* has the highest frequency of convergent indels  
331 in the CAO.

332 **An intergenic short insertion is located upstream of the ABR-conferring *nudC*.** A compatible 1bp  
333 insertion located in an intergenic region 39bp upstream of *nudC* was observed in four related samples  
334 (Fig 2, Fig 4d). NudC is an NAD(+) diphosphatase, where antibiotic resistance to isoniazid and  
335 ethionamide was observed when overexpressing *nudC* (41). Of note, the *nudC* gene in the outbreak  
336 reference genome 49-02 has a 239P->239R polymorphism compared to the reference H37Rv. This  
337 position has been found to disrupt the NudC dimer formation, hence it is expected to affect the NudC  
338 catalytic activity. The observed indel is compatible; hence, we hypothesize that the short insertion is  
339 advantageous to MTB, for example by altering the *nudC* expression level. An increased *nudC*  
340 expression might confer antibiotic resistance to isoniazid and/or ethionamide.

341 **A short deletion disrupts the ABR-conferring *ethA*.** A compatible 1bp deletion was observed in the  
342 *ethA* gene in 17 samples with an additional uncertain sample (Fig 2). This deletion results in a  
343 frameshift leading to a truncated protein (Fig 4e). EthA, an FAD-containing monooxygenase, is  
344 involved in the activation of ethionamide, a second-line antibiotic prodrug. The downregulation of  
345 *ethA* has been demonstrated to generate an ethionamide resistance phenotype (42). We observed  
346 two additional co-occurring variants in the same samples, where one is a non-synonymous  
347 substitution in *rpoB* (Table 2). We hypothesize that the disruption of EthA confers a strong selective  
348 advantage by mediating resistance to ethionamide. RpoB is known to confer resistance to rifampicin  
349 and mutations outside the rifampicin resistance determining region might compensate the cost of the  
350 resistance (10). Here we observed a substitution outside the rifampicin resistance determining region  
351 and thus hypothesize that it might be involved in compensation, resulting in the vertical inheritance  
352 of the variants. Interestingly, we find 20 additional variants in *ethA* throughout the tree, of which 17  
353 are non-synonymous SNPs and three are frameshift single base pair deletions. The presence of SNPs  
354 that are exclusively non-synonymous shows that this gene is under strong positive selection (Table  
355 S7).

356 **An incompatible short insertion disrupts the type VII secretion system gene *espK*.** We observed a  
357 1bp insertion in the *espK* gene in eight samples including four related samples and four unrelated  
358 samples (Fig 2). Hence this insertion likely emerged five times independently, indicating convergent  
359 evolution. The variant emerged in a homopolymer region of seven cytosines (Fig 4f), resulting in a  
360 frameshift and a premature stop codon. The *espK* gene is located in the ESX-1 locus, a type VII  
361 secretion system. The locus additionally comprises PE and PPE genes, encoding for proteins that are  
362 exported or found in the cell membrane (43). EspK is thought to act as a chaperone of the neighboring  
363 *espB* gene (44), and is found dispensable for growth *in vitro*. EspB acts as a repressor of the host  
364 immune response, thereby increasing MTB survival (45). Notably, it was shown that inhibition of *espK*  
365 and *espB* results in reduced virulence in comparison to the wild-type (45). The 1bp insertion likely  
366 renders EspK nonfunctional and hence has a direct effect on EspB function as well. Previous studies  
367 observed convergent substitutions in another type VII secretion system gene (*esxW* in ESX-5) that  
368 increased MTB transmissibility (21). Hence, the 1bp indel in the type VII secretion system might be  
369 related to MTB transmissibility as well.

370 Our examples demonstrate that indels contribute to the evolution and diversification of MTB  
371 CAO strains, by affecting essential metabolic pathways and antibiotic resistance with potential  
372 pathobiological consequences. Furthermore, we found that compatible indels often co-occur with  
373 substitutions that affect related functions or pathways (Table 2). These co-occurrences could be  
374 explained by epistatic interactions in which either indels compensate the effects of substitutions or  
375 *vice versa*. In addition, the significant enrichment of short indels in ABR-conferring genes that are  
376 dispensable shows that indels contribute significantly to ABR evolution in the multi-drug resistant  
377 CAO, likely by frameshifts that disrupt the protein sequences. Our study demonstrates that, even if  
378 rare, including indels in outbreak genome analyses supplies crucial evidence for the profiling of  
379 antibiotic resistant strains and might reveal epistatic interactions.

## 380 **Discussion**

381 The contribution of indels to genome evolution is often understudied, mainly due to difficulties in  
382 reliable indel detection methodology. Our approach allows to infer high-quality indels by estimating  
383 the level of inference uncertainty, which was used to identify unreliable genetic variants. Applying this  
384 approach to MTB sequencing data, we show that indels can be employed to increase the resolution  
385 of MTB strain comparisons in genomic epidemiology approaches, e.g., for outbreak investigations. The  
386 accurate detection of indels in the CAO revealed an indel evolutionary rate that is lower than the  
387 substitution rate. Finally, indels are an important factor in the evolution of antibiotic resistance in  
388 MTB, where compatible and convergent indels represent putative targets for positive selection and  
389 where co-occurring variants highlight epistatic interactions.

390 MTB outbreak reconstructions so far mainly relied on the estimation of phylogenies based on  
391 SNPs. Here we show that six branches of the CAO tree and two branches of the Hamburg outbreak  
392 tree were reconstructed solely by short indels. Furthermore, the comparison of the CAO tree inferred  
393 solely from SNPs and inferred from SNPs and indels revealed similar topologies. Thus, the inclusion of  
394 indels can refine outbreak phylogenies.

395 The MTB mutation rate, that is, the rate at which mutations arise in the genome, is in the  
396 range of bacterial mutation rates (between 1.4e-10 for *Thermus thermophilus* and 4e-9 for *Buchnera*  
397 *aphidocola*; 1.9e-10 mutations/bp/generation for *M. tuberculosis* (46)). However, MTB strain  
398 evolution is characterized by a long generation time (1) and strong purifying selection that eliminates  
399 most genetic variants from the population, with only few mutations being fixed (47). Both of these  
400 processes contribute to a low substitution rate in MTB strains compared to strains of other bacterial  
401 species (6). A previous comparison between evolutionary rates of mutations and indels in multiple  
402 bacterial species showed a 2.8 to 9.7-fold decrease of indel rates compared to mutation rates (48).  
403 Notably, the comparison in the latter study is based on *de novo* rates, i.e., variants that are arise in a  
404 bacterial individual per generation, which aim to include all variants before selection. In contrast,  
405 outbreak analyses include only the variants that are observed after the effect of selection. Hence, the  
406 23-fold decrease of the MTB indel rate compared to the substitution rate can reflect both a lower *de*  
407 *novo* indel rate, as observed for other bacterial species, and a stronger effect of purifying selection on  
408 indels. The latter is expected since indels incur a higher fitness cost than substitutions as they often  
409 disrupt genes and render a truncated gene product (48).

410 Notably, MTB is evolving strictly vertically without the contribution of recombination. The  
411 advantage of sex and recombination is widely discussed (e.g., (49)). On the one hand, recombination  
412 is beneficial by combining advantageous alleles from different genotypes in the population; whereas  
413 in the absence of recombination, advantageous alleles are linked to the genetic background where  
414 they arise (the Hill-Robertson effect). This genetic linkage might result in the fixation of neutral or  
415 slightly deleterious alleles by genetic hitchhiking with advantageous alleles. In addition, clonal  
416 interference between beneficial alleles in different genetic backgrounds slows down adaptation (the  
417 Fisher-Muller effect). On the other hand, in the presence of positive epistasis, i.e., when the double  
418 mutant has a higher fitness than expected from the individual alleles, recombination can lead to a  
419 decrease in fitness by breaking up advantageous allele combinations (resulting in recombination load).  
420 It has been observed that the magnitude of recombination impacts the genetic architecture of a  
421 species, where positive epistasis evolved in a bacteriophage model system under low recombination  
422 but not under high recombination (50). In addition, an artificial gene network model has been used to  
423 demonstrate that positive epistasis evolves in asexual populations whereas negative epistasis can

424 evolve in sexual populations (51). It is thus expected that the asexual lifestyle of MTB results in a  
425 genetic architecture with widespread positive epistasis.

426 Indeed, epistatic interactions between genetic variants are widespread in MTB, where the  
427 emergence of ABR-conferring mutations is often accompanied by compensatory mutations (3). The  
428 fixation of compensatory variants might even be favored over the reversal of antibiotic resistance in  
429 the absence of ongoing antibiotic treatment, because compensatory mutations might appear with a  
430 higher rate compared to the very specific target of a reversal mutation (52). After the compensatory  
431 mutation increased in frequency, a newly appearing reversal mutation will not establish in the  
432 population due to clonal interference. Subsequently, transmission bottlenecks might contribute to the  
433 fixation of the compensatory mutation, after which the reversal mutation has only a low or no  
434 selective advantage precluding its establishment in the population (52). Thus, in the MTB genetic  
435 architecture with widespread positive epistasis, compensatory variants might have a higher likelihood  
436 of being fixed compared to reversal mutations, even when the combination of resistance and  
437 compensatory mutation has a lower fitness than the reversal mutation. This evolved genetic  
438 architecture thus supports the fixation of compensatory mutations instead of reversal mutations.

439 Here we highlight that inherited indels were found to co-occur with substitutions. Although  
440 some of the co-occurring variants might be explained by genetic hitchhiking, the presence of co-  
441 occurring indels and substitutions in related gene functions or pathways supports that these variants  
442 interact epistatically. We thus conclude that MTB evolved a genetic architecture with widespread  
443 positive epistasis, where epistatic interactions between substitutions and indels contribute to the  
444 establishment of indels in the population.

445 Taken together, our results demonstrate the interplay between substitutions and indels in the  
446 evolution of biological functions that are essential for MTB infection and antibiotic resistance. We  
447 identified short and long indels that improve the resolution of outbreak phylogenies and that are  
448 crucial for the prediction of drug resistance in MTB strains. Especially for new hallmark drugs to treat  
449 multi-drug resistant MTB, such as bedaquiline and clofazimine, indels play a major role in collateral  
450 resistance towards both drugs (53,54). Thus, increasing knowledge on interactions between all variant  
451 types is paramount for our understanding of the fundamental evolutionary principles that govern the  
452 spread of antibiotic resistance and the associated compensatory mechanisms in MTB.

## 453 **Methods**

### 454 **Sample collection and variants calling**

455 We analyzed 353 multi-drug resistant MTB strains sampled longitudinally from the Central Asian  
456 Outbreak (MTB lineage 2, first referenced in (25)). Inclusion criteria were the presence of genetic

457 markers defining the CAO clade (25). The strain collection was mainly assembled from a previously  
458 published collection derived from a drug resistance survey in Karakalpakstan, Uzbekistan (10) and  
459 from routine multi-drug resistance TB surveillance data from German patients (Table S1a) with 199  
460 newly generated datasets and covering a sampling time from 1995 to 2015 (Table S1a). The closely  
461 related and fully drug-susceptible strain 49-02 (RefSeq: NZ\_HG813240.1 version 11-MAR-2017) serves  
462 as the reference genome for variant calling. In addition, we performed the analysis for an outbreak of  
463 64 fully drug-susceptible isolates (Table S1b) in the Hamburg region (MTB lineage 4, first referenced  
464 in (26)). Variants were inferred on the complete reference genome 7199-99 (RefSeq: NC\_020089.1  
465 version 19-MAY-2017).

466 We first trimmed the reads using trimmomatic v. 0.36 (55), with parameters  
467 SLIDINGWINDOW:4:15 MINLEN:36 LEADING:3 TRAILING:3. We mapped the trimmed reads to  
468 the outbreak reference genome using BWA MEM v0.7.16 (56), realigned around indels with GATK  
469 v3.8-0-ge9d806836 (57), and marked duplicates with PICARD v2.13.2. The median coverage ranges  
470 from 41 to 255. To detect SNPs and indels, we combined seven variant calling tools: GATK v3.8-0-  
471 ge9d806836 (57), FreeBayes v1.1.0-50-g61527c5 (58), Delly v0.7.7 (59), Pindel v0.2.5b9 (60), SvABA  
472 FH Version 134 (61), Scalpel v0.5.3 (62), and MindTheGap v2.0.2 (63). Tools were run and variants  
473 filtered according to tool-specific quality scores (Table S8) and we retained variants with a frequency  
474 over 75%.

475 Many indels occur in genomic regions of high GC content or that contain tandem repeats (37).  
476 The alignment of reads in these regions is therefore more difficult, and we expect to find higher  
477 uncertainty levels for indels compared to SNPs. Notably, in the case of MTB, this has led to the  
478 systematic exclusion of variants in PE and PPE genes (19). Here we implement two filtering steps, re-  
479 genotyping and back-genotyping, to obtain high-quality genomic variants.

## 480 **Re-genotyping**

481 Since variants can be missed by variant calling, we performed re-genotyping to ascertain presences  
482 and absences of each variant in every sample. Re-genotyping determines whether the read alignment  
483 contains sufficient signal to support the variant. Additionally, we performed re-genotyping with a  
484 single tool per variant category. Therefore, the re-genotyped variant support values (i.e., the read  
485 support for the alternative allele) standardize the variant scores, allowing comparable assessment of  
486 variants from similar category. We used GATK to re-genotype SNPs and short indels and svtyper (64)  
487 for long deletions (Table S8). Recommended hard filtering was applied to the variants genotyped with  
488 GATK and all re-genotyped variants having a frequency over 75% were retained. Since no tool, to our  
489 knowledge, can be used to genotype long insertions, we used the breakpoints identified by  
490 MindTheGap as evidence of presence.

491 **Back-genotyping**

492 To quantify the uncertainty associated with the variant calling, we implemented an additional layer of  
493 filtering for ambiguous signal. The idea behind the back-genotyping is to identify the “inverse signal”  
494 of a variant to confirm the absence of a variant (Fig S1). The approach consists of (i) generating  
495 multiple modified genomes of the outbreak reference that contain the detected variants, (ii) mapping  
496 the reads of each sample to each of the modified genomes, and (iii) genotyping the variant positions  
497 as for re-genotyping. For SNPs, we genotype the variants as they were introduced. For insertions, we  
498 genotype the putative deletions; for deletions, we genotype the putative insertions.

499 Variants in genomic regions that are difficult to align are expected to exhibit many  
500 uncertainties, i.e. contradicting results between genotyping and back-genotyping. Therefore, we  
501 filtered SNPs if they had more than five uncertainties, and the remaining variants if they had more  
502 than 20 uncertainties. The threshold has been determined in order to limit the incompatible variants  
503 in the final set of variants (Fig S2).

504 **Functionality assignments & enrichment tests**

505 Since functions of MTB genes are determined based on experiments in H37Rv, we first retrieved  
506 homologs between the outbreak reference and H37Rv (NC\_000962.3). For this, we performed a blast  
507 all-vs-all (65) between both sets of protein sequences and retrieved the significant hits (e-value < 1e-  
508 10). After computing the global identity between significant hits with the needle algorithm (66,67),  
509 we considered proteins as homologs if they shared a global identity higher than 30%. We then  
510 assigned the six essentiality categories described in (27) to the homologous proteins in the outbreak  
511 reference and grouped the gene categories into two main ones: essential genes are genes that are  
512 required for growth *in vitro*, and dispensable genes are genes that are not required for the growth *in*  
513 *vitro*. The latter category includes genes annotated as non-essential, conferring growth advantage,  
514 conferring growth defect, uncertain and containing an essential domain and genes without homolog  
515 in H37Rv.

516 A list of genes that were found to confer drug-resistance upon mutations in H37Rv are  
517 additionally used for annotation (Table S2). We identified the homologs in the outbreak reference and  
518 assigned the ABR-conferring or non-ABR-conferring categories accordingly.

519 **Phylogeny inference and evolutionary rate estimation**

520 The back-genotyping allowed us to generate a presence-absence matrix, where uncertainties are  
521 represented as gaps. We estimated the phylogeny based on the presence-absence patterns of the  
522 final variants with iqtree v1.6.1 (68), using the GTR2+FO binary model, and the ultrafast bootstrap.

523 We displayed the tree using iTol v4.3.3 (69). The position of the root was estimated by the least-  
524 squares dating method implemented in LSD (70). Variants that are absent and present at least two  
525 times each in the presence-absence pattern contain information on the phylogenetic relationships,  
526 i.e. they are parsimony informative.

527 We used LSD v0.3beta (70) to estimate evolutionary rates. We followed Menardo et al. (2019)  
528 to determine the significance of the temporal signal (22). For each class of variants, we extracted their  
529 presence-absence sub-matrix, estimated the branch lengths on the pre-estimated phylogeny (iqtree  
530 option –pre), and performed the date-randomization test (DRT) (29). The calculated evolutionary  
531 rates are considered significant with three grades of significance. The stringent test assigns  
532 significance if the calculated confidence intervals of the rate do not overlap with the DRT confidence  
533 intervals. The intermediate test assigns significance if the calculated rates do not overlap with the DRT  
534 confidence intervals. The simple test assigns significance if the calculated rates do not overlap with  
535 the DRT rates (22). The hard limit of 1e-10 is imposed by LSD.

## 536 **Acknowledgements**

537 We thank V. Mohr, F. Boysen and T. Ubben for excellent technical assistance for next generation  
538 sequencing of isolates. We thank Claudio Köser from the University of Cambridge for providing the list  
539 of antibiotic resistance-conferring genes in MTB. We thank Tobias Marschall from the Max-Planck  
540 Institute for Informatics Saarbrücken for discussions on variant calling and genotyping.

## 541 **Funding statement**

542 Parts of the work have been funded by grants from the European Union PathoNgenTrace (FP7-  
543 278864-2) project, from the German Center for Infection Research, from Deutsche  
544 Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy –  
545 EXC 22167-390884018, and grants from the Leibniz Science Campus EvoLUNG.

## 546 **Author contributions**

547 AK, TD, SN, MG, MM, and TAK designed the study; MM, TAK, RD, FM, and SN collected the data; MG  
548 analyzed the data; MG, AK, and TD interpreted the results; MG and AK wrote the manuscript with  
549 contributions from all authors.

## 550 **References**

- 551 1. Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Microbiology*. 2018 Feb 19;(16):202–13.
- 553 2. World Health Organization. Global Tuberculosis Report. WHO report. 2019;

- 554 3. Gygli SM, Borrell S, Trauner A, Gagneux S. Antimicrobial resistance in *Mycobacterium*  
555 tuberculosis: mechanistic and evolutionary perspectives. *FEMS Microbiology Reviews*.  
556 2017 Mar 25;41(3):354–73.
- 557 4. Godfroid M, Dagan T, Kupczok A. Recombination Signal in *Mycobacterium tuberculosis*  
558 Stems from Reference-guided Assemblies and Alignment Artefacts. *Genome Biol Evol*.  
559 2018 Aug 1;10(8):1920–6.
- 560 5. Boritsch EC, Khanna V, Pawlik A, Honoré N, Navas VH, Ma L, et al. Key experimental  
561 evidence of chromosomal DNA transfer among selected tuberculosis-causing  
562 mycobacteria. *Proceedings of the National Academy of Sciences of the United States of*  
563 *America*. 2016 Aug 30;113(35):9876–81.
- 564 6. Eldholm V, Balloux F. Antimicrobial Resistance in *Mycobacterium tuberculosis*: The Odd  
565 One Out. *Trends in Microbiology*. 2016 Aug;24(8):637–48.
- 566 7. Andersson DI, Hughes D. Antibiotic resistance and its cost: is it possible to reverse  
567 resistance? *Nat Rev Microbiol*. 2010 Apr;8(4):260–71.
- 568 8. Telenti A, Imboden P, Marchesi F, Matter L, Schopfer K, Bodmer T, et al. Detection of  
569 rifampicin-resistance mutations in *Mycobacterium tuberculosis*. *The Lancet*. 1993 Mar  
570 13;341(8846):647–51.
- 571 9. Vos M de, Müller B, Borrell S, Black PA, Helden PD van, Warren RM, et al. Putative  
572 Compensatory Mutations in the *rpoC* Gene of Rifampin-Resistant *Mycobacterium*  
573 tuberculosis Are Associated with Ongoing Transmission. *Antimicrobial Agents and*  
574 *Chemotherapy*. 2013 Feb 1;57(2):827–32.
- 575 10. Merker M, Barbier M, Cox H, Rasigade J-P, Feuerriegel S, Kohl TA, et al. Compensatory  
576 evolution drives multidrug-resistant tuberculosis in Central Asia. Storz G, editor. *eLife*.  
577 2018 Oct 30;7:e38200.
- 578 11. Wilson BA, Garud NR, Feder AF, Assaf ZJ, Pennings PS. The population genetics of drug  
579 resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens.  
580 *Mol Ecol*. 2016 Jan 1;25(1):42–66.
- 581 12. Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, Caugant DA, et al.  
582 Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible  
583 ancestor in a single patient. *Genome Biology*. 2014;15:490.
- 584 13. Coll F, Preston M, Guerra-Assunção JA, Hill-Cawthorn G, Harris D, Perdigão J, et al.  
585 PolyTB: A genomic variation map for *Mycobacterium tuberculosis*. *Tuberculosis*. 2014  
586 May 1;94(3):346–54.
- 587 14. McEvoy CRE, Falmer AA, van Pittius NCG, Victor TC, van Helden PD, Warren RM. The  
588 role of IS6110 in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis*. 2007  
589 Sep;87(5):393–404.

- 590 15. Zeng X, Kwok JS-L, Yang KY, Leung KS-S, Shi M, Yang Z, et al. Whole genome sequencing  
591 data of 1110 *Mycobacterium tuberculosis* isolates identifies insertions and deletions  
592 associated with drug resistance. *BMC Genomics*. 2018 May 16;19:365.
- 593 16. Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, et al. GWAS for  
594 quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance  
595 genes and regulatory regions. *Nature Communications*. 2019 May 13;10(1):2128.
- 596 17. Mortimer TD, Weber AM, Pepperell CS. Signatures of Selection at Drug Resistance Loci  
597 in *Mycobacterium tuberculosis*. *mSystems*. 2018 Feb 27;3(1):e00108-17.
- 598 18. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the  
599 biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*.  
600 1998 Jun 11;393(6685):537–44.
- 601 19. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole  
602 genome sequencing of *Mycobacterium tuberculosis* : current standards and open  
603 issues. *Nature Reviews Microbiology*. 2019 Jun 17;1.
- 604 20. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic  
605 analysis identifies targets of convergent positive selection in drug-resistant  
606 *Mycobacterium tuberculosis*. *Nat Genet*. 2013 Oct;45(10):1183–9.
- 607 21. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent  
608 transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection  
609 for the EsxW Beijing variant in Vietnam. *Nature Genetics*. 2018 May 21;1.
- 610 22. Menardo F, Duchêne S, Brites D, Gagneux S. The molecular clock of *Mycobacterium*  
611 tuberculosis. *PLOS Pathogens*. 2019 Sep 12;15(9):e1008067.
- 612 23. Ektefaie Y, Dixit A, Freschi L, Farhat MR. Tuberculosis resistance acquisition in space  
613 and time: an analysis of globally diverse *M. tuberculosis* whole genome sequences.  
614 *bioRxiv*. 2019 Nov 11;837096.
- 615 24. Liu Q, Ma A, Wei L, Pang Y, Wu B, Luo T, et al. China's tuberculosis epidemic stems  
616 from historical expansion of four strains of *Mycobacterium tuberculosis*. *Nature*  
617 *Ecology & Evolution*. 2018 Nov 5;1.
- 618 25. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary  
619 history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nature*  
620 *Genetics*. 2015 Mar;47(3):242–9.
- 621 26. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole Genome  
622 Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium*  
623 tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLOS*  
624 *Medicine*. 2013 Feb 12;10(2):e1001387.

- 625 27. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive  
626 Essentiality Analysis of the *Mycobacterium tuberculosis* Genome via Saturating  
627 Transposon Mutagenesis. *mBio*. 2017 Mar 8;8(1):e02133-16.
- 628 28. Eldholm V, Pettersson JH-O, Brynildsrud OB, Kitchen A, Rasmussen EM, Lillebaek T, et  
629 al. Armed conflict and population displacement as drivers of the evolution and  
630 dispersal of *Mycobacterium tuberculosis*. *PNAS*. 2016 Nov 29;113(48):13881–6.
- 631 29. Duchêne S, Duchêne D, Holmes EC, Ho SYW. The Performance of the Date-  
632 Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data. *Mol Biol*  
633 *Evol*. 2015 Jul 1;32(7):1895–906.
- 634 30. Gómez-Valero L, Silva FJ, Christophe Simon J, Latorre A. Genome reduction of the  
635 aphid endosymbiont *Buchnera aphidicola* in a recent evolutionary time scale. *Gene*.  
636 2007 Mar 1;389(1):87–95.
- 637 31. Singh A. Guardians of the mycobacterial genome: A review on DNA repair systems in  
638 *Mycobacterium tuberculosis*. *Microbiology*. 2017;163(12):1740–58.
- 639 32. Zhang J. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys.  
640 *Nat Genet*. 2006 Jul;38(7):819–23.
- 641 33. Fishbein S, van Wyk N, Warren RM, Sampson SL. Phylogeny to function: PE/PPE protein  
642 evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Molecular*  
643 *Microbiology*. 2015 Jun 1;96(5):901–16.
- 644 34. Wilson DJ, Consortium TCr. GenomeMap: within-species genome-wide dN/dS  
645 estimation from over 10,000 genomes. *bioRxiv*. 2019 Jan 17;523316.
- 646 35. Takayama K, Wang C, Besra GS. Pathway to Synthesis and Processing of Mycolic Acids  
647 in *Mycobacterium tuberculosis*. *Clinical Microbiology Reviews*. 2005 Jan 1;18(1):81–  
648 101.
- 649 36. Solovyev V. V. Solovyev, A Salamov (2011) Automatic Annotation of Microbial  
650 Genomes and Metagenomic Sequences. In *Metagenomics and its Applications in*  
651 *Agriculture, Biomedicine and Environmental Studies* (Ed. R.W. Li), Nova Science  
652 Publishers. In 2011. p. 61–78.
- 653 37. Williams LE, Wernegreen JJ. Sequence Context of Indel Mutations and Their Effect on  
654 Protein Evolution in a Bacterial Endosymbiont. *Genome Biol Evol*. 2013 Mar  
655 1;5(3):599–605.
- 656 38. Reddy PV, Puri RV, Khera A, Tyagi AK. Iron Storage Proteins Are Essential for the  
657 Survival and Pathogenesis of *Mycobacterium tuberculosis* in THP-1 Macrophages and  
658 the Guinea Pig Model of Infection. *Journal of Bacteriology*. 2012 Feb 1;194(3):567–75.
- 659 39. Brown NL, Stoyanov JV, Kidd SP, Hobman JL. The MerR family of transcriptional  
660 regulators. *FEMS Microbiol Rev*. 2003 Jun 1;27(2–3):145–63.

- 661 40. Campbell PJ, Morlock GP, Sikes RD, Dalton TL, Metchock B, Starks AM, et al. Molecular  
662 Detection of Mutations Associated with First- and Second-Line Drug Resistance  
663 Compared with Conventional Drug Susceptibility Testing of *Mycobacterium*  
664 *tuberculosis*. *Antimicrobial Agents and Chemotherapy*. 2011 May 1;55(5):2032–41.
- 665 41. Wang X-D, Gu J, Wang T, Bi L-J, Zhang Z-P, Cui Z-Q, et al. Comparative analysis of  
666 mycobacterial NADH pyrophosphatase isoforms reveals a novel mechanism for  
667 isoniazid and ethionamide inactivation. *Molecular Microbiology*. 2011;82(6):1375–91.
- 668 42. Welzen L de, Eldholm V, Maharaj K, Manson AL, Earl AM, Pym AS. Whole transcriptome  
669 and genomic analysis of extensively drug-resistant *Mycobacterium tuberculosis* clinical  
670 isolates identifies downregulation of *ethA* as a mechanism of ethionamide resistance.  
671 *Antimicrob Agents Chemother*. 2017 Oct 9;AAC.01461-17.
- 672 43. Gröschel MI, Sayes F, Simeone R, Majlessi L, Brosch R. ESX secretion systems:  
673 mycobacterial evolution to counter host immunity. *Nature Reviews Microbiology*. 2016  
674 Nov;14(11):677–91.
- 675 44. McLaughlin B, Chon JS, MacGurn JA, Carlsson F, Cheng TL, Cox JS, et al. A  
676 *Mycobacterium* ESX-1-Secreted Virulence Factor with Unique Requirements for  
677 Export. *PLOS Pathogens*. 2007 Aug 3;3(8):e105.
- 678 45. Mishra SK, Shankar U, Jain N, Sikri K, Tyagi JS, Sharma TK, et al. Characterization of G-  
679 Quadruplex Motifs in *espB*, *espK*, and *cyp51* Genes of *Mycobacterium tuberculosis* as  
680 Potential Drug Targets. *Molecular Therapy - Nucleic Acids*. 2019 Jun 7;16:698–706.
- 681 46. Lynch M. Evolution of the mutation rate. *Trends in Genetics*. 2010 Aug;26(8):345–52.
- 682 47. Trauner A, Liu Q, Via LE, Liu X, Ruan X, Liang L, et al. The within-host population  
683 dynamics of *Mycobacterium tuberculosis* vary with treatment efficacy. *Genome  
684 Biology*. 2017;18:71.
- 685 48. Sung W, Ackerman MS, Dillon MM, Platt TG, Fuqua C, Cooper VS, et al. Evolution of the  
686 Insertion-Deletion Mutation Rate Across the Tree of Life. *G3: Genes, Genomes,  
687 Genetics*. 2016 Aug 1;6(8):2583–91.
- 688 49. Barton NH, Charlesworth B. Why Sex and Recombination? *Science*. 1998 Sep  
689 25;281(5385):1986–90.
- 690 50. Malmberg RL. The Evolution of Epistasis and the Advantage of Recombination in  
691 Populations of Bacteriophage T4. *Genetics*. 1977 Jul 25;86(3):607–21.
- 692 51. Azevedo RBR, Lohaus R, Srinivasan S, Dang KK, Burch CL. Sexual reproduction selects  
693 for robustness and negative epistasis in artificial gene networks. *Nature*. 2006  
694 Mar;440(7080):87–90.
- 695 52. Levin BR, Perrot V, Walker N. Compensatory Mutations, Antibiotic Resistance and the  
696 Population Genetics of Adaptive Evolution in Bacteria. *Genetics*. 2000 Mar  
697 1;154(3):985–97.

- 698 53. Hartkoorn RC, Uplekar S, Cole ST. Cross-Resistance between Clofazimine and  
699 Bedaquiline through Upregulation of MmpL5 in *Mycobacterium tuberculosis*.  
700 *Antimicrobial Agents and Chemotherapy*. 2014 May 1;58(5):2979–81.
- 701 54. Hoffmann H, Kohl TA, Hofmann-Thiel S, Merker M, Beckert P, Jaton K, et al. Delamanid  
702 and Bedaquiline Resistance in *Mycobacterium tuberculosis* Ancestral Beijing Genotype  
703 Causing Extensively Drug-Resistant Tuberculosis in a Tibetan Refugee. *American*  
704 *Journal of Respiratory and Critical Care Medicine*. 2016 Feb 1;193(3):337–40.
- 705 55. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence  
706 data. *Bioinformatics*. 2014 Aug 1;30(15):2114–20.
- 707 56. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler  
708 transform. *Bioinformatics*. 2009 Jul 15;25(14):1754–60.
- 709 57. McKenna N, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The  
710 Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA  
711 sequencing data. *Genome Res*. 2010 Sep 1;20(9):1297–303.
- 712 58. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing.  
713 arXiv:12073907 [q-bio] [Internet]. 2012 Jul 17 [cited 2017 Oct 19]; Available from:  
714 <http://arxiv.org/abs/1207.3907>
- 715 59. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant  
716 discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012 Sep  
717 15;28(18):i333–9.
- 718 60. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to  
719 detect break points of large deletions and medium sized insertions from paired-end  
720 short reads. *Bioinformatics*. 2009 Nov 1;25(21):2865–71.
- 721 61. Wala JA, Bandopadhyay P, Greenwald NF, O'Rourke R, Sharpe T, Stewart C, et al.  
722 SvABA: genome-wide detection of structural variants and indels by local assembly.  
723 *Genome Res*. 2018 Apr 1;28(4):581–91.
- 724 62. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee Y, Wang Z, et al. Accurate de novo and  
725 transmitted indel detection in exome-capture data using microassembly. *Nat Meth*.  
726 2014 Oct;11(10):1033–6.
- 727 63. Rizk G, Gouin A, Chikhi R, Lemaitre C. MindTheGap: integrated detection and assembly  
728 of short and long insertions. *Bioinformatics*. 2014 Dec 15;30(24):3451–7.
- 729 64. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq:  
730 ultra-fast personal genome analysis and interpretation. *Nature Methods*. 2015  
731 Oct;12(10):966–8.
- 732 65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool.  
733 *Journal of Molecular Biology*. 1990 Oct 5;215(3):403–10.

- 734 66. Needleman SB, Wunsch CD. A general method applicable to the search for similarities  
735 in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970  
736 März;48(3):443–53.
- 737 67. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software  
738 Suite. *Trends in Genetics*. 2000 Jun 1;16(6):276–7.
- 739 68. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective  
740 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*.  
741 2015 Jan 1;32(1):268–74.
- 742 69. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new  
743 developments. *Nucleic Acids Res*. 2019 Jan 4;47(W1):W256–9.
- 744 70. To T-H, Jung M, Lycett S, Gascuel O. Fast Dating Using Least-Squares Criteria and  
745 Algorithms. *Syst Biol*. 2016 Jan 1;65(1):82–97.

746