A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region

Authors:

Jean Claude Semuto Ngabonziza[1,2,3#], Chloé Loiseau[4,5#], Michael Marceau[6#], Agathe Jouet[7], Fabrizio Menardo[4,5], Oren Tzfadia[2], Rudy Antoine[6], Esdras Belamo Niyigena[1], Wim Mulders[2], Kristina Fissette[2], Maren Diels[8], Cyril Gaudin[7], Stéphanie Duthoy[7], Willy Ssengooba[9], Emmanuel André[10], Michel K Kaswa[11], Yves Mucyo Habimana[12], Daniela Brites[4,5], Dissou Affolabi[13], Jean Baptiste Mazarati[14], Bouke Catherine de Jong[2], Leen Rigouts[2,3], Sebastien Gagneux[4,5]*, Conor Joseph Meehan[2,15]*, Philip Supply[6]*

Affiliations:

[1]National Reference Laboratory Division, Department of Biomedical Services, Rwanda Biomedical Center, Kigali, Rwanda. [2]Mycobacteriology Unit, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium. [3]Department of Biomedical Sciences, Antwerp University, Antwerp, Belgium. [4]Swiss Tropical and Public Health Institute, Basel, Switzerland. [5]University of Basel, Basel, Switzerland. [6]Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019 - UMR 9017 - CIIL - Center for Infection and Immunity of Lille, F-59000 Lille, France. [7]Genoscreen, Lille, France. [8]BCCM/ITM, Mycobacterial culture collection, Institute of Tropical Medicine, Antwerp, Belgium. [9]Department of Medical Microbiology, College of Health Sciences, Makerere University, Kampala, Uganda. [10]Laboratory of Clinical Bacteriology and Mycology, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium. [11]National Tuberculosis Program, Kinshasa, DR Congo. [12]Tuberculosis and Other Respiratory Diseases Division, Institute of HIV/AIDS Disease Prevention and Control, Rwanda Biomedical Center, Kigali, Rwanda. [13]Laboratoire de Référence des Mycobactéries, Cotonou, Benin. [14]Department of Biomedical Services, Rwanda Biomedical Center. [15]School of Chemistry and Biosciences, University of Bradford, Bradford, UK.

[#]contributed equally

[*]Joint senior authors

Materials and correspondence: Conor Meehan: c.meehan2@bradford.ac.uk, Philip Supply: philip.supply@ibl.cnrs.fr, Sebastien Gagneux: sebastien.gagneux@swisstph.ch

Abstract

The human- and animal-adapted lineages of the *Mycobacterium tuberculosis* complex (MTBC) are thought to have clonally expanded from a common progenitor in Africa. However, the molecular events that accompanied this emergence remain largely unknown. Here, we describe two MTBC strains isolated from patients with multidrug-resistant tuberculosis, representing an as-yet-unknown lineage, named Lineage 8 (L8), seemingly restricted to the African Great Lakes region. Using genome-based phylogenetic reconstruction, we show that L8 is a sister clade to the known MTBC lineages. Comparison with other complete mycobacterial genomes indicate that the divergence of L8 preceded the loss of the *cobF* genome region - involved in the cobalamin/vitamin B12 synthesis - and gene interruptions in a subsequent common ancestor shared by all other known MTBC lineages. This discovery further supports an East African origin for the MTBC and provides additional molecular clues on the ancestral genome reduction associated with adaptation to a pathogenic lifestyle.

45 **Introduction**

46 Tuberculosis (TB), caused by members of the *Mycobacterium tuberculosis* complex (MTBC),
47 is among the ancient scourges of humankind[1], and remains the leading cause of mortality
48 globally due to an infectious disease[2]. Intense research has been dedicated to decipher the
49 evolutionary history of the MTBC and to understand the causes underlying the worldwide
50 spread of TB[3–5]. Current genome data show that the MTBC is comprised of the five human-
51 adapted lineages representing *M. tuberculosis sensu stricto* (L1–4, and L7), two other
52 human-adapted lineages traditionally referred to as *M. africanum* (L5-6) and at least nine
53 animal-adapted lineages[6]. Africa is the only continent where all MTBC lineages are present,
54 suggesting that the MTBC emerged from a common ancestor therein and then clonally
55 expanded to the rest of the world following human migrations[3,7–10]. However, the genomic
56 traits of this common ancestor and the region from which this expansion took place in Africa
57 remain unknown. Whole genome sequencing (WGS) analyses showed that rare human TB
58 bacilli with a smooth colony morphotype, highly restricted to the Horn of Africa and named
59 *Mycobacterium canettii* (alias smooth tubercle bacilli or STB) represent early evolutionary
60 branching lineages that predate the emergence of the most recent common ancestor
61 (MRCA) of the MTBC (or of the rest of the MTBC, if *M. canettii* is considered to be part of the
62 complex)[4,11,12] Indeed, whereas known MTBC strains differ by no more than ~2,000 Single
63 Nucleotide Polymorphisms (SNPs)[13], *M. canettii* strains are 10 to 25-fold more genetically
64 diverse and separated by at least 14,000 SNPs from the hitherto known MTBC MRCA[4,12].
65 Moreover, *M. canettii* strains are less virulent and possess highly mosaic genomes, possibly
66 reflecting adaptation to an environmental reservoir favouring active lateral gene flow[4,14–16].
67 These biological differences support the existence of lineages that reflect intermediate
68 stages in the evolution from *M. canettii* towards the obligate MTBC pathogens.

69 Here, we describe two exceptional strains representing a new clade, diverging before the
70 MRCA of the other MTBC lineages. These two strains were discovered in two independent
71 analyses, and were both multidrug-resistant (MDR; i.e. resistant to at least rifampicin and
72 isoniazid). One was isolated from a TB patient in Rwanda through an ongoing MDR-TB
73 diagnostic trial in Africa. The second isolate was from a patient in Uganda, and was
74 discovered upon screening publicly available draft genome datasets, where it was
75 misclassified as an *M. bovis* strain. We used PacBio and Illumina WGS to reconstruct the full
76 circular genome of the Rwandan strain. We used these data and the available Illumina
77 sequencing data of the Ugandan strain to reconstitute the phylogeny of this novel lineage,
78 which we named Lineage 8 (L8), and further investigate molecular and evolutionary events
79 associated with the emergence of the MTBC.

Results

Patient with the L8 TB strain in Rwanda

The strain in Rwanda was isolated from a male patient, aged 77 years, HIV-negative, resident of Rulindo district bordering with the Southwest of Uganda, and who had lived in Uganda previously. The patient was diagnosed with rifampicin-resistant TB based on standard Xpert MTB/RIF testing (Xpert; Cepheid, Sunnyvale, CA, USA), which probes for mutations in the rifampicin resistance-determining region of the *rpoB* gene of MTBC strains[17]. The results of the assay showed a rare delayed probe B reaction (~3% prevalence in Rwanda)[18], presumed (and later confirmed; see below) to be due to the rifampicin resistance-associated Asp435Tyr mutation[19].

Per routine practice, the patient was initiated on standard short-course MDR-TB treatment (i.e. 9-month WHO-endorsed MDR-TB regimen, including moxifloxacin, kanamycin, protionamide, ethambutol, clofazimine, high dose isoniazid and pyrazinamide)[20]. However, the patient developed hypotension, and eventually died due to probable cardiac failure, after 20 days of treatment. Phenotypic drug-susceptibility testing (DST) confirmed resistance to both rifampicin and isoniazid, and susceptibility to other anti-TB drugs including ethambutol, fluoroquinolones, and second-line injectables.

Growth characteristics and biochemical properties of the Rwandan strain

The strain from Rwanda was grown in 12.5 days on Mycobacterial Growth Indicator Tubes. Colonies were observed on the fifth week after initial inoculation on Löwenstein-Jensen medium, indicating a slow grower phenotype with rough colonies (Figure 1). The strain also displayed archetypal biochemical characteristics of *M. tuberculosis sensu stricto*, including niacin production combined with urease hydrolysis (Table 1).

103 **Genotypic resistance and SNP profile of the Rwandan strain by Deeplex Myc-TB**

104 Following the MDR-TB diagnosis, the strain was included in the first set of tests for an
105 ongoing MDR-TB diagnostic trial "DIAgnostics for MDR-TB in Africa (DIAMA) Clinicaltrials.gov,
106 NCT03303963", evaluating a new targeted deep sequencing assay, called Deeplex-MycTB.
107 Deeplex-MycTB testing confirmed the presence of the *rpoB* Asp435Tyr mutation conferring
108 rifampicin resistance, along with the *inhA* Ser94Ala mutation conferring isoniazid resistance,
109 consistently with the MDR phenotype identified by phenotypic DST (**Figure 2**). This strain
110 also harboured two alleles in phylogenetic positions in *embB* (Ala378) and *gidB* (Ala205) not
111 associated with resistance to ethambutol or streptomycin, which were both shared by
112 several MTBC lineages (L1, 5, 6, 7, and animal lineages) and *M. canettii*[21]. In addition, nine
113 other - so far uncharacterized - SNPs were identified in six of the 18 gene targets
114 interrogated by the assay (**Figure 2**). Moreover, this test detected an atypical spoligotype
115 pattern, 111110000000000000000000000000000001110000000 (**Figure 2**), which was further
116 confirmed by conventional membrane-based spoligotyping testing. This spoligotype pattern
117 was unique in the global spoligotype database that comprises 111,637 MTBC isolates from
118 131 countries[22].

119 **WGS analysis and phylogenetic position of the Rwandan and Ugandan strains**

120 Results from WGS analysis of the Rwandan strain using Illumina sequencing confirmed all
121 Deeplex-MycTB findings.

122 The strain isolated in Uganda was discovered independently upon screening global MTBC
123 genome datasets publicly available on NCBI/EBI, comprising approximately 20,000 genomes
124 in total. From subsequent processing with our WGS analysis pipeline, we found 1 genome
125 that did not classify in any of the 7 human-adapted lineages or 9 animal-adapted ecotypes
126 known at the time, but had been misclassified as *M. bovis* isolated from a human patient[23].

127 These WGS data revealed a similar spoligotype

128 111110000000000000000000000000000001111000111, characterized by the presence of

129 spacers 1 to 5 and 34 to 37 (vs 34-36 in the Rwandan strain) with all intervening spacers

130 missing. Moreover, the Ugandan strain also shared the same *rpoB* Asp435Tyr and *inhA*

131 Ser94Ala mutations and the same sequence alleles in *embB* and *gidB*. The Ugandan strain

132 contained an additional *katG* Ser315Thr mutation conferring isoniazid resistance, as well as

133 the *embA* C-11A and *embB* Asp328Tyr mutations, and two *pncA* missense mutations,

134 predictive of pyrazinamide resistance. Moreover, only three of the nine aforementioned

135 uncharacterized SNPs detected by Deeplex-MycTB were shared between both strains.

136 To further assess the relationships between both strains and in comparison to other MTBC

137 strains, a maximum likelihood phylogeny was inferred from 241 MTBC genomes, including

138 representatives of all known human- and animal- adapted lineages[6] and using a *M. canettii*

139 strain as an outgroup. *M. canettii* represents the closest outgroup to the MTBC including L8,

140 as shown by subsequent comparative analysis of a complete L8 genome (see below), and

141 previous observations of ~2.0 Mb larger genomes and substantially lower average nucleotide

142 identities of phylogenetically closest non-tuberculous mycobacterial species such as *M.*

143 *marinum* and *M. kansasii*[4,24,25]. This reconstruction revealed a unique phylogenetic position

144 of the two new genomes from Rwanda and Uganda (**Figure 3**), representing a newly

145 characterised monophyletic clade in which none of the known MTBC genomes are

146 contained. A core genome-based phylogeny was also constructed from representatives of

147 the MTBC lineages as well as *M. canettii, M. marinum* and *M. kansasii* (Supplementary figure

148 1). This phylogeny confirmed the placement of the L8 Rwandan strain as being a new clade

149 between *M. canettii* and the other lineages of the MTBC. Based on these phylogenies, this

150 clade shares a MRCA with the rest of the MTBC, thus representing a new sister clade to the

151 known MTBC, which we named Lineage 8 (L8). Comprehensive SNP analysis identified a total
152 of 189 SNPs separating both L8 genomes, which is within the range of zero to 700 SNPs
153 found between any two strains within any of the lineages 1 to 7 of the MTBC[13]. On average,
154 two strains of the MTBC including L8 differed by 1443 SNPs (corresponding to 0.04% of the
155 genome, excluding repetitive/'problematic' regions), thus at least an order of magnitude
156 lower than the SNP distance separating any MTBC strain from *M. canettii*[4,13].
157 The absence of any matching pattern in the global spoligotype database, as well as the lack
158 of detection of this clade in previous large WGS datasets of MTBC strains from global
159 sources, suggests that L8 is rare and seemingly geographically restricted to the African Great
160 Lakes region. Specifically, the L8 spoligotype signature and the 3 SNPs shared by both L8
161 strains were not detected in any of 115 MTBC genomes from a previous drug resistance
162 survey in Uganda[26], nor in 380 rifampicin-resistant strains from Rwanda collected between
163 1991 and 2018, from routine drug resistance surveillance as well as various drug resistance
164 surveys[27–29]. Furthermore, among 14 other available isolates out of 27 from Uganda and
165 Rwanda tested by Gene Xpert MTB/RIF that showed the same delayed probe B as L8, none
166 displayed the L8 signatures when tested by Deeplex Myc-TB or by classical spoligotyping.
167 Likewise, none of > 1,500 clinical samples from TB patients tested by Deeplex-MycTB from a
168 recent nationwide drug resistance survey performed in the Democratic Republic of Congo
169 displayed the L8 spoligotype signature or the specific SNPs.

170 **Defining features of a complete L8 genome**

171 To further assess the sister position of L8 and split from the remaining MTBC inferred from
172 SNP analysis, the Rwandan strain was subjected to long read-based PacBio sequencing.
173 Comparison of the obtained assembly with 36 available complete genomes of MTBC
174 members, comprising L1-L4 (including H37Rv), *M. africanum* (L6) and *M. bovis* strains,

175 showed a highly syntenic organization, with no major structural rearrangement between
176 both groups. Although the assembled L8 genome of 4,379,493 bp was within the 4.34-4.43
177 Mb size range of the other MTBC genomes, it was 30 kb smaller than the 4.41-Mb mean size
178 of genomes of *M. tuberculosis sensu stricto*[30]. However, the largest part of this gap was
179 accounted by the absence of three genomic regions in L8, corresponding to regions of
180 difference (RDs) known to be variably present or absent in other MTBC
181 (sub)lineages[31,32](Supplementary Table 2). These include a 9.3-kb PhiRv1 prophage region
182 (RD3), as well as 10.0-kb and 8.5-kb segments corresponding to RD14 and RD5, comprising
183 the *plcABC* gene cluster and the *plcD* gene regions, respectively[31]. In L8, each of the two
184 latter regions only contained one copy of the IS*6110* insertion sequence, devoid of direct
185 repeats (DRs) that normally flank IS*6110* copies after transposition, indicating that these
186 deletions in L8 resulted from recombination between two adjacent IS*6110* copies with loss
187 of the intervening sequences[33]. These mobile DNA-related deletions, which also arose
188 independently in several other MTBC branches[31,34], probably occurred after the divergence
189 of L8 from the other MTBC lineages. Apart from these three deletions and two dozen
190 repetitive/multicopy genes (IS6110-related, PE/PPE-, or Mce-encoding), we only found 5
191 non-repetitive genes, included in two small segments (3.4 and 4.4 kb), which were
192 undetected in the complete L8 genome while being present in reference MTBC genomes
193 (Supplementary Table 2).
194 Conversely, a particular 4.4 kb genome region was present in the genomes of both L8 strains
195 and in *M. canettii*, but absent in the 36 available complete genomes of MTBC members
196 (Supplementary Table 3). This region comprises the *cobF* gene (**Figure 4**), encoding the
197 precorrin 6A synthase involved in the cobalamin/vitamin B12 synthesis, along with two other
198 genes, respectively encoding a PE-PGRS protein family member and a protein of unknown

199 function. This region shared by the L8 and the *M. canettii* genomes is also present in the
200 phylogenetically proximal non-tuberculous mycobacterial species *M. marinum* and *M.*
201 *kansasii* (Supplementary Table 3). Via BLAST analysis, we further confirmed the systematic
202 absence of *cobF* in any of 6,456 quality draft genome assemblies available as of January 2020
203 from the NCBI, from strains belonging to lineages 1 to 7 or the animal lineages of the MTBC.
204 Moreover, we thereby determined that the junction between the sequences flanking the
205 *cobF* deletion was at the same nucleotide position in all but 6 of these genomes, resulting in
206 the truncation of *rv0943c* and *rv0944* genes as seen in the complete MTBC genomes (Fig. 4).
207 Consistent with the clonal evolution of the MTBC with negligible, if any, horizontal gene
208 transfer between strains[1,14,32,35] this perfect conservation of this sequence junction suggests
209 that *cobF* was lost in the MRCA of the other MTBC lineages, after its divergence from L8. The
210 6 exceptions were 3 strains from lineage 4.3 and 3 from lineage 3 that showed slightly larger
211 deletions, including the 5' region of *rv0943c* or the 5' region of *rv0943c*, *rv0944* and the 5'
212 region of *rv0945*, respectively, suggesting probable subsequent deletion events in particular
213 sub-branches of these lineages.
214 In contrast, none of the almost 900 other genes specifically identified in the *M. canettii*
215 genomes, and absent in the other MTBC genomes, were found in the L8 genome. The latter
216 finding thus supports the close relationship with the previously known MTBC lineages
217 indicated by the SNP-based phylogeny, as well as the outgroup position of *M. canettii*
218 relatively to the MTBC including L8.
219 Further evidence for the early branching of L8 relative to the rest of the MTBC comes from
220 examination of interrupted coding sequences (ICDSs). These ICDSs correspond to frameshifts
221 or in-frame stop codons detected in genes originally intact in a common progenitor, thus
222 putatively representing so-called molecular scars inherited during progressive

223 pseudogenization of the MTBC genomes[36,37]. Four orthologues of MTBC ICDSs were
224 previously found to be intact in the genomes of *M. canettii* strains, as well as in *M. marinum*
225 and *M. kansasii* [4]. One of these four orthologues (*pks8*), which belongs to a multigene family
226 encoding polyketide synthases involved in the biosynthesis of important cell envelope
227 lipids[38], was also intact in the genomes of both L8 strains (Figure 5 and Supplementary Table
228 4). Moreover, we found an additional orthologue of MTBC ICDSs (i.e. *rv3899c-rv3900c*),
229 coding for a conserved hypothetical protein, which was intact in the genomes of *M. canettii*,
230 *M. kansasii*, *M. marinum* and both L8 strains (Supplementary Table 4). These two molecular
231 scars were also likely acquired by the other MTBC lineages after their divergence from the
232 common progenitor shared with L8.
233 The assembled L8 genome also included 48 out of 50 genes (the exceptions are rv3513c
234 encoding the probable fatty-Acid-Coa ligase FadD18 and the PhiRv1 region; see above)
235 present in MTBC members but not found in any of the *M. canettii* genomes, including a
236 number of genes putatively acquired through horizontal gene transfer by the common
237 ancestor of the MTBC after its separation from *M. canettii*[4] (Supplementary Table 5). This
238 observation additionally supports both the close relationship with the previously known
239 MTBC lineages and the outgroup position of *M. canettii* relatively to the MTBC including L8.
240 Likewise, consistent with the rough colony morphotype of the Rwandan strain, both L8
241 strains displayed the single polyketide-synthase-encoding *pks5* gene configuration shared by
242 all MTBC members, instead of the dual *pks5* conformation found in *M. canettii* strains
243 involved in the smooth colony phenotype of the latter strains[15]. Thus, the recombination
244 between the two *pks5* genes and the loss of the intervening *pap* gene, thought to have
245 resulted in surface remodelling and incremental gain of virulence after the phylogenetic
246 separation from *M. canettii*[15], already existed in the common progenitor of L8 and the rest

247 of the MTBC. Moreover, both L8 strains also contained the intact TbD1 and RD9 regions,
248 shared by the other "ancestral" *M. tuberculosis* lineages (L1, L7) but subsequently lost by the
249 so-called "modern" lineages of *M. tuberculosis* (TbD1 lost in L2-4), *M. africanum* (L5 and L6)
250 and the animal lineages (RD9)[31].
251 In contrast to the highly clonal population structure of the MTBC, *M. canettii* strains are
252 highly recombinogenic, as apparent from mosaic sequence arrangements in their genomes
253 and functional DNA transfer between *M. canettii* strains mediated by a distributive conjugal
254 transfer (DCT)-like mechanism[4,39]. However, no significant genome-wide recombination
255 signal was detected by ClonalFrameML analysis[40] between L8 and other MTBC strains
256 (Supplementary Figure 2). In particular, and in contrast to the numerous recombination
257 segments in *M. canettii*[4], the complete L8 genome only contained 26 possible recombination
258 segments, yet the longest of these was 607bp and the average length was 142bp.

259 **Discussion**

260 The discovery of L8 provides unique insights into an ancestor of the MTBC that existed after
261 the *pks5*-recombination-mediated surface remodelling, which occurred after separation of
262 the MTBC MRCA from the *M. canettii* clade, but preceded the loss of the *cobF* region and
263 gene interruptions in a later common ancestor of the other MTBC lineages. The seeming
264 restriction of this lineage to the African Great Lakes region represents new evidence
265 supporting an origin for the MTBC in the eastern part of the African continent. These
266 findings reinforce results from previous work suggesting an East- rather than a West African
267 origin of the MTBC[3,4,9,11,41].

268 A distinct ecological niche, linked to a potential environmental reservoir, has been
269 hypothesized to explain the marked geographic restriction of *M. canettii* strains to the Horn
270 of Africa, the lower persistence of these strains in infection models as well as their genome
271 mosaicism implying multiple DNA recombination events within the *M. canettii* strain pool[4, 12].
272 However, although our analysis is limited to two genomes identified to date, our results
273 suggest that L8 is as clonal as the rest of the MTBC[3,14,34,42]. Moreover, the observation that
274 both L8 strains share two uncommon rifampicin- and isoniazid-resistance conferring
275 mutations in *rpoB* and *inhA* suggests that multidrug resistance was already acquired in the
276 common ancestor of these two strains. Isoniazid and rifampicin were introduced in TB
277 treatments in the African Great Lakes region in the late fifties and 1983, respectively (Dr.
278 Armand Van Deun, personal communication). These shared MDR-defining mutations, and
279 the detection of these isolates in human patients in both cases (with reported absence of
280 previous TB history for the Rwandan patient), suggest that these patients were infected with
281 an already-resistant strain, which was exposed to drug selective pressure already decades
282 ago and had been likely circulating in the community for some time. Overall, this pattern

283 thus suggests human-to-human transmission rather than infection from a non-human
284 source. While based on only two initial strains, these results are consistent with the
285 presumed scenario of a human rather than a zoonotic origin for the MTBC[31,43].
286 Of note, given the above timeline of introduction of rifampicin and isoniazid in both
287 countries, the ~100 SNPs distance separating these two strains from their MRCA would imply
288 a rapid molecular clock for L8, in the range of the upper bound of 2.2 SNPs/genome/year
289 most recently estimated for other MTBC clades[44]. However, this mutation rate cannot be
290 confirmed until additional L8 samples are uncovered.
291 Remarkably, the absence of other L8 strains in datasets from Uganda, Rwanda and DRC,
292 together comprising more than 2,000 strains, suggests that L8 is rare even within the African
293 Great Lakes region. Such scarcity is compatible with selective sweeps of later branching
294 MTBC strains, introduced more recently into the region. Similar scenarios have also been
295 proposed to explain the slow apparent replacement of MTBC L5 and L6 by L4 in West
296 Africa[45–47] and the restriction of L7 to Northern Ethiopia[48].
297 Loss-of-function linked to the deletion of *cobF* is a plausible candidate molecular event
298 involved in such a replacement scenario for L8. Indeed, loss-of-function appears to be an
299 important mechanism driving the pathoadaptive evolution of the TB pathogen, as shown for
300 the role of the loss of lipo-oligosaccharide production (via recombination in the *pks5* locus)[15]
301 in the evolution towards increased virulence from *M. canettii* to MTBC strains. Likewise, loss
302 of secretion of PPE-MPTR and PE_PGRS proteins by the type VII secretion system ESX-5 (via
303 mutations of the *ppe38* locus) has been involved in the hypervirulence of recent branches of
304 L2 (alias "modern" Beijing) strains[49]. The loss of the *cobF* region in the other MTBC lineages,
305 inferred from comparative genomics with *M. canettii* and non-tuberculous mycobacteria[4],
306 was previously hypothesised to reflect enhanced adaptation to an intracellular parasitic

307 lifestyle[50]. Indeed, the cobalamin/vitamin B12 synthesis pathway, of which the *cobF*-
308 encoded precorring-6a synthase is a component, represents a highly complex and energy
309 consuming process with about 30 enzymatic steps[51]. While the absence of this component
310 may not entirely ablate cobalamin biosynthesis[52,53], its loss might have resulted in gain of
311 fitness and reflect enhanced pathogenic professionalisation, by economical reliance upon
312 the mammalian host environment as source of vitamin B12. As an additional plausible but
313 not necessarily mutually exclusive hypothesis, such selective sweep of L8 might have been
314 (further) enhanced by the loss of theTbD1 region in later branching MTBC strains. This
315 region, which we also found intact in L8, as is the case in the "ancestral" *M. tuberculosis*
316 lineages L1 and L7, encodes members of the mycobacterial membrane protein families
317 MmpL. Very recent findings indicate that the loss of this region in later branching MTBC
318 strains was also associated with a gain of virulence, and the deletion of TbD1 at the origin of
319 the "modern" *M. tuberculosis* lineages L2/L3/L4 has therefore been suggested as a key driver
320 for their global epidemic spread[54]. If true, more recently emerged or introduced *cobF*- and
321 TbD1-deleted strains might conceivably have largely outcompeted L8 strains. This hypothesis
322 could be explored by assessing the growth and the virulence/fitness in cellular and animal
323 models, of recombinant *cobF*- and/or TbD1-knock-out, as well as *cobF*- and/or TbD1-knock-
324 in strains, derived from the available Rwandan L8 strain and other MTBC strains,
325 respectively.
326 In conclusion, our genomic data, on an as-yet-unknown ancestral stage between the MTBC
327 and the putative progenitor pool of *M. canettii*-like mycobacteria, thus suggest further
328 experiments to examine candidate molecular events potentially involved in the
329 pathoadaptive evolution of *M. tuberculosis*. The discovery of such rare strains raises the
330 possibility for the existence of further extant strains, especially in Eastern Africa,

representing other clades further closing the biological gap between the MTBC and *M. canettii.*

333 **Methods**

334 **Rwandan patient recruitment and ethics statement**

335 The patient in Rwanda was recruited into, and gave informed consent for, the DIAMA study which
336 was approved by the Rwanda National Ethical committee (IRB 00001497 of IORG0001100; Ref
337 No.0069/RNEC/2017).

338 **Phenotypic characterization**

339 We studied conventional mycobacterial growth and biochemical characteristics including
340 colony morphology, niacin production, nitrate reduction, p-nitro benzoic acid growth
341 inhibition, catalase production, urea hydrolysis, tween 80 hydrolysis, and thiophene
342 carboxylic acid hydrazide growth inhibition[55]. For comparative purpose, a reference set of
343 the seven known human-adapted MTBC lineages[56], together with *M. canettii*
344 (BCCM/ITM500199), *M. bovis* (BCCM/ITM500324), *M. bovis BCG* (BCCM/ITM500101), and *M.*
345 *orygis* (BCCM/ITM501026) strains were processed with the novel strain isolated in Rwanda.
346 Moreover, phenotypic drug-susceptibility testing to first- and second-line anti-TB drugs was
347 done using the proportion method[57]. The strain isolated in Uganda was not available for
348 phenotypic characterization.

349 **Targeted and whole genome sequencing of the Rwandan strain**

350 Targeted sequencing was performed by using the Deeplex-MycTB assay[58] (Genoscreen,
351 France). Briefly, this assay relies on a 24-plexed PCR amplification of mycobacterial species
352 identification (*hsp65*), genotyping (spoligotyping and phylogenetic single nucleotide
353 polymorphisms (SNPs)) and 18 *M. tuberculosis* complex drug resistance-associated gene
354 targets. This test and short-read Illumina-based WGS were performed on the Rwandan strain
355 as follows. A bead beating method was used to extract DNA from colonies (Supplementary
356 method 1). Libraries of Deeplex-MycTB amplicons or genome fragments were constructed

357 using the Nextera XT kit and sequenced on an Illumina MiSeq platform with paired end, 150-
358 bp read lengths (Illumina, CA, USA). DNA extraction suitable for PacBio SMRT sequencing
359 was performed using the Genomic DNA Buffer Set (Qiagen Inc, Germantown, Maryland,
360 USA) (Supplementary method 2). Sequencing was performed on a PacBio RS II using the
361 SMRT technology.

362 **Deeplex-MycTB analysis and spoligotyping**

363 Analysis of the Deeplex-MycTB sequencing data, including SNP calling and spoligotype
364 identification, was performed by read mapping on *M. tuberculosis* H37Rv sequence
365 references, using a parameterized web application (GenoScreen)[58]. Membrane-based
366 spoligotyping was performed as described previously[59].

367 **Illumina whole genome sequencing analysis**

368 Raw genomic reads from the newly sequenced L8 genome from Rwanda and the L8 genome
369 from Uganda (SAMN02567762) were processed as previously described[60]. Briefly, the reads
370 were trimmed with Trimmomatic v0.33.22[61] and reads larger than 20 bp were kept. The
371 software SeqPrep (https://github.com/jstjohn/SeqPrep) was used to identify and merge any
372 overlapping paired-end reads. The resulting reads were aligned to the reconstructed
373 ancestral sequence of the MTBC[62] using the mem algorithm of BWA v0.7.13[63]. Duplicated
374 reads were marked using the MarkDuplicates module of Picard v2.9.1
375 (https://github.com/broadinstitute/picard) and local realignment of reads around InDels was
376 performed using the RealignerTargetCreator and IndelRealigner modules of GATK v3.4.0[64].
377 SNPs were called with Samtools v1.2 mpileup[65] and VarScan v2.4.1[66] using the following
378 thresholds: minimum mapping quality of 20, minimum base quality at a position of 20,
379 minimum read depth at a position of 7X, maximum strand bias for a position 90%.

380 The spoligotype pattern of the strain from Uganda was extracted *in silico* from the raw reads

381 using kvarQ[67].

382 **Phylogenetic reconstruction**

383 The maximum likelihood phylogeny was inferred with RAxML v.8.2.8[68] using an alignment
384 containing only polymorphic sites and the branch lengths of the tree were rescaled using
385 invariant sites (rescaled_branch_length = (branch_length * alignment_length) /
386 (alignment_length +invariant_sites))[44,69].
387 A position was considered polymorphic if at least one genome had a SNP at that position.
388 Deletions and positions not called according to the minimum threshold of 7x were encoded
389 as gaps. We excluded positions with more than 20% missing data, positions falling in PE-
390 PGRS genes, phages, insertion sequences and in regions with at least 50 bp identity to other
391 regions in the genome. We also excluded variable positions falling in drug resistance-related
392 genes. The phylogeny was computed using the general time-reversible model of sequence
393 evolution (-m GTRCAT -V options), 100 bootstrap inferences and *M. canettii* (SRR011186)
394 was used as an outgroup to root the phylogeny.

395 **Whole genome *de novo* assembly, annotation and comparative genomics**

396 Raw PacBio reads obtained from the Rwandan strain were assembled with Canu v1.6[70], using
397 default settings and an expected genome size of 4.4 Mbp, typical of MTBC strains. After
398 discarding 60,272 reads below minimal quality parameters, 106,681 reads were used for the
399 assembly. Based on the expected genome size, the average coverage depth was estimated at
400 186x using raw reads, and 39x and 38x using corrected and trimmed reads, respectively. The
401 obtained unique contig of 4,387,285 bp was circularized with Circlator v1.5.5[71] using default
402 settings, resulting in an assembly of 4,379,493 bp. Additional sequence verification and
403 correction was then performed by mapping Illumina reads obtained from the same strain,
404 using pacbio-utils version 0.2[72] (https://github.com/douglasgscofield/PacBio-utilities) and snippy
405 version 4.4[73] (https://github.com/tseemann/snippy). Alignments of the final assembly were

406 performed against an ensemble of complete genome sequences available from 38 strains of
407 tubercle bacilli. This set included 34 *M. tuberculosis* strains from lineages 1, 2, 3 and 4
408 (comprising H37Rv), *M. africanum* L6 GM041182, *M. bovis* AF2122/97, as well as the closest
409 STB-A (CIPT 140010059) and most distant (STB-K) *M. canettii* strains (Supplementary Table
410 1). Comparative alignments and genome annotation were performed based on BLAST
411 searches and analysis of gene synteny, using Artemis and Artemis comparison tool[74], as well
412 as a custom Multiple Annotation of Genomes and Differential Analysis (MAGDA) software
413 previously used for annotation of *M. canettii* and *Helicobacter pylori* genomes[4,75].
414 Comparisons with orthologues from *M. canettii* STB-D, -E, -G, -H, -I, and -J in addition to STB-
415 A and -K, and from *M. marinum* type strain M and *M. kansasii* genomes were additionally
416 done using the Microscope platform v3.13.3[76]. When applicable, annotations were
417 transferred from those of *M. tuberculosis* or *M. canettii* orthologs in the
418 TubercuList/Mycobrowser database, using BLAST matches of > 90% protein sequence
419 identity, an alignable region of >80% of the shortest protein length in pairwise comparisons
420 and visual inspection of the gene synteny. Genome completeness was assessed using
421 CheckM[77] using the lineage-specific workflow and *Mycobacterium* as the genus. The PacBio
422 assembled Rwandan strain was found to have 98.74% completeness and 0% contamination
423 or strain heterogeneity, making it suitable for further analyses.
424 ACT comparison files were generated using MAUVE 2015-02-25 software to visualize the
425 genome-wide distribution of SNP densities between the assembled L8 genome from Rwanda
426 and *M. tuberculosis* H37Rv and *M. canettii* STB-A and STB-K genomes. Recombination
427 between L8 and other MTBC lineages or *M. canettii* was assessed from a progressive MAUVE
428 alignment of the PacBio assembled L8 genome and previously published closed genomes[73]
429 using ClonalFrameML[40]. To further assess the phylogenetic placement of the Lineage 8

430 strain, a core gene alignment was constructed using the completed genomes of the
431 Rwandan L8 strain, representatives of the MTBC lineages 1-4 and 6, *M. canettii, M. bovis, M.*
432 *marinum* (accession GCF_000419315.1) and *M. kansassii* (accession GCF_000157895.3). The
433 GFF files of each genome were input to roary[78] with an 80% identity cut-off, as has been
434 done for previous genus-level mycobacterial core alignments[79]. A phylogenetic tree was
435 constructed from the core gene superalignment using RAxML-NG v 0.9.0[80] under the
436 GTR+Gamma model of evolution with 20 starting trees. Bootstrapping was run until
437 autoMRE converged with a value of 0.03 (50 replicates).

438 **Accession codes**

439 The complete genome sequence of the L8 strain from Rwanda was deposited in the NCBI
440 repository under project PRJNA598991 with SRR10828835 and SRR10828834 accession
441 codes for Illumina- and PacBio-derived genome sequences, respectively.

442 **References**

443 1. Gagneux, S. Ecology and evolution of Mycobacterium tuberculosis. *Nature Reviews*
444 *Microbiology* **16**, 202–213 (2018).
445 2. World Health Organization. *Global Tuberculosis Report 2018*. (2018).
446 3. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium
447 tuberculosis with modern humans. *Nature genetics* **45**, 1176–82 (2013).
448 4. Supply, P. *et al.* Genomic analysis of smooth tubercle bacilli provides insights into
449 ancestry and pathoadaptation of Mycobacterium tuberculosis. *Nature genetics* **45**,
450 172–9 (2013).
451 5. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New
452 World human tuberculosis. *Nature* **514**, 494–7 (2014).
453 6. Brites, D. *et al.* A New Phylogenetic Framework for the Animal-Adapted
454 Mycobacterium tuberculosis Complex. *Frontiers in Microbiology* **9**, 2820 (2018).
455 7. Rutaihwa, L. K. *et al.* Multiple Introductions of Mycobacterium tuberculosis Lineage 2–
456 Beijing Into Africa Over Centuries. *Frontiers in Ecology and Evolution* **7**, 112 (2019).
457 8. O'Neill, M. B. *et al.* Lineage specific histories of Mycobacterium tuberculosis dispersal
458 in Africa and Eurasia. *Molecular Ecology* **28**, 3241–3256 (2019).
459 9. Wirth, T. *et al.* Origin, spread and demography of the Mycobacterium tuberculosis
460 complex. *PLoS pathogens* **4**, e1000160 (2008).
461 10. Gagneux, S. *et al.* Variable host-pathogen compatibility in Mycobacterium
462 tuberculosis. *Proceedings of the National Academy of Sciences of the United States of*
463 *America* **103**, 2869–73 (2006).
464 11. Gutierrez, M. C. *et al.* Ancient origin and gene mosaicism of the progenitor of
465 Mycobacterium tuberculosis. *PLoS pathogens* **1**, e5 (2005).

466 12. Blouin, Y. *et al.* Progenitor "Mycobacterium canettii" clone responsible for lymph
467 node tuberculosis epidemic, Djibouti. *Emerging infectious diseases* **20**, 21–8 (2014).
468 13. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in Mycobacterium
469 tuberculosis. *Seminars in Immunology* **26**, 431–444 (2014).
470 14. Boritsch, E. C. *et al.* Key experimental evidence of chromosomal DNA transfer among
471 selected tuberculosis-causing mycobacteria. *Proceedings of the National Academy of*
472 *Sciences of the United States of America* **113**, 9876–81 (2016).
473 15. Boritsch, E. C. *et al.* pks5-recombination-mediated surface remodelling in
474 Mycobacterium tuberculosis emergence. *Nature Microbiology* **1**, 15019 (2016).
475 16. Mortimer, T. D. & Pepperell, C. S. Genomic signatures of distributive conjugal transfer
476 among mycobacteria. *Genome biology and evolution* **6**, 2489–500 (2014).
477 17. Helb, D. *et al.* Rapid detection of Mycobacterium tuberculosis and rifampin resistance
478 by use of on-demand, near-patient technology. *Journal of Clinical Microbiology* **48**,
479 229–237 (2010).
480 18. Ng, K. C. S. *et al.* Automated algorithm for early identification of rifampicin-resistant
481 tuberculosis transmission hotspots in Rwanda [abstract]. *The International Journal of*
482 *Tuberculosis and Lung Disease* **22**, 605 (2018).
483 19. Torrea, G. *et al.* Variable ability of rapid tests to detect Mycobacterium tuberculosis
484 rpoB mutations conferring phenotypically occult rifampicin resistance. *Scientific*
485 *Reports* **9**, (2019).
486 20. Trébucq, A. *et al.* Treatment outcome with a short multidrug-resistant tuberculosis
487 regimen in nine African countries. *The International Journal of Tuberculosis and Lung*
488 *Disease* **22**, 17–25 (2018).
489 21. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex

490 strains. *Nature Communications* **5**, (2014).

491 22. Couvin, D., David, A., Zozio, T. & Rastogi, N. Macro-geographical specificities of the
492 prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the
493 Mycobacterium tuberculosis genotyping database. *Infection, Genetics and Evolution*
494 (2018). doi:10.1016/j.meegid.2018.12.030

495 23. Wanzala, S. I. *et al.* Retrospective Analysis of Archived Pyrazinamide Resistant
496 Mycobacterium tuberculosis Complex Isolates from Uganda-Evidence of Interspecies
497 Transmission. *Microorganisms* **7**, (2019).

498 24. Stinear, T. P. *et al.* Insights from the complete genome sequence of Mycobacterium
499 marinum on the evolution of Mycobacterium tuberculosis. *Genome Research* **18**, 729–
500 741 (2008).

501 25. Wang, J. *et al.* Insights on the emergence of Mycobacterium tuberculosis from the
502 analysis of Mycobacterium kansasii. *Genome biology and evolution* **7**, 856–70 (2015).

503 26. Ssengooba, W. *et al.* Whole genome sequencing to complement tuberculosis drug
504 resistance surveys in Uganda. *Infection, genetics and evolution : journal of molecular*
505 *epidemiology and evolutionary genetics in infectious diseases* **40**, 8–16 (2016).

506 27. Carpels, G. *et al.* Drug resistant tuberculosis in sub-Saharan Africa: an estimation of
507 incidence and cost for the year 2000. *Tubercle and lung disease : the official journal*
508 *of the International Union against Tuberculosis and Lung Disease* **76**, 480–6 (1995).

509 28. Umubyeyi, A. N. *et al.* Results of a national survey on drug resistance among
510 pulmonary tuberculosis patients in Rwanda. *The international journal of tuberculosis*
511 *and lung disease : the official journal of the International Union against Tuberculosis*
512 *and Lung Disease* **11**, 189–94 (2007).

513 29. World Health Organization & WHO. WHO | Global tuberculosis report 2016. *WHO*

514 (2019).

515 30. Yang, T. *et al.* Pan-genomic study of Mycobacterium tuberculosis reflecting the
516 primary/secondary genes, generality/individuality, and the interconversion through
517 copy number variations. *Frontiers in Microbiology* **9**, (2018).

518 31. Brosch, R. *et al.* A new evolutionary scenario for the Mycobacterium tuberculosis
519 complex. *Proceedings of the National Academy of Sciences of the United States of*
520 *America* **99**, 3684–9 (2002).

521 32. Tsolaki, A. G. *et al.* Functional and evolutionary genomics of Mycobacterium
522 tuberculosis: Insights from genomic deletions in 100 strains. *Proceedings of the*
523 *National Academy of Sciences of the United States of America* **101**, 4865–4870 (2004).

524 33. Brosch, R. *et al.* Genomic analysis reveals variation between Mycobacterium
525 tuberculosis H37Rv and the attenuated M. tuberculosis H37Ra strain. *Infection and*
526 *Immunity* **67**, 5768–5774 (1999).

527 34. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W. & Small, P. M. Stable
528 association between strains of Mycobacterium tuberculosis and their human host
529 populations. *Proceedings of the National Academy of Sciences of the United States of*
530 *America* **101**, 4871–6 (2004).

531 35. Chiner-Oms *et al.* Genomic determinants of speciation and spread of the
532 Mycobacterium tuberculosis complex. *Science Advances* **5**, (2019).

533 36. Deshayes, C. *et al.* Detecting the molecular scars of evolution in the Mycobacterium
534 tuberculosis complex by analyzing interrupted coding sequences. *BMC evolutionary*
535 *biology* **8**, 78 (2008).

536 37. Smith, N. H., Hewinson, R. G., Kremer, K., Brosch, R. & Gordon, S. V. Myths and
537 misconceptions: the origin and evolution of Mycobacterium tuberculosis. *Nature*

538 *Reviews Microbiology* **7**, 537–544 (2009).

539 38. Etienne, G. *et al.* Identification of the polyketide synthase involved in the biosynthesis
540 of the surface-exposed lipooligosaccharides in mycobacteria. *Journal of Bacteriology*
541 **191**, 2613–2621 (2009).

542 39. Boritsch, E. C. & Brosch, R. Evolution of Mycobacterium tuberculosis: New Insights
543 into Pathogenicity and Drug Resistance. in *Tuberculosis and the Tubercle Bacillus,*
544 *Second Edition* **4**, 495–515 (American Society of Microbiology, 2016).

545 40. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in
546 whole bacterial genomes. *PLoS computational biology* **11**, e1004041 (2015).

547 41. Hershberg, R. *et al.* High Functional Diversity in Mycobacterium tuberculosis Driven by
548 Genetic Drift and Human Demography. *PLoS Biology* **6**, e311 (2008).

549 42. Supply, P. *et al.* Linkage disequilibrium between minisatellite loci supports clonal
550 evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area.
551 *Molecular Microbiology* **47**, 529–538 (2003).

552 43. Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A. & Behr, M. A. Genomic Deletions
553 Suggest a Phylogeny for the *Mycobacterium tuberculosis* Complex. *The Journal of*
554 *Infectious Diseases* **186**, 74–80 (2002).

555 44. Menardo, F., Duchêne, S., Brites, D. & Gagneux, S. The molecular clock of
556 Mycobacterium tuberculosis. *PLoS pathogens* **15**, e1008067 (2019).

557 45. Niobe-Eyangoh, S. N. *et al.* Genetic biodiversity of Mycobacterium tuberculosis
558 complex strains from patients with pulmonary tuberculosis in Cameroon. *Journal of*
559 *Clinical Microbiology* **41**, 2547–2553 (2003).

560 46. Godreuil, S. *et al.* First molecular epidemiology study of Mycobacterium tuberculosis
561 in Burkina Faso. *Journal of Clinical Microbiology* **45**, 921–927 (2007).

562 47. Groenheit, R. *et al.* The Guinea-Bissau Family of Mycobacterium tuberculosis Complex
563 Revisited. *PLoS ONE* **6**, e18601 (2011).
564 48. Comas, I. *et al.* Population Genomics of Mycobacterium tuberculosis in Ethiopia
565 Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa.
566 *Current Biology* **25**, 3260–3266 (2015).
567 49. Ates, L. S. *et al.* Mutations in ppe38 block PE_PGRS secretion and increase virulence of
568 Mycobacterium tuberculosis. *Nature Microbiology* **3**, 181–188 (2018).
569 50. Boritsch, E. C. *et al.* A glimpse into the past and predictions for the future: the
570 molecular evolution of the tuberculosis agent. *Molecular Microbiology* **93**, 835–852
571 (2014).
572 51. Martens, J. H., Barg, H., Warren, M. & Jahn, D. Microbial production of vitamin B12.
573 *Applied Microbiology and Biotechnology* **58**, 275–285 (2002).
574 52. Gopinath, K. *et al.* A vitamin B_{12} transporter in *Mycobacterium tuberculosis*. *Open*
575 *Biology* **3**, 120175 (2013).
576 53. Minias, A., Minias, P., Czubat, B. & Dziadek, J. Purifying Selective Pressure Suggests
577 the Functionality of a Vitamin B12 Biosynthesis Pathway in a Global Population of
578 Mycobacterium tuberculosis. *Genome Biology and Evolution* **10**, 2326–2337 (2018).
579 54. Bottai, D. *et al.* TbD1 deletion as a driver of the evolutionary success of modern
580 epidemic Mycobacterium tuberculosis lineages. *Nature Communications* **11**, 1–14
581 (2020).
582 55. LEÃO, S. C. *et al.* Practical handbook for the phenotypic and genotypic identification of
583 mycobacteria. (2004).
584 56. Borrell, S. *et al.* Reference set of Mycobacterium tuberculosis clinical strains: A tool for
585 research and product development. *PLOS ONE* **14**, e0214088 (2019).

586 57. Kent, P. & Kubica, G. *Public Health Mycobacteriology: A Guide for the Level III*
587 *Laboratory US Department of Health and Human Services, Centres for Disease Control*.
588 (1985).
589 58. Makhado, N. A. *et al.* Outbreak of multidrug-resistant tuberculosis in South Africa
590 undetected by WHO-endorsed commercial tests: an observational study. *The Lancet*
591 *Infectious Diseases* **18**, 1350–1359 (2018).
592 59. van der Zanden, A. G. M. *et al.* Improvement of differentiation and interpretability of
593 spoligotyping for Mycobacterium tuberculosis complex isolates by introduction of
594 new spacer oligonucleotides. *Journal of clinical microbiology* **40**, 4628–39 (2002).
595 60. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with
596 minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).
597 61. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
598 sequence data. *Bioinformatics (Oxford, England)* **30**, 2114–20 (2014).
599 62. Comas, I. *et al.* Human T cell epitopes of Mycobacterium tuberculosis are
600 evolutionarily hyperconserved. *Nature genetics* **42**, 498–503 (2010).
601 63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
602 transform. *Bioinformatics (Oxford, England)* **25**, 1754–60 (2009).
603 64. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
604 analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–303
605 (2010).
606 65. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping
607 and population genetical parameter estimation from sequencing data. *Bioinformatics*
608 *(Oxford, England)* **27**, 2987–93 (2011).
609 66. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration

610 discovery in cancer by exome sequencing. *Genome research* **22**, 568–76 (2012).
611 67. Steiner, A., Stucki, D., Coscolla, M., Borrell, S. & Gagneux, S. KvarQ: targeted and
612 direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* **15**, 881
613 (2014).
614 68. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
615 large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
616 69. Duchene, S. *et al.* Inferring demographic parameters in bacterial genomic data using
617 Bayesian and hybrid phylogenetic methods. *BMC evolutionary biology* **18**, 95 (2018).
618 70. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer
619 weighting and repeat separation. *Genome research* **27**, 722–736 (2017).
620 71. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long
621 sequencing reads. *Genome Biology* **16**, 294 (2015).
622 72. Scofield, D. G. GitHub - douglasgscofield/PacBio-utilities: Collection of utilities for
623 working with PacBio-based assemblies. Available at:
624 https://github.com/douglasgscofield/PacBio-utilities. (Accessed: 20th December
625 2019)
626 73. Seemann, T. GitHub - tseemann-snippy Rapid bacterial SNP calling and core genome
627 alignments. Available at: https://github.com/tseemann/snippy. (Accessed: 20th
628 December 2019)
629 74. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences
630 stored in a relational database. *Bioinformatics (Oxford, England)* **24**, 2672–6 (2008).
631 75. Li, H. *et al.* East-Asian Helicobacter pylori strains synthesize heptan-deficient
632 lipopolysaccharide. *PLoS Genetics* **15**, (2019).
633 76. Médigue, C. *et al.* MicroScope-an integrated resource for community expertise of

634 gene functions and comparative analysis of microbial genomic and metabolic data.
635 *Briefings in bioinformatics* **20**, 1071–1084 (2019).
636 77. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
637 Assessing the quality of microbial genomes recovered from isolates, single cells, and
638 metagenomes. *Genome Research* **25**, 1043–1055 (2015).
639 78. Page, A. J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis.
640 *Bioinformatics* **31**, 3691–3693 (2015).
641 79. Fedrizzi, T. *et al.* Genomic characterization of Nontuberculous Mycobacteria. *Scientific*
642 *Reports* **7**, (2017).
643 80. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast,
644 scalable and user-friendly tool for maximum likelihood phylogenetic inference.
645 *Bioinformatics (Oxford, England)* **35**, 4453–4455 (2019).

646 **Acknowledgements**

647 Calculations were partially performed at sciCORE (http://scicore.unibas.ch/) scientific computing core
648 facility at University of Basel. This work was supported by EDCTP2 grant DRIA2014-326—
649 DIAMA of the European Union, the Belgian General Directorate for Development
650 Cooperation (PhD fellowship to JCSN), grant ANR-16-CE35-0009 from Agence Nationale de la
651 Recherche, the Swiss National Science Foundation (grants 310030_188888, IZRJZ3_164171,
652 IZLSZ3_170834 and CRSII5_177163), and the European Research Council (309540-EVODRTB).
653 The views and opinions of authors expressed herein do not necessarily state or reflect those
654 of EDCTP. P.S., C.G. and S.D. declare the following competing interests: P.S. was a consultant
655 of Genoscreen; C.G. and S.D. were employees of the same company. The other authors
656 declare no competing interests. The funders had no role in study design, data collection and
657 analysis, decision to publish or preparation of the manuscript.

658 **Author contributions**

659 S.G., P.S., C.M., B.C.d.J., L.R. and J.C.S.N. designed the study. P.S., J.C.S.N., C.L., M.M., C.M.
660 and S.G. analyzed data and wrote the manuscript, with comments from all authors. A.J. and
661 C.M. performed the assembly of sequences. M.M. annotated the L8 genome. C.L., F.M., D.B.
662 and A.J. performed SNP analyses and phylogenetic reconstruction. M.M., A.R. and P.S.
663 conducted comparative analyses of complete mycobacterial genomes, with support from
664 C.L. and O.T. J.C.S.N., E.B.N., I.M.H., J.B.M., W.M., K.F. and M.D. performed and/or analyzed
665 data from mycobacterial isolation, growth assays, phenotypic characterization and/or
666 molecular tests. S.D., C.G., J.C.S.N., E.B.N., E.A. and M.K.K. conducted targeted deep
667 sequencing analyses. S.D., C.G. and W.S. L. Majlessi, F.S., C. Locht and C. Leclerc conducted
668 and/or analyzed immune assays. J.T., A. Criscuolo and S.B. conducted MLST, recombination
669 and/or phylogenetic analyses. L.F. conducted histopathological analyses. V.K., M.O. and C.P.

670 created bioinformatics tools and analyzed data. M.F. isolated STB strains. T.S. and T.P.S.

671 conducted core genome and NeighborNet analyses.

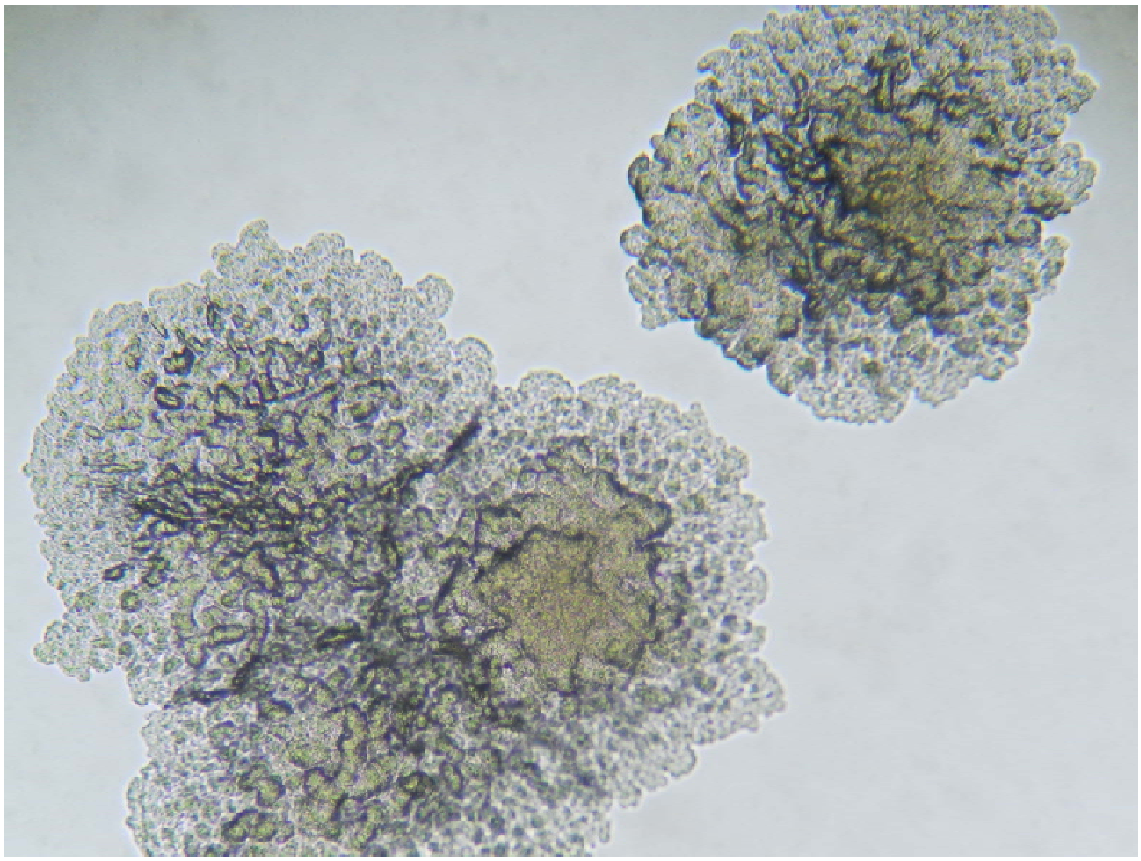
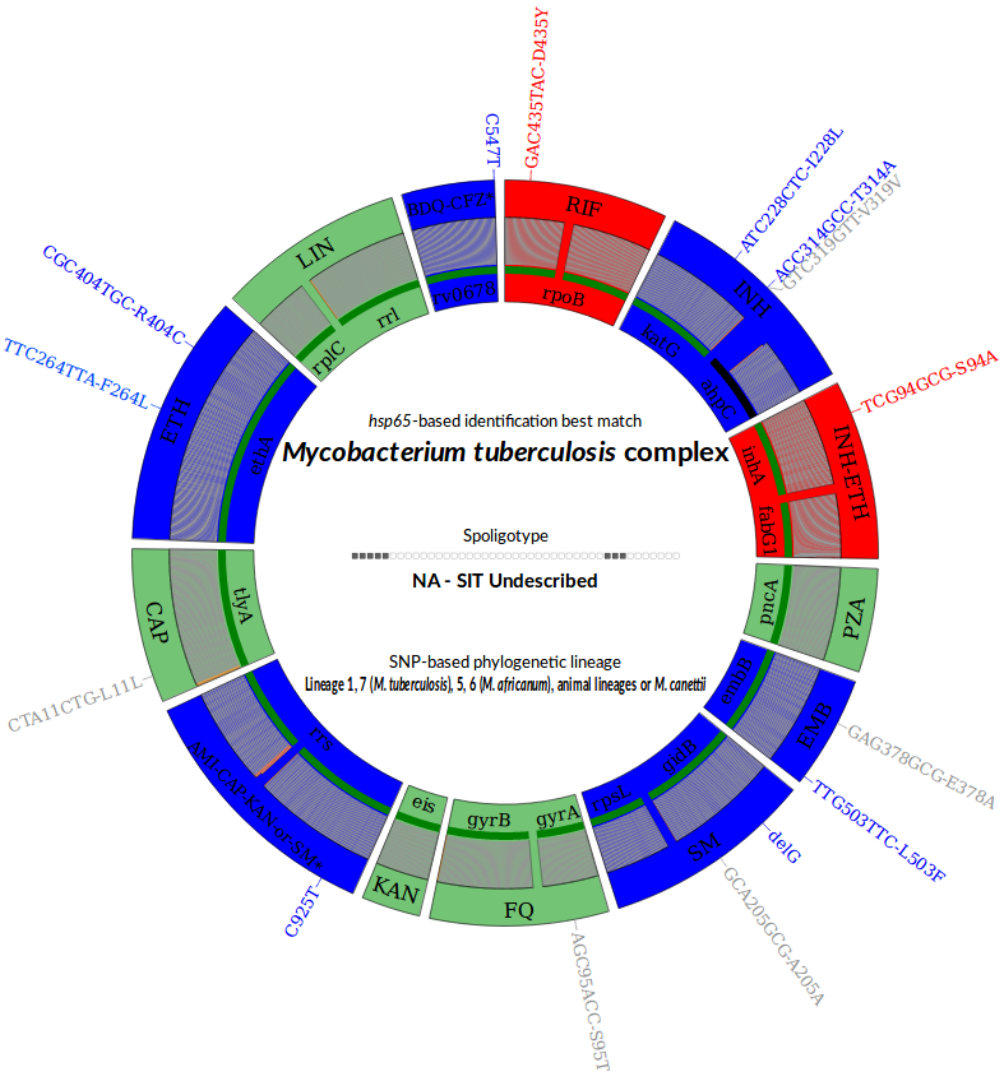


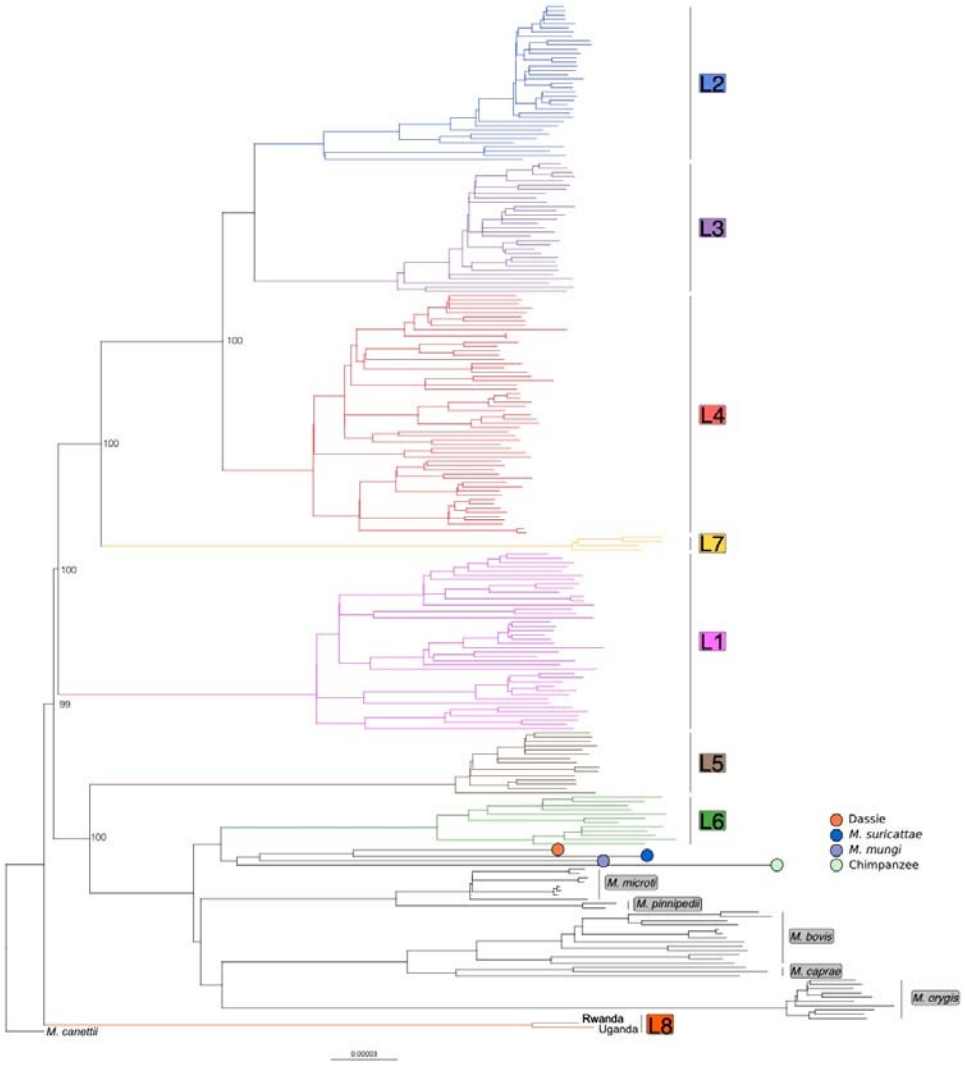
Figure 1. Microscopic image of L8 on Dubos agar medium showing typical rough colonies (read at 100x).



675

676 **Figure 2.** Deeplex-MycTB results identifying a multidrug resistant strain from Rwanda with
677 an atypical genotypic profile in the *M. tuberculosis* complex. Target gene regions are
678 grouped within sectors in a circular map according to the tuberculous drug resistance with
679 which they are associated. The two sectors in red indicate regions where rifampicin and
680 isoniazid resistance associated mutations are detected. The multiple sectors in blue refer to
681 regions where as yet uncharacterized mutations are detected, while sectors in green
682 indicate regions where no mutation or only mutations not associated with resistance (shown
683 in gray around the map) were detected. Green lines above gene names represent the

684 reference sequences with coverage breadth above 95%. Limits of detection (LOD) of
685 potential heteroresistance (reflected by subpopulations of reads bearing a mutation),
686 depending on the coverage depths over target sequence positions, are represented by grey
687 (LOD 3%) and orange zones (variable LOD >3%–80%; only seen in extremities of a few
688 targets, such as the two *rrs* regions) above the reference sequences within the sectors.
689 Information on an unrecognized spoligotype, an equivocal SNP-based and on mycobacterial
690 species identification, based on *hsp65* sequence best match, are shown in the centre of the
691 circle. *AMI, amikacin; BDQ, bedaquiline; CAP, capreomycin; CFZ, clofazimine; EMB,
692 ethambutol; ETH, ethionamide; FQ, fluoroquinolones; KAN, kanamycin; LIN, linezolid; INH,
693 isoniazid; PZA, pyrazinamide; RIF, rifampin; SM, streptomycin; SIT, spoligotype international
694 type.



695

696 **Figure 3** . Maximum likelihood phylogeny of 241 MTBC genomes, inferred from 43,442

697 variable positions. The scale bar indicates the number of substitutions per polymorphic site.

698 Branches corresponding to human-adapted strains are coloured and branches corresponding

699 to animal-adapted strains are depicted in black. The phylogeny is rooted on *M. canettii* and

700 bootstrap values are shown for the most important splits.

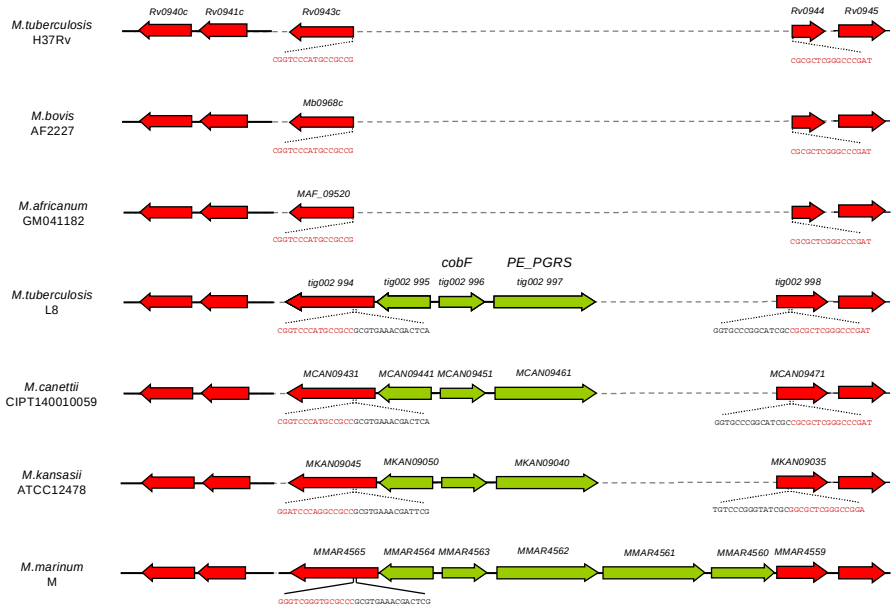
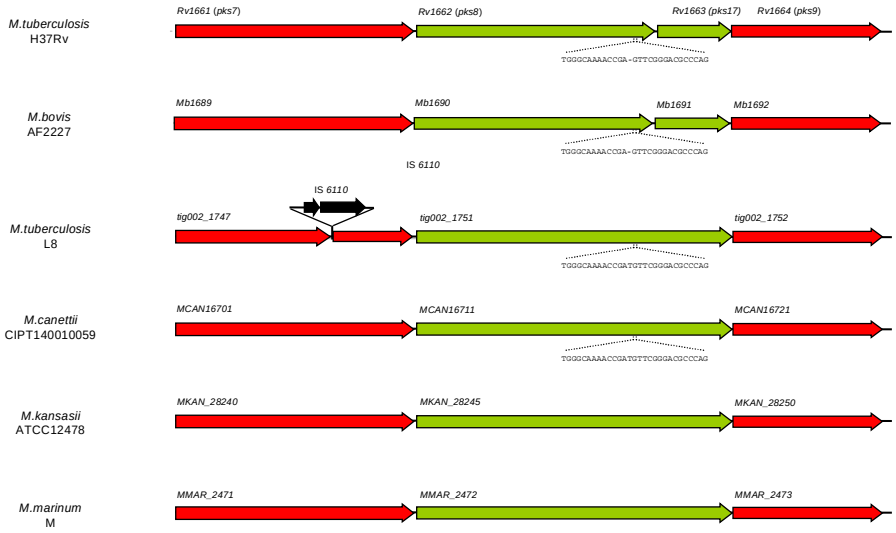


Figure 4. Aligned genome segments showing the *cobF* gene region in *M. tuberculosis* L8, M. canettii CIPT140010059 (alias STB-A), M. kansasii ATCC12478 and M. marinum M strains, and the corresponding deletion in *M. tuberculosis* H37Rv, M. bovis AF2122/97, and *M. africanum* GM041182. Coding sequences of this region are shown in green, and flanking coding sequences in red. Sequences flanking the deletion point in truncated genes in *M. tuberculosis*, *M. africanum* and *M. bovis*, and in the cobF region present in *M. canettii*, *M. kansasii* and *M. marinum* are indicated in red and black, respectively. Dashed lines correspond to missing segment parts relatively to the longest segment found in *M. marinum*.



710

711 **Figure 5.** Aligned genome segments showing the interrupted coding sequences *pks8/17* in

712 *M. tuberculosis* H37Rv and *M. bovis* AF2122/97 gene region, and complete *pks8* genes in L8,

713 *M. canettii* CIPT140010059 (alias STB-A) and *M. kansasii* ATCC12478. *pks8/17* and *pks8*

714 coding sequences are shown in green, and flanking genes in red. Sequences flanking the 1-

715 nucleotide deletion and resulting in a frameshift in *M. tuberculosis* complex strains are

716 indicated. Dashed lines correspond to missing segment parts relatively to the longest

717 segment found in *M. canettii.*

Table 1. Standard biochemical characteristics of selected mycobacterial species or *M. tuberculosis* complex lineages/subspecies versus L8

Mycobacterial species/lineage	BCCM/ITM REF N°	Niacin production	Nitrate reduction	Urease hydrolysis	Tween hydrolysis	Catalase production	Arylsulfatase
L8	500961	+	+	+	weak +	-	-
L1	500941	+	+	+	-	weak +	-
L2	500945	+	-	+	-	-	-
L3	500947	+	+	+	-	-	-
L4	500951	+	+	+	+	-	-
L5	500953	+	-	-	+	-	-
L6	500957	+	-	weak +	weak +	-	-
L7	500959	+	-	weak +	-	-	-
M. bovis	500324	-	-	weak +	-	-	-
M. bovis BCG	500101	-	-	+	-	weak +	-
M. orygis	501026	+	-	+	-	-	-
M. canettii	500199	-	-	+	+	weak +	-
M. fortuitum	500465	-	+	+	-	weak +	+

+: positive reaction; -: negative