Pks5-recombination-mediated surface remodelling in Mycobacterium tuberculosis emergence

Eva C. Boritsch, Wafa Frigui, Alessandro Cascioferro, Wladimir Malaga, Gilles Etienne, Françoise Laval, Alexandre Pawlik, Fabien Le Chevalier, Mickael Orgeur, Laurence Ma, et al.

1 *pks5*-recombination-mediated surface remodelling in *Mycobacterium tuberculosis*

2 emergence

3

4 Eva C. Boritsch[1], Wafa Frigui[1], Alessandro Cascioferro[1], Wladimir Malaga[2,3], Gilles

5 Etienne[2,3], Françoise Laval[2,3], Alexandre Pawlik[1], Fabien Le Chevalier [1,5], Mickael Orgeur[1],

6 Laurence Ma[4], Christiane Bouchier[4], Timothy P. Stinear[6], Philip Supply[7], Laleh Majlessi[1],

7 Mamadou Daffé[2,3], Christophe Guilhot[2,3,*] and Roland Brosch[1,*].

8

9 [1]Institut Pasteur, Unit for Integrated Mycobacterial Pathogenomics, 75015 Paris, France;

10 [2]CNRS, IPBS (Institut de Pharmacologie et de Biologie Structurale), 31000 Toulouse, France;

11 [3]Université de Toulouse, UPS, IPBS, 31000 Toulouse, France;

12 [4]Institut Pasteur, PF1-Plate-Forme Génomique, Paris, France;

13 [5]University Paris Diderot, Sorbonne Paris Cité, Cellule Pasteur, Paris, France

14 [6]Department of Microbiology and Immunology, University of Melbourne, Parkville,

15 Victoria, Australia.

16 [7]Inserm U1019, CNRS UMR8204, Université de Lille, Institut Pasteur de Lille, Center for

17 Infection and Immunity, 59000 Lille, France.

18

19 *Corresponding authors: Christophe.Guilhot@ipbs.fr and roland.brosch@pasteur.fr

20

Mycobacterium tuberculosis is a major, globally spread, aerosol-transmitted human pathogen, thought to have evolved by clonal expansion from a *Mycobacterium canettii*-like progenitor. In contrast, extant *M. canettii* strains are rare, genetically diverse and geographically restricted mycobacteria of only marginal epidemiological importance. Here we show that the contrasting evolutionary success of these two groups is linked to loss of lipooligosaccharide (LOS) biosynthesis and subsequent morphotype changes. Spontaneous smooth-to-rough *M. canettii* variants were found mutated in the polyketide-synthase-encoding *pks5* locus and deficient in LOS synthesis; a phenotype restored by complementation. Importantly, these rough variants showed altered host-pathogen interaction and increased virulence in cellular- and animal-infection models. In one variant, LOS deficiency occurred via homologous recombination between two *pks5* copies and removal of the intervening acyltransferase-encoding gene. The resulting single *pks5* configuration is similar to that fixed in *M. tuberculosis*, known to lack LOS. Our results suggest that *pks5*-recombination-mediated bacterial surface remodelling increased virulence, driving evolution from putative generalist mycobacteria towards professional pathogens of mammalian hosts.

36   Tuberculosis is a major human infectious disease. Although many aspects of the disease-
37   causing potential of its etiological agent *Mycobacterium tuberculosis* are known[1], our
38   understanding is scant of the molecular events that favoured its evolutionary success as one of
39   the most widely distributed human pathogens. New insights into this question are important
40   for uncovering mechanisms of pathogenesis and new drug targets[2]. Strains of the closely
41   related and phylogenetically early branching *Mycobacterium canettii*, also named smooth
42   tubercle bacilli (STB)[3] are powerful resources to investigate the evolution of *M. tuberculosis*
43   and the *M. tuberculosis* complex (MTBC)[3]. The first strain of *M. canettii* was isolated by
44   Georges Canetti in 1969 and since then less than 100 isolates have been described, most of
45   which have been isolated from tuberculosis patients with a connection to the Horn of Africa[4-
46   7]. Despite their geographic restriction, *M. canettii* strains show much greater genetic
47   variability and are less virulent/persistent than *M. tuberculosis*. Genome comparisons suggest
48   that *M. tuberculosis* evolved by clonal expansion from a pool of *M. canettii*-like tubercle
49   bacilli through the gain of virulence and persistence mechanisms[3,8,9]. While some genomic
50   differences were found to be specific for a single *M. canettii* strain, apparently due to isolated
51   horizontal gene transfer (e.g. the *mce5* operon[3] or the *eptABCD* operon[10], present exclusively
52   in strain STB-J)[3], others are conserved throughout all *M. canettii* strains as a result of
53   phylogenetic ancestry (e.g. *cobF,* present in *M. canettii* and deleted from the MTBC)[3].

54   Here we investigated phenotypic differences between *M. canettii* and *M. tuberculosis*
55   and focused on the unique, conserved, smooth (S) colony morphotype of *M. canettii*, as
56   opposed to the rough (R) morphotype of MTBC members. Previously, S-morphotypes of non-
57   tuberculous mycobacterial species such as *Mycobacterium avium*[11], *Mycobacterium*
58   *abscessus*[12], *Mycobacterium kansasii*[13] or *Mycobacterium marinum*[14] have been found to be
59   less virulent than R-morphotypes, raising the question of whether the highly conserved *M.*
60   *tuberculosis* R-morphotype might have been the result of evolutionary selection, based on
61   host-pathogen interactions favouring a more virulent or persistent phenotype. In other
62   mycobacterial species, S/R variation is often attributed to different kinds of cell surface
63   glycolipids, such as glycopeptidolipids (GPL) for *M. avium*[15] and *M. abscessus*[16,17] or
64   lipooligosaccharides (LOS) for *M. kansasii*[13] and *M. marinum*[14,18,19]. Insights from earlier
65   studies for *M. canettii* have remained abstruse as no specific lipid exclusively present in the S-
66   morphotype has been identified[20], nor has the genetic basis for morphotype variation been
67   determined[4].

68   Building on recent data from several *M. canettii* genomes[3], here we studied smooth
69   and spontaneously converted R-variants of two different *M. canettii* strains, STB-K (CIPT

70   140070010) and STB-I (CIPT 140070007)[3], hereafter named $K_{S/R}$ and $I_{S/R}$, respectively. We

71   used whole genome sequencing (WGS) and identified differences in genes of the *pks5* locus,

72   which in *Mycobacterium smegmatis*, *M. marinum* or *M. kansasii* are implicated in LOS

73   biosynthesis[14,21-23]. In this report we uncover the mechanisms underlying the S-to-R

74   morphology change of tubercle bacilli and explore the biological consequences with emphasis

75   on the patho-evolution of *M. tuberculosis.*

76

77   **RESULTS**

78   **Genome comparison of *M. canettii* S- and R-variants**

79   During passaging of different smooth *M. canettii* strains, we observed spontaneous R-

80   morphotype variants for two *M. canettii* strains (I and K). These R-variants were re-passaged

81   three times on solid medium to ensure stable phenotypes and subjected to Illumina-based

82   WGS. The results obtained pointed to morphotype-linked alterations in the *pks5* locus of *M.*

83   *canettii* S/R variants (Supplementary Tables 1 and 2, Supplementary Note), which was of

84   particular interest given suggested roles for Pks5 polyketide synthases in morphotype changes

85   and LOS production in different mycobacterial species[14,21]. In the LOS-producers *M.*

86   *marinum*[24] and *M. kansasii*[25], the Pks5-encoding locus harbours two proximal *pks5* homologs

87   separated by one or more interspaced *pks5*-associated gene(s) (Fig. 1A). The genome of *M.*

88   *canettii* STB-A[3] also shows a twin-*pks5* configuration similar to *M. marinum*. In contrast *M.*

89   *tuberculosis*, which does not produce LOS only has a single *pks5* and lacks a full-length *pap*

90   ortholog (Fig. 1A).

91

92   **Recombination of two *pks5* genes in *M. canettii* K_R and *M. tuberculosis***

93   To characterize the *pks5* locus in strains $I_{S/R,}$ $K_{S/R}$ and *M. tuberculosis* H37Rv we performed

94   long-range PCR using *pks5*-flanking primers, yielding amplicons of ca. 15 kb for both the S-

95   and R-variants of strain I, and 6 kb for *M. tuberculosis* H37Rv, corresponding to the twin- and

96   single-*pks5* configuration, respectively (Fig. 1 A, B). For strain K, however, the prevalent

97   amplicon size of the S-variant was 15 kb, while the size for the R-variant was 6 kb, which was

98   surprising because the previously released STB-K genome sequence assembly[3] indicated only

99   a single *pks5*. The PCR result suggests that the previous WGS-based genome assembly of

100   strain K (NC_019951) contained a misassembly in that region probably due to high sequence

101   identity (> 94 %) between the two *pks5* genes. Similarly, we also confirmed the presence of

102   two *pks5* genes by long-range PCR amplicon-sequencing for *M. canettii* strains D and J, for

103   which previously only one *pks5* was indicated in the released genome sequences NC_019950

104 and NC_019952, respectively (Supplementary Fig. 1). Hence, the results obtained for S- and

105 R-variants of strain K suggest that K_S contains two *pks5* genes, flanking a gene (*pap*) coding

106 for a putative polyketide-synthase-associated acyltransferase[26], whereas K_R only harbours a

107 single *pks5* and no interspaced *pap* gene.

108 These findings were supported by results obtained by alignment to the reference

109 sequences of *M. canettii* STB-A and *M. tuberculosis* H37Rv (Supplementary Fig. 2 and

110 Supplementary Note). This analysis suggested that homologous recombination in strain K_R

111 occurred at a site that left intact the typical polyketide-synthase domain structures of Pks5[26,27]

112 (Fig. 1). The observed *pks5* configuration in K_R can thus serve as a model for the situation

113 found in *M. tuberculosis,* where two former *pks5* genes also seem to have coalesced into a

114 single *pks5* gene. In the latter case, the proposed recombination event also left the domain

115 organisation of Pks5 intact, although the site of recombination was apparently different from

116 the site observed for K_R (Fig. 1 and Supplementary Fig. 2B). Further comparisons among

117 different MTBC members[9] showed > 99.93% sequence identity of the single *pks5* gene in

118 these strains, suggesting a unique recombination event in the most recent common ancestor of

119 the MTBC after the separation from an *M. canettii*-like progenitor.

120 Finally, for strain I, amplicon-sequencing of the *pks5* locus of I_R and I_S confirmed the

121 presence of *pap* and two *pks5* copies in both morphotypes. However, alignment of amplicon-

122 derived sequences indicated the presence of multiple non-synonymous SNPs in both *pks5*

123 genes (6 in *pks5-2* and 3 in *pks5-1*) for the R-variant (Supplementary Fig. 3A), which likely

124 resulted from recombination of the ketosynthase domain regions of the *pks5* copies

125 (Supplementary Fig. 3B). From these data we predicted that these changes likely caused the

126 R-morphotype of I_R.

127

128 **Pap and Pks5-2 restore the S-morphotype in *M. canettii***

129 To test if changes in the *pks5* locus were responsible for the S/R variations in *M. canettii,* we

130 transformed strains K_R and I_R, as well as *M. tuberculosis* H37Rv with the integrating cosmid

131 C9, selected from an *M. canettii* STB-A large-fragment genomic library constructed for this

132 purpose in pYUB412. This vector integrates into the *attB* site located in the *glyV*-tRNA of

133 mycobacterial genomes[28]. The 30.5 kb insert of cosmid C9 spans the *pks5* locus of STB-A

134 (Supplementary Fig. 4). Fig. 2 shows that recombinant *M. canettii* strains I_R::C9 and K_R::C9

135 regained S-morphology that was indistinguishable from their respective S-variants. To refine

136 the region required for S-morphotype restitution, we isolated a second cosmid, named H6 that

137 contains full-length *pap* and *pks5-2* but a 3'-truncated *pks5-1* (Supplementary Fig. 4).

138 Transformation of I_R and K_R with H6 yielded smooth colonies for both I_R::H6 and K_R::H6

139 (Supplementary Fig. 5A).

140   Complementation with a *pks5-2* single gene expression construct restored an S-

141 morphotype for strain I_R::*pks5-2*, but not for K_R::*pks5-2* (Supplementary Figs. 5B and 5C).

142 Moreover, transformation with *pap* was unable to restore smooth colonies in either of the

143 strains despite appropriate protein expression (Supplementary Figs. 5B and 5D). Together, the

144 complementation experiments suggest that *pks5-2* and *pap* are necessary for the S-

145 morphotype in *M. canettii*.

146   In contrast, R-morphology of *M. tuberculosis* H37Rv remained unchanged upon

147 transformation with cosmid C9 (Fig. 2A) despite correct genomic integration of the cosmid

148 (Fig. 3). Similarly, transformation of other MTBC lineage members with C9 did not yield S-

149 morphotypes (Supplementary Fig. 6). These results suggest that apart from an appropriate

150 *pks5-pap* locus, additional genes intervene in the formation of S-morphotypes in tubercle

151 bacilli.

152

153 **_M. canettii_ strains synthesize LOS**

154   Analysis of total lipid extracts from a large panel of *M. canettii* strains by thin-layer

155 chromatography (TLC) showed glycoconjugate spots, with TLC mobilities similar to that of

156 LOS from *M. canettii* reference strain CIPT-140010059 (STB-A or *M. canettii* strain A)[29] in

157 all tested strains. Substantial differences in the quantity produced by each strain were

158 observed (Fig. 4A). In contrast, no such glycoconjugates were detected in *M. tuberculosis*

159 H37Rv, consistent with the long-standing knowledge that *M. tuberculosis* does not synthesize

160 LOS (Fig. 4A)[30]. Purification and MALDI-TOF mass spectrometry analysis of LOS from *M.*

161 *canettii* strains A, K_S and I_S revealed pseudomolecular-ion mass peaks at *m/z* 2530, 2572, and

162 2614 for LOS from strain K_S (Fig. 4B) and strain A[29], whereas the pattern of strain I_S showed

163 peaks at *m/z* 2516, 2558 and 2600 (Fig. 4B), consistent with the different mobility on TLC

164 (Figs. 4A and 4C). The 14 mass-unit difference indicates strain-specific structural differences,

165 likely due to the absence of one methyl-group from the carbohydrate or the lipid moiety.

166   The structure of the simplest carbohydrate domain of the *M. canettii* LOS is 2-O-Me-

167 α-L-Fuc*p*(1->3)-β-D-Glc*p*(1->3)-2-O-Me-α-L-Rha*p*(1->3)-2-O-Me-α-L-Rha*p*(1->3)-β-D-

168 Glc*p*(1->3)-4-O-Me-α-L-Rha*p*(1->3)-6-O-Me-α-D-Glc(1<->1)-tri-O-acyl-α-D-Glc (Fig. 4D).

169 This octaglycosyl unit is usually further glycosylated by an incompletely defined *N*-acyl

170 derivative of a 4-amino-4,6-dideoxy-Gal*p* residue to generate a second nonasaccharide-

171 containing glycolipid (LOS 9) (Fig. 4D)[29]. The MALDI-TOF mass spectra of LOS purified

6

172 from strains K and A are consistent with the LOS 9 structure. Of note, similar patterns of ion
173 mass peaks were detected in spectra of LOS purified from *M. canettii* strains D, E, F, G, L,
174 and H, while LOS of strain J showed a 14 mass difference, similar to strain I_S (Supplementary
175 Fig. 7).

Rough *M. canettii* variants I and K are deficient in LOS biosynthesis

177 Having established that all tested *M. canettii* strains synthesized LOS, we next compared the
178 production of these glycoconjugates in the R-variants of I and K. As predicted, I_R was
179 severely impaired in LOS production with only trace amounts detected, and LOS was
180 undetectable in K_R (Fig. 4C and Supplementary Fig. 8). No other differences in polyketide-
181 derived-lipids were observed between the S- and R-variants (Supplementary Fig. 9).
182 Complementation with cosmid C9 restored LOS biosynthesis in recombinant I and K strains
183 (Fig. 4C), whereas recombinant expression of *pks5-2* alone restored the LOS profile in
184 I_R::*pks5-2*, but not in K_R::*pks5-2* (Supplementary Fig. 10A), likely due to the absence of full-
185 length *pap* in K_R::*pks5-2*. Indeed, based on current knowledge on acyl-trehalose biosynthesis
186 in *M. tuberculosis*[31,32], a *pap*-encoded acyltransferase is needed to catalyze the transfer of
187 Pks-produced polymethyl-branched fatty acids onto trehalose.
188   We also tested LOS production in the MTBC strains complemented with cosmid C9,
189 despite unaffected R-morphotype. LOS was not detected in any of the recombinant *M.*
190 *tuberculosis* strains (Supplementary Fig. 10B), suggesting that apart from *pks5* recombination
191 and *pap* deletion, in the MTBC additional loss-of-function mutations or insertions/deletions
192 (indels) have occurred in adjacent genes of the LOS-encoding locus that are not covered by
193 cosmid C9 (Supplementary Tables 3 and 4). These results establish that point mutations or
194 recombination/deletion events in the core *pks5-pap* locus of *M. canettii* strains I_R and K_R led
195 to LOS deficiency and R-morphology. The results also explain the R-morphotype of *M.*
196 *tuberculosis* and open new perspectives for detailed research on the functions of *pks5*-
197 adjacent genes.
198

Increased fitness and virulence of R-variants in cellular and animal models

200 Given the identified altered cell surface structure associated with S/R morphology, the
201 question arose whether these changes had an impact on host-pathogen interactions. We thus
202 undertook a series of infection experiments, starting with tests of intracellular replication, a
203 hallmark of mycobacterial virulence. Differentiated human THP-1 macrophages were infected
204 with Sauton-grown *M. canettii* strains K_S, K_R and K_R::C9, as well as K_R::vector-control, up to

205 five days. While no intracellular growth was observed for K_S and K_R::C9, the R-morphotypes

206 K_R and $K_{R::}$vector-control replicated readily, analogous to *M. tuberculosis* H37Rv (Figs. 5A

207 and 5B). Similar results were obtained when human peripheral blood monocyte-derived

208 macrophages from healthy donors were infected with the different morphotypes

209 (Supplementary Fig. 11A), confirming that the increased fitness advantage of the R-

210 morphotypes was a general phenomenon in human macrophages and not just in THP-1 cells.

211 For murine-derived RAW macrophages a similar trend, but without significant difference was

212 observed (Supplementary Fig. 11B).

213   To test whether an increased fitness advantage could also be observed *in vivo*, we first

214 infected SCID mice intravenously with 1 x 10^6 CFU of each morphotype. Mice infected with

215 *M. tuberculosis* H37Rv or *M. canettii* K_R had a median survival of 16 and 24 days pi,

216 respectively, while mice infected with strains K_S and K_R::C9 had a longer survival time; on

217 average 31 and 36 days, respectively (Fig. 5C). Higher virulence of R-morphotypes was also

218 observed in the sensitive guinea pig model, using low-dose aerosol infection (Fig. 5D). While

219 in this highly susceptible model, the S-variants of *M. canettii* were already substantially

220 virulent and able to replicate in infected animals, the ratios of the bacterial burden in the lungs

221 at day 42 relative to day 1 were significantly higher for strain K_R than those for the S-

222 morphotypes. Interestingly, for animals infected with K_R::C9 (Fig. 2) we observed more than

223 half of the colonies showing a re-converted R-morphotype, which was never observed in any

224 of our previous *in vitro* experiments, suggesting possible reversion of the usually very stable

225 pYUB412-based integrated genetic complementation[33], due to *in vivo* fitness advantages of R-

226 morphotypes during the long-duration *in vivo* infection assays.

227   To further investigate the host-pathogen interaction of S- and R-variants, we

228 characterized the inflammatory responses induced in infected phagocytes, as recently

229 performed for S- and R-forms of *M. abscessus*[34,35]. As shown in Figs. 5E and 5F, a potential

230 LOS-dependent anti-inflammatory effect was observed in *M. canettii*-infected bone-marrow

231 derived dendritic cells (BM-DCs) of C57BL/6 mice. Strains K_S and K_R::C9 induced

232 significantly lower secretion of NF-κB-dependent inflammatory cytokines IL-6 and IL-12p40

233 than K_R. Results with these two cytokines, known to be induced in response to infection with

234 *M. tuberculosis*[36], were also representative for TNF-α and NF-κB-dependent CCL5 and

235 CXCL10 chemokines (Supplementary Fig. 12). In our experimental conditions, most of the

236 responses were dependent on interaction via Toll-like receptor 2 (TLR-2). Smooth variants

237 showed reduced interaction with this crucial innate immune sensor of numerous

238 mycobacterial ligands (Figs. 5 G, 5H and Supplementary Fig. 13). Presence or absence of

239   LOS thus seems to strongly influence the complex pattern of host-tubercle bacilli interaction,

240   with likely consequences for virulence and pathogenesis.

241

242   **DISCUSSION**

243   The evolutionary transition for a bacterial population from low-virulence/high genetic

244   diversity towards high-virulence/low genetic diversity is well established for some major

245   human pathogens such as *Yersinia pseudotuberculosis* vs *Yersinia pestis* or *Salmonella*

246   *enteritidis* vs *Salmonella typhi*[37]. In contrast, the specific genetic changes that led to the

247   emergence of *M. tuberculosis* are less well understood. Here we have exploited the ancestral-

248   state characteristics of *M. canettii* to address this gap in our knowledge[3,6,9]. We show that the

249   transition of *M. tuberculosis* from likely generalist to obligate pathogen was associated with a

250   dramatic change in mycobacterial surface glycolipid composition. The S-to-R morphotype

251   variation we have studied here has been linked with higher virulence in opportunistic

252   pathogenic mycobacteria[11-14] and variation in LOS content has been shown to account for

253   morphology differences in *M. kansasii*[13] or *M. marinum*[14,18,19]. However, earlier studies on *M.*

254   *canettii* colony morphology have provided inconclusive results. While a study from van

255   Soolingen and coworkers found differences in LOS content between S/R colony variants of

256   an *M. canettii* strain isolated from a 2-year-old Somali child[4], Lemassu and colleagues did not

257   observe a correlation between LOS content and colony morphology in various strains tested[20],

258   likely because the used strains were poor LOS producers with intermediate morphotypes. In

259   this current work we resolve these ambiguities and clearly link LOS biosynthesis with the

260   smooth colony morphotype. We also provide a genetic mechanism for the phenotype that

261   helps explain the evolution of the tubercle bacilli. Our results indicate that recombination of

262   two *pks5* genes associated with deletion of the interspersed *pap* gene evidently was a

263   molecular event that became fixed in the most recent common ancestor of the MTBC after the

264   phylogenetic separation from the *M. canettii* clade (Fig. 6). Indeed, while *M. canettii* strains

265   possess a twin *pks5* and *pap* conformation and synthesize LOS, all members of the classical

266   MTBC contain an abridged *pks5* locus missing a full length *pap* gene and do not produce

267   LOS. Identification of a similar recombination event in a spontaneous R-variant of *M. canettii*

268   K reinforces the likelihood of this evolutionary scenario. Comparable recombination-deletion

269   events also occurred for other genes, e.g. *pknH* (Supplementary Note).

270   Pks5 is a type-I polyketide synthase and a member of the broader polyketide-synthase

271   family, which in mycobacteria have important functions that range from catalyzing the

272   essential last condensation step of mycolic acid biosynthesis (Pks13)[38] to the synthesis of the

273   polyketide backbone of phenolic glycolipids (Pks15/1)[39] or the synthesis of mycoketides

274   (Pks12)[40]. In *M. marinum*, the two Pks5 polyketide synthases (Pks5 and Pks5.1) are thought

275   to be involved in the synthesis of polymethyl-branched fatty acids, which are further modified

276   by glycosyltransferases and methyltransferases encoded by genes adjacent to the *pks5*

277   genes[22,23,41] to produce LOS. Indeed*,* disruption of *pks5* (*MMAR_2340*), analogous to

278   disruptions of the adjacent genes *fadD25* (*MMAR_2341*), *papA4 (MMAR_2343),* and *papA3*

279   *(MMAR_2355*) leads to a complete loss of LOS, suggesting that besides the core *pks5-pap*

280   locus, several flanking genes are also involved in the early steps of LOS-

281   biosynthesis[14,18,19,22,23,41,42] (Supplementary Table 3).

282   Within the group of tubercle bacilli, the overall genetic organization of the

283   orthologous LOS locus is similar between *M. canettii* and the non-functional version in *M.*

284   *tuberculosis*. However, apart from the recombined *pks5* and the deleted *pap* gene, several

285   flanking genes show SNPs and small indels (Supplementary Table 4), some of which might

286   represent loss-of-function mutations involved in the non-complementation phenotype with

287   cosmid C9 in MTBC members. These findings suggest that after the phylogenetic separation

288   of the MTBC from the *M. canettii*-like progenitor pool, in addition to *pks5* recombination and

289   *pap* deletion, other mutations in genes of the LOS locus accumulated in the MTBC in the

290   absence of selective pressure for sustaining LOS biosynthesis. While LOS production might

291   play an important role in generalist mycobacteria, it appears that its absence might provide a

292   selective advantage for specialized pathogens in the mammalian host.

293   This conclusion is supported by our different infection experiments, where loss of

294   LOS by *M. canettii* R-variants led to increased virulence in human macrophages and under *in*

295   *vivo* conditions, particularly for the sensitive guinea pig infection model. These data support a

296   scenario wherein the recombination of *pks5* genes in a common ancestor of the MTBC has

297   contributed to the evolutionary success and host adaptation of the resulting LOS-deprived

298   strains. This finding is in agreement with data for other mycobacterial species, where

299   increased virulence of R-morphotypes relative to S-morphotypes has been reported[11,12,14],

300   although the involved glycolipids are different. In *M. avium,* for example, homologous

301   recombination within the *ser2* gene cluster likely led to phenotype changes linked to loss of

302   GPLs [43]. For *M. abscessus,* a masking effect of GPLs on the outermost layer of the cell wall

303   of the S-morphotype strains has been suggested, which was related to repression of TLR2

304   responses[34]. Similarly, LOS is located in the outermost layer of the mycobacterial cell

305   envelope[44] and thus might be interacting with other cell wall-associated

306   lipids/glycoconjugates such as phthiocerol dimycocerosates, lipoarabinomannans or the 19-

307   kD lipoprotein[45], which are important for host-pathogen interaction[46]. At present, it is difficult

308   to speculate by which mechanisms R-variants of tubercle bacilli might have gained a fitness

309   advantage during infection. However, differences we show here in TLR2-mediated

310   inflammatory responses between S- and R-variants of *M. canettii* K suggest that bacterial

311   surface remodelling could well have played an important role in the ancestor of the MTBC

312   during adaptation to mammalian hosts. In this respect, the proposed recombination of *pks5*

313   and deletion of *pap* in a common ancestor of the MTBC adds to previously described

314   differences between *M. canettii* and the MTBC, such as the putative loss of vitamin B12

315   production due to the *cobF* deletion[3,7,9,47,48], the integration of gene *pe_pgrs33 (rv1818c)*[3], the

316   molecular scars in *pks8/17* and three other gene pairs[3,7,9], or the exchange of the exotoxin-

317   encoding domain in CpnT (Rv3903)[9,49] (Fig. 6). While these reported differences are mainly

318   based on hypotheses formulated during comparative genomics and genome analysis

319   approaches, the virulence differences observed for rough *M. canettii* relative to isogenic LOS-

320   producing smooth *M. canettii* strains now provide first experimental support of our postulated

321   evolutionary model (Fig. 6). Recombination of two ancestral *pks5* genes and consequent

322   remodelling of the bacterial cell surface thus represent key events in the emergence of the

323   professional pandemic pathogen *M. tuberculosis*.

324   **Authors to whom correspondence and requests for materials should be addressed:**

325   Roland Brosch (roland.brosch@pasteur.fr) or Christophe Guilhot

326   (Christophe.Guilhot@ipbs.fr)

343   **Author contributions**

344   E.C.B., C.G., L.M. and R.B. designed the study. E.C.B., W.F., F.LC., and A.P. performed

345   mycobacterial phenotypic assays and/or infection experiments. E.C.B., A.C. and R.B.

346   established genetic constructs. W.M., G.E., F.L., M.D. and C. G. generated and/or analysed

347   mycobacterial lipid and lipooligosaccharide profiles. E.C.B., L.Ma, C.B., M.O., T.P.S. and

348   P.S. generated and/or analysed sequence data. E.C.B and L.Majlessi conducted and analysed

349   immune assays. E.C.B., T.P.S., P.S., C.G. and R.B wrote the manuscript, with comments

350   from all authors.

351

352   **Competing financial interests**

355   **Accession codes**

360   **Figure legends**

361
362   **Fig. 1.** Comparison of *pks5* genomic region of various mycobacterial strains. (A) *pks5* genetic

363   region of *M. kansasii* ATCC 12478, *M. marinum* M, *M. canettii* A and *M. tuberculosis*

364   H37Rv. Percentages on gray lines represent identity values as defined by ClustalW2 between

365   genes of *M. kansasii*, and *M. marinum* or *M. canettii* A and *M. marinum* or *M. tuberculosis*

366   H37Rv. The red arrows indicate location of primers used for long range PCR of *pks5* gene(s)

367   in *M. canettii* strains I and K, as well as in *M. tuberculosis* H37Rv. (B) Results of long range

368   PCR using above-mentioned primers that bind outside the *pks5* genes. *M. tuberculosis* H37Rv

369   (*Mtb*)*; M. canettii* strains: I_S , I_R, K_S and K_R. Note that for strains I_S and I_R unspecific

370   fragments at ~ 12 kb are visible. Data are representative of at least three repetitions. (C)

371   Genetic locus of *pks5* genes in *M. canettii* strains K_S and K_R and potential recombination

372   region in K_R (depicted by black arrows). Percentages on gray lines represent identity values as

373   defined by ClustalW2. (D) Domain organization of the *pks5* genes of *M. canettii* strains K_S

374 and K_R. Domains were predicted according to the organization *of pks5* of *M. tuberculosis*

375 H37Rv[50] and their respective amino acid positions (AA) are depicted either above or below

376 the particular domains. Sequences of individual domains of *pks5-1* and *pks5-2* were aligned

377 using ClustalW2 and identity values are shown as percentage. Origin of the domains of the

378 recombined *pks5* of *M. canettii* strain K_R is represented in different nuances of gray (light

379 gray originating from *pks5-2* and dark gray from *pks5-1;* medium gray indicates that domains

380 are identical between *pks5-1* and *pks5-2*). Domain abbreviations: ketosynthase (KS),

381 acyltransferase (AT), dehydratase (DH), enoylreductase (ER), ketoreductase (KR) and acyl-

382 carrier protein (ACP).

383

384 **Figure 2.** Complementation of morphology phenotypes of R variants. (A) Colony

385 morphologies of *M. canettii* strains K_S and I_S and their rough mutants K_R and I_R

386 complemented with the whole *pks5* locus (C9) as well as *M. tuberculosis* H37Rv WT and *M.*

387 *tuberculosi*s::C9. Data are representative of at least ten repetitions (plating). Scale bar = 2.5

388 mm. (B) *M. canettii* strains K_S, K_R and K_R::C9 grown in 7H9 medium supplemented with

389 ADC. Data are representative of at least ten repetitions. Scale bar = 5.0 mm. (C) Ziehl-

390 Neelsen staining of *M. canettii* strains K_S, K_R and K_R::C9. Data are representative of 2

391 repetitions. Scale bar = 10 μm.

392

393 **Figure 3.** PFGE and Southern hybridization analyses with PCR-derived probes binding either

394 in a conserved domain of *pks5* (middle panel) or within the *pap* gene (right panel). Genomes

395 of strains were digested with *MfeI*, resulting in either a 20 or a 12 kb fragment (black arrows).

396 For all C9-complemented strains, the probe hybridized twice, with fragments of 20 kb and 12

397 kb confirming the presence of the natural and the vector-integrated *pks5* loci. PFGE gels were

398 run for 16 h with a pulse of 1 s; *M. canettii* A; *M. tuberculosis* H37Rv (*Mtb*); *M. canettii*

399 strains K_S, K_R, K_R::C9; K_R::vector-control; *M. tuberculosis* H37Rv::C9; cosmid C9; M: low-

400 range PFG Marker (NEB). Data are representative of three repetitions.

401

402 **Figure 4.** Deficient LOS production in rough morphotypes. (A) TLC analysis of lipid extracts

403 from various *M. canettii* strains and *M. tuberculosis* H37Rv. Lipid extracts were dissolved in

404 $CHCl_3$ and run in ($CHCl_3$:CH_3OH:H_2O, 60:24:2). Glycolipids were visualized by spraying

405 with anthrone, followed by charring. Insert corresponds to an enhanced contrast picture

406 showing a faint blue spot corresponding to LOS in *M. canettii* A. This experiment was

407 performed at least twice with similar results. (B) MALDI-TOF mass spectra of purified LOS

13

408   from *M. canettii* strains I_S and K_S. Note that the mass peak distribution for the LOS

409   component of strain I was 14 mass-units lower than that observed for K and A. Further

410   MS/MS fragmentation analysis showed successive losses of the O-methyl acylated trehalose

411   and oligosaccharides from the 3 major precursor-ions for strain I, and thus suggests that the

412   difference of 14 mass-units resulted from the presence of a rhamnosyl unit in LOS of strain I,

413   instead of a 2-O-Me-rhamnosyl residue linked to the β-glucosyl unit (L-Rha*p*-α1->3-D-Glc*p*-

414   β1->3) present in LOS of strains K. (C) TLC analysis ($CHCl_3$:CH_3OH:H_2O, 60:24:2) of lipid

415   extracts of *M. canettii* strains I_S, I_R, I_R::vector-control, I_R::C9, as well as K_S, K_R, K_R::vector

416   control, K_R::C9 and *M. tuberculosis* H37Rv::C9.. This experiment is representative of two

417   experiments performed independently. LOS = lipooligosaccharide. (D) Structure of

418   lipooligosaccharides (LOS) from *M. canettii*. The acyl substituents, primarily 2L-,4L-

419   dimethylhexadecanoate,   2L-,4L-,6L-,8L-tetramethyloctadecanoate,   and   2-methyl-3-

420   hydroxyeicosanoate, are at positions 2,3,6 and 3,4,6 of the terminal glucosyl unit[29].

421

422   **Figure 5.** Survival of morphotypes in different infection models. **(**A-B) Intracellular growth

423   of *M. tuberculosis* H37Rv, *M. canettii* strains K_S, K_R, K_R::C9 and K_R::vector-control in

424   differentiated THP-1 macrophages. THP-1 cells were infected with various strains at an MOI

425   of 0.05 (~ 1 bacterium per 20 cells). CFU of intracellular bacteria were determined 3 h and 3,

426   4 and 5 days post infection (pi). The figure shows CFU numbers (A) and fold growth rates

427   (B). Data are represented as means and standard deviation of at least four independent

428   experiments. Significance in difference was determined using Two-way *ANOVA* (*$p < 0.05$,

429   **$p < 0.001$, ***$p < 0.0001$). (C) Survival of SCID mice infected with different morphotypes.

430   1 x 10^6 CFU per mouse of *M. tuberculosis* H37Rv, *M. canettii* strains K_S, K_R or K_R::C9 were

431   IV-injected and survival of mice was monitored. Endpoints were defined as loss of > 20 % of

432   body weight. The initial dose of all strains was comparable with 4.6 x 10^6 CFU/ml for

433   H37Rv, 3.6 x 10^6 CFU/ml for K_S, 4.1 x 10^6 CFU/ml for K_R and 3.3 x 10^6 CFU/ml for K_R::C9.

434   Data represent one experiment with 10 mice per group. (D) Bacterial load of morphotypes in

435   lungs of infected guinea pigs. Guinea pigs were infected via aerosol route with 2.5x10^6 CFU.

436   Lungs of infected animals were homogenized at day 1 and day 42 pi and CFU of different

437   morphotypes were determined. Data represent median with interquartile range of CFU at day

438   42 divided by CFU at day 1 of two biological replicates with 4 guinea pigs per group (strains

439   K_S and K_R) or 1 experiment with 4 guinea pigs per group (K_R::C9). Significant difference was

440   observed using Kruskal-Wallis test with Dunn's correction (*$p < 0.05$, **$p < 0.01$). (E-F) IL-6

441   and IL-12p40 production in BM-DCs upon infection with different morphotypes. BM-DCs of

442 C57BL/6 WT were infected with Sauton-grown *M. tuberculosis* H37Rv, *M. canettii* strains

443 K_S, K_R or K_R::C9 at an MOI of 1. 24 h pi, levels of IL-6 and IL-12p40 in the cell supernatants

444 were determined by ELISA. Note that a dominant inhibitory effect of LOS on the induction of

445 inflammatory responses was excluded since co-infection of BM-DCs with strains K_S and K_R

446 did not result in less cytokine production than infection with *M. canettii* K_R alone. Data are

447 represented as means and standard deviation of three independent experiments. Significance

448 in difference was determined using Mann Whitney test (*$p < 0.05$). (G-H) Differential

449 interaction of S- and R-morphotypes with TLR2. Human TLR2- or TLR4-transfected

450 HEK293 cells were incubated with S- and R-morphotypes at an MOI of 1. At 24 h pi, levels

451 of secreted embryonic alkaline phosphatase reporter gene under the control of NF-κB were

452 measured with a spectrophotometer. PAM3CSK4 (10 µg/ml) and LPS (100 ng/ml) were used

453 as positive controls for TLR2- or TLR4-mediated stimulation, respectively. Data are

454 represented as means and standard deviation of at least three independent experiments.

455 Significance in difference was determined using Mann-Whitney test (***$p < 0.001$). n.s. non-

456 stimulated.

457

458 **Figure 6.** Scheme showing supposed molecular key events in mycobacterial evolution from

459 the recombinogenic *M. canettii* strain pool of putative environmental origin, towards

460 professional pathogens of mammalian hosts evolved by clonal expansion of one emerging

461 sublineage. Network phylogeny inferred among eight *M. canettii* strains used in this study,

462 and 46 MTBC strains by NeighborNet analysis based on genome sequence data. Gene names

463 and arrow shown in grey refer to previously described differences between *M. canettii* strains

464 and MTBC members[9], whereas the *pks5* recombination event is marked by a red arrow.

465 Figure adapted from Bottai et al.[1] and Boritsch et al.[9]

466

467

METHODS

Mycobacterial strains and growth conditions

Cloning was performed in LB-grown *Escherichia coli* XL-2 (Stratagene) using hygromycin (200 µg·ml⁻¹) or kanamycin (50 µg·ml⁻¹) selection. MTBC and *M. canettii* strains were grown in Middlebrook 7H9 broth (Becton-Dickinson) containing albumin-dextrose-catalase (ADC) or on Middlebrook 7H11 medium (Becton-Dickinson) containing oleic acid-albumin-dextrose-catalase (OADC) or in Sauton medium (when specified) at 37°C. Hygromycin (50 µg·ml⁻¹) was used for mycobacterial selection.

Genome analysis of S/R morphotype isolates

Mycobacterial DNA was prepared by standard procedures[16,17]. After final ethanol precipitation and washing, DNA was resuspended in TE buffer and used for library preparation and Illumina-based genome sequencing. This approach yielded on average 21 million (I_S and I_R) and 8 million (K_S and K_R) reads per strain, which were then mapped against the corresponding reference sequences of STB-I (ENA WGS project CAOO00000000) and STB-K (NC_019951). Reads were aligned against the reference genomes[3] using SHRiMP[51]. Alignment maps were visualized with Tablet[52] and SNPs were called according to coverage sums and variant frequencies. At least 10 reads had to be matched with a substitution frequency > 0.89 to guarantee the detection of high confidence SNPs. Additional mapping and SNP analysis was performed using BioNumerics v7.6 software (Applied Maths), using parameters pre-calibrated based on re-sequencing data of reference genomes[53].

Network phylogenic analysis of *M. canettii* and MTBC strains

Network phylogeny among *M. canettii* and MTBC strains was inferred by NeighborNet analysis, based on pairwise alignments of whole-genome SNP data. Phylogenetic groupings were identified by split decomposition analysis using the SplitsTree4 software[54].

Long-range PCR and *pks5*-sequencing

pks5 and/or *pap* genes were amplified from chromosomal DNA using Platinum-Taq polymerase (Life Technologies) according to the manufacturer's manual with primers EB-1 TTTATTAATCAGGGAAAGGCGACATCGGA and EB-2 TTTTTATAACCGCCAAGAC AAACTTCATC for annealing at 55°C and elongation at 68°C (10 min). Amplicons were Sanger-sequenced after QIAquick (Quiagen) purification, using primers listed in

502　supplementary Table **S5**. Amplification of *pks5-1* with the last part of *pap* or amplification of

503　*pks5-2* with the first part of *pap* to verify the presence of both *pks5* genes in the various *M.*

504　*canettii* strains was performed using Pwo Polymerase (Roche) and the oligos

505　EB-3　　　　　　GACGAACTACTAGTCGTTGGTAGCG　　　　and　　　EB-4

506　TTTTTATAACCGCCAAGACAAACTTCATC,　　　　　　and　　　　EB-5

507　TTTCAGGGAAAGGCGACATCGGA　　　　　　and　　　　EB-6

508　TTTCGCTACCAACGACTAGTAGTTCGTC, respectively.

509

Complementation of R-variants

511　R-variants of *M. canettii* strains K and I, as well as *M. tuberculosis* H37Rv and other MTBC

512　members were transformed with the integrating cosmid C9 and H6 spanning the extended or

513　partial *pks5* locus of *M. canettii* strain A. The C9 and H6 clones were selected from a cosmid

514　library that was constructed in pYUB412[28] by using partially-digested, agarose-embedded

515　genomic DNA[55,56] of *M. canettii* reference strain CIPT 140010059 (STB-A). Clone-inserts

516　were PCR-verified using primers binding in vector- or insert-sequences: (T7-F

517　AGGCATGCAAGCTCAGGATA;　T7-R　GGATCGGTCCAGTAATCGT　and　T3-F

518　GCAGAAGCACTAGACGATCC; T3-R GCCGCAATTAACCCTCACTA). Transformed,

519　hygromycin-resistant R-variants that exhibited no S-morphology served as vector-controls.

520　Insert-termini sequencing of these latter clones showed partial insert-loss including the *pks5*

521　region, whereas genes encoding MCAN_15411, MCAN_15421 and part of MCAN 15431

522　remained present.

523　　　　For construction of vectors expressing HA-tagged proteins, corresponding genes were

524　amplified with Pwo polymerase (Roche). For the construction of the HA-tagged *pks5-2*

525　expression plasmid, pTTP1b[57] was cut with *Hin*dIII (NEB) to remove the kanamycin

526　resistance cassette and blunt ends were generated using Klenow fragment (NEB). A

527　hygromycin resistance cassette was amplified from pAL70[58] with the oligos AC123

528　ACAGGCCTGTCGTCGAGGTCCACCAA　　　　　　and　　　　　　AC124

529　ACAGGCCTGGATGCCAGGGCCTTTCA, cut with *Stu*I and cloned into the digested vector

530　to generate pAL232. The strong mycobacterial promoter of the gene *hsp60* was amplified

531　from chromosomal DNA of *M. tuberculosis* H37Rv with the oligos NheI_HindII_hsp60_F

532　aaaGCTAGCAAGCTTggtgaccacaacgacgcgcccgctttgatc　and　hsp60_SpeI_EcoRV_XbaI_R

533　aaaTCTAGAgatatcACTAGTtgtcttggccattgcgaagtgattcctcc, the amplified product was cut

534　with *Nhe*I and *Xba*I (NEB) and cloned into *Xba*I-digested pAL232 to generate pEB18. The

535　gene *pks5-2* was amplified from chromosomal DNA of *M. canettii* strain K_S with the oligos

536 pks5-2_SpeI_F aaaACTAGTGTGGGTAAGGAGAGAACAAAG and pks5-2_HA_PsiI_R

537 aaaTTATAAttaAGCATAATCAGGAACATCATACGGATATGAAGGTGCTGCAATGTC

538 GG. The amplified product was cut with *Spe*I and *Psi*I and cloned into *Spe*I-*Eco*RV digested

539 pEB18. For construction of the HA-tagged *pap* expression system, the plasmid pMV10-25

540 was first cut with *Nco*I, blunt ended using Klenow fragment (NEB) and subsequently cut with

541 *Nhe*I. The *pap* gene was amplified using the oligos

542 *pap_SpeI_F*   aaaACTAGTGTGATCATTGGCGGGGGC   and   pap_HA_*PsiI_R*

543 aaaTTATAAttaAGCATAATCAGGAACATCATACGGATAGCTAGATACGCGAACTGC

544 TG. The amplified product was digested with *Spe*I and *Psi*I and cloned into the *Nhe*I/ blunt

545 end pMV10-25[59] vector.

546 Whole cell lysates were generated from cultures grown in Middelbrook 7H9 supplemented

547 with ADC to an OD of 0.6-0.8. Cells were disrupted using a TissueLyser II (Quiagen) and

548 Zirkonia/ Silica beads (BioSpec) and passed through 0.2 µm filter units. Proteins were

549 separated on NuPAGE Novex 10% Bis-Tris gels (Invitrogen) for HA-tagged *pap* and

550 NuPAGE Novex 4-12% Bis-Tris gels (Invitrogen) for HA-tagged *pks5-2*. HA-tagged *pap*

551 was blotted on a nitrocellulose membrane using an iBlot Dry Blotting System

552 (LifeTechnologies), while *pks5-2* was blotted on nitrocellulose membranes using a wet

553 transfer system for 2 h at 100 V at 4°C. Membranes were blocked in 5% milk powder in TBS

554 and subsequently incubated with either mouse anti-HA (anti-HA.11; Covance) or rabbit anti-

555 SigA (Statens Serum Institute, Copenhagen, Denmark) primary antibody. Bound antibodies

556 were detected using a horseradish peroxidase-conjugated sheep anti-mouse or donkey anti-

557 rabbit (Amersham ECL) antibody and the chemiluminescent signal was developed with a

558 SuperSignal West Femto Maximum Sensitivity Substrate (Thermo Scientific). Image

559 acquisition was performed with an Azure c300 (Azure Biosystems).

560

561 **Ziehl-Neelsen staining**

562 Mycobacteria were grown to an OD of 0.4, smeared on glass slides and killed at 95°C (1 h).

563 After heat-fixation with 90% ethanol, mycobacteria were stained with carbol-fuchsin and

564 heated (10 min). After washing, glass slides were incubated with sulphuric acid (25%) (2

565 min), washed and incubated with 90% ethanol (5 min), washed and counter-stained with

566 methylene blue (2 min).

567

568 **PFGE and Southern Blot hybridization**

569   Genomic DNA in agarose-plugs, prepared as described[55], was *Mfe*I-digested, separated by

570   pulsed-field gel electrophoresis (Biorad CHEF II, pulse 1 s for 16 h at 6 V·cm^{-1}) and

571   transferred onto Hybond-C-Extra nitrocellulose (GE), as described[55]. Hybridization was

572   performed with [α-^{32}P]dCTP-labeled PCR-probes at 68°C in 6xSSC/ 0,5% SDS/ 0.01 M

573   EDTA/ 5xDenhardt's solution/ 100 mg·ml^{-1}salmon-sperm DNA. After washing, membranes

574   were exposed to phosphorimager screens, which were scanned in a STORM phosphorimager.

575   Probes were amplified from *M. tuberculosis* DNA using *pks5* primers (pks5-F

576   GTTGTGGGAGGCGTTGCT; pks5-R GAAACGTCGAACGCATGAC) or from *M. canettii*

577   K$_S$ using *pap* primers (pap-F CTCGATTATTCACGGCTGGT; pap-R

578   CGTATAGCCCGGTGATCAAC).

579

580   **Extraction and analysis of lipooligosaccharides**

581   Mycobacterial cells obtained from 7H9-grown cultures were exposed to CHCl$_3$/CH$_3$OH (1:2,

582   v/v) for 48h to kill bacteria. Lipids were extracted following previously described

583   protocols[14,21] and separated by TLC. Finally, glycolipids were visualized by spraying the

584   plates with a 0.2% anthrone solution (w/v) in concentrated H$_2$SO$_4$, followed by heating. The

585   previously described LOS of *M. canettii* strain CIPT 140010059 was used as standard[29].

586   Crude lipid extracts were subjected to chromatography on a Sep-Pak Florisil cartridge and

587   eluted at various concentrations of CH$_3$OH (0, 5, 10, 15, 20, 30%) in CHCl$_3$. Each fraction

588   was analyzed by TLC on Silica Gel G60 using CHCl$_3$/CH$_3$OH/H$_2$O (30:12:1, v/v/v) as solvent

589   and glycolipids were visualized as described above.

590   MALDI-TOF/TOF-MS and MS/MS analyses were conducted in the positive

591   ionization and reflectron mode by accumulating 10 spectra of 250 laser shots, using the 5800

592   MALDI TOF/TOF Analyser (Applied Biosystems/Absciex) equipped with a Nd:Yag laser

593   (349nm). For MS and MS/MS data acquisitions, uniform, continuous, and random stage

594   motion was selected at a fixed laser intensity of 4000 (instrument-specific units) and 400 Hz

595   pulse rate and 6000 (instrument-specific units) and 1000Hz, respectively. For MS/MS data

596   acquisition, the fragmentation of selected precursor-ions was performed at collision energy of

597   1 kV. Lipid samples were dissolved in chloroform and spotted onto the target plate as 0.5 μl

598   droplets, followed by the addition of 0.5 μl matrix solution (10 mg of 2,5-dihydroxybenzoic

599   acid [Sigma-Aldrich]/ml in CHCl$_3$/CH$_3$OH, 1/1 [vol/vol]. Samples were allowed to crystallize

600   at room temperature. Spectra were externally calibrated using lipid standards.

601   Alkaline hydrolysis of purified LOS and total extractable lipids from the various

602   strains was performed with 1M sodium methanolate for 1h at 37°C. After neutralization with

603 glacial acetic acid, the mixture was dried under stream of nitrogen and lipids were extracted

604 with diethyl ether and washed twice with water. The resulting fatty acid methyl esters were

605 converted to trimethylsilyl (TMS) derivatives, using a mixture of

606 pyridine/hexamethyldisilazane/trimethylchlorosilane (6:4:2, v/v/v).

607 GC-MS analyses were performed using a Thermo TraceGCultra chromatograph

608 coupled with an ISQ mass spectrometer. Chromatographic separations of the TMS derivatives

609 of the fatty acid methyl esters were obtained using an Inferno ZB5HT column of 15m. Helium

610 was the carrier gas at constant flow rate of 1.2mL min^{-1}. The oven temperature program was

611 started at 120°C, ramped to 380°C at 10°C/min (with final isothermal step of 5min at 380°C).

612 The temperature of the injector was 220°C and injection of 1μl of samples in petroleum ether

613 was performed in a split mode (ratio of 20:1). EI mass spectra were recorded using electron

614 energy of 70eV from 60 to 600 with a transfer line maintained at 275°C.

615 To visualize various polyketide-derive lipids, including acyl-trehaloses such

616 diacyltrehaloses (DAT), polyacyltrehaloses (PAT), sulfolipids (SL), phenolic glycolipids

617 (PGL) or lipooligosaccharides (LOS) from different *M. canettii* strains metabolic labelling

618 with ^{14}C propionate was used. For this TLC analysis, *M. canettii* strains were grown to the

619 exponential phase in 10 mL 7H9 liquid medium supplemented with ADC and 0.05% Tween

620 80 and labelled by incubation with 0.4 μCi.ml^{-1} [1-^{14}C] propionate for 24 h. The TLC plates

621 were run in CHCl$_3$/CH$_3$OH/H$_2$O (60/16/2) for DAT and SL and CHCl$_3$/CH$_3$OH (99/1) for

622 PAT. Labelled lipids were visualized with a Typhoon PhosphorImager (Amersham

623 Biosciences).

624

625 **Macrophage infections**

626 THP-1 (TIB-202D) human monocyte-like cells were purchased from ATCC, directly

627 amplified and stocked in liquid nitrogen. Only low passage cells (passage number <11) were

628 used in the experiments. The purchased THP-1 cell line has been authenticated and tested for

629 microbial contaminants, including mycoplasma, by ATCC. For the experiments, THP-1 were

630 cultivated in RPMI 1640, GlutaMAX (Life Technologies) containing 10 % heat-inactivated

631 fetal bovine serum (Life Technologies), seeded at a density of 7.5x10^4 cells per well in 96

632 well plates and differentiated into macrophages through incubation with 50 mM PMA for 3

633 days. For infection, bacteria grown in Sauton medium without shaking were sonicated, added

634 to the macrophages at an MOI of 0.05, (~ 1 bacterium per 20 THP-1 cells) and incubated for 2

635 h. Sauton medium was used, as it allows the production of more complex polar lipids[18]. After

636 phagocytosis, 0.1 mg·ml^{-1} amikacin was added for 1 h to remove extracellular bacteria and

cells were incubated for up to 6 days at 37°C and 5% CO_2. At various time points macrophages were lysed with 0.1% Triton-X100 in PBS and lysates plated in serial dilutions on 7H11+OADC plates to determine intracellular survival of bacteria in CFU. Experiments were performed as at least four biological replicates, each done in triplicate (technical replicates).

Additional infection experiments were conducted using Raw murine macrophages as well as Human monocyte-derived macrophages (hMDMs). Raw cells were cultivated in RPMI 1640 Medium, GlutaMAX (Gibco, Life Technologies) supplemented with 5 % heat-inactivated fetal bovine serum (Gibco, Life Technologies) and seeded one day before infection. In parallel, hMDMs were obtained from buffy coats by centrifugation of blood from healthy human donors in lymphocyte separation medium (Eurobio) and further isolation of CD14[+] monocytes from the mononuclear cell fraction using CD14 microbeads (Miltenyi Biotec). Monocytes were differentiated into macrophages in the presence of rhM-CSF (50 ng/ml; R&D Systems). At the day of infection mycobacterial strains grown in Sauton medium without shaking were sonicated, added to the macrophages at a MOI of 1:20, and incubated for 2 hrs at 37°C and 5% CO_2. After phagocytosis, infected macrophages were incubated with 0.1 mg·ml^{-1} amikacin for 1 h to remove extracellular bacteria and finally incubated for 6 days in new RPMI supplemented with 10 % FBS for hMDMs and 1% for Raw cells at 37°C and 5 % CO_2. At selected time points macrophages were lysed with 0.1 % Triton-X100 in PBS and bacteria were plated in serial dilutions on 7H11 plates supplemented with OADC to determine CFU counts. Buffy coats were obtained from healthy donors after informed consent (Etablissement Français du Sang). Three biological replicates were performed, each done in triplicate (technical replicates).

Animal infection studies

Six-week-old female SCID mice (Charles River) were infected intravenously with 200 µl of 5×10^6 bacteria/mouse and survival of mice was monitored. Humane endpoints were defined as loss of > 20 % of body-weight. The experiment was performed once, using 10 mice per strain (technical replicates). In a second well-established mycobacterial infection model, five to six week-old female guinea pigs (Hartley; Charles River) were aerosol-infected using 5 ml of a suspension containing 5×10^5 bacteria·ml^{-1}. Six weeks post-infection, animals were killed and organs homogenized using a gentleMACS Dissociator (Miltenyi Biotec) and gentleMACS M tubes. Bacterial loads in organs were determined by plating serial dilutions of organ homogenates on solid medium. The number of animals included in the experiments

671 (sample size choice) was determined by taking into account the rule of 3Rs (replacement,

672 reduction, refinement) and statistical requirements. As regards randomization and blinding,

673 CFU counts were recorded in parallel by different investigators.

674

675 **Cytokine and chemokine assays and HEK-Blue cell reporter assay**

676 ELISA and ProcartaPlex Luminex immunoassay from culture supernatants of infected BM-

677 DCs was performed as previously described[60], and/or according to the manufacturer's

678 instructions (ProcartaPlex Luminex Immunoassay). mAbs specific to IL-12p40 and IL-6 were

679 from BD Biosciences. Reagents for quantitative ProcartaPlex Luminex immunoassay were

680 from affymetrix eBioscience. ELISA experiments were performed as three biological

681 replicates, while Luminex analyses were performed as two independent biological replicates,

682 each done in duplicate. Signal acquisition was performed on pooled duplicates.

683 HEK-TLR2 and HEK-TLR4 cells (InvivoGen) were grown in DMEM medium (Gibco Life

684 Technologies) supplemented with 10% FBS, 100 $\mu g \cdot ml^{-1}$ Normocin and 1X HEK-Blue

685 Selection. TLR stimulation assay was performed according to the manufacturer's protocol. 10

686 $\mu g \cdot ml^{-1}$ PAM3CSK4 served as positive control for HEK-TLR2 and 100 $ng \cdot ml^{-1}$ LPS as

687 positive control for HEK-TLR4. Experiments were performed as at least three biological

688 replicates.

689

690 **Compliance with regulations on animal welfare and ethics**

691 Animal studies were performed in agreement with European and French guidelines (Directive

692 86/609/CEE and Decree 87–848 of 19 October 1987) after approval by the Institut Pasteur

693 Safety Committee (Protocol 11.245) and local ethical committees (CNREEA 2012-0061;

694 CETEA 2013-0036).

695

696 **References:**

697 1 Bottai, D., Stinear, T. P., Supply, P. & Brosch, R. Mycobacterial Pathogenomics and
698 Evolution *Microbiology Spectrum* **2**, MGM2-0025-2013 (2014).
699 2 Le Chevalier, F., Cascioferro, A., Majlessi, L., Herrmann, J. L. & Brosch, R.
700 *Mycobacterium tuberculosis* evolutionary pathogenesis and its putative impact on
701 drug development. *Future Microbiol.* **9:969-85.**, (2014).
702 3 Supply, P. *et al.* Genomic analysis of smooth tubercle bacilli provides insights into
703 ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat Genet* **45**, 172-179
704 (2013).
705 4 van Soolingen, D. *et al.* A novel pathogenic taxon of the *Mycobacterium tuberculosis*
706 complex, Canetti: characterization of an exceptional isolate from Africa. *Int J Syst*
707 *Bacteriol* **47**, 1236-1245. (1997).

708	5	Gutierrez, M. C. *et al.* Ancient origin and gene mosaicism of the progenitor of
709		Mycobacterium tuberculosis. *PLoS Pathog.* **1**, e5.
710	6	Koeck, J. L. *et al.* Clinical characteristics of the smooth tubercle bacilli
711		'Mycobacterium canettii' infection suggest the existence of an environmental reservoir.
712		*Clin Microbiol Infect* **17**, 1013-1019 (2011).
713	7	Blouin, Y. *et al.* Progenitor "*Mycobacterium canettii"* Clone Responsible for Lymph
714		Node Tuberculosis Epidemic, Djibouti. *Emerg Infect Dis* **20**, 21-28, (2014).
715	8	Dormans, J. *et al.* Correlation of virulence, lung pathology, bacterial load and delayed
716		type hypersensitivity responses after infection with different *Mycobacterium*
717		*tuberculosis* genotypes in a BALB/c mouse model. *Clin Exp Immunol.* **137**, 460-468.
718		(2004).
719	9	Boritsch, E. C. *et al.* A glimpse into the past and predictions for the future: the
720		molecular evolution of the tuberculosis agent. *Mol Microbiol.* **93**, 835-852. (2014).
721	10	Panas, M. W. *et al.* Noncanonical SMC protein in Mycobacterium smegmatis restricts
722		maintenance of *Mycobacterium fortuitum* plasmids. *Proc Natl Acad Sci U S A.* **111**,
723		13264-13271. (2014).
724	11	Kansal, R. G., Gomez-Flores, R. & Mehta, R. T. Change in colony morphology
725		influences the virulence as well as the biochemical properties of the *Mycobacterium*
726		*avium* complex. *Microb Pathog* **25**, 203-214, (1998).
727	12	Catherinot, E. *et al.* Hypervirulence of a rough variant of the *Mycobacterium*
728		*abscessus* type strain. *Infect Immun* **75**, 1055-1058, (2007).
729	13	Belisle, J. T. & Brennan, P. J. Chemical basis of rough and smooth variation in
730		mycobacteria. *J Bacteriol.* **171**, 3465-3470. (1989).
731	14	van der Woude, A. D. *et al.* Unexpected link between lipooligosaccharide biosynthesis
732		and surface protein release in *Mycobacterium marinum*. *J Biol Chem* **287**, 20417-
733		20429, (2012).
734	15	Barrow, W. W. & Brennan, P. J. Isolation in high frequency of rough variants of
735		Mycobacterium intracellulare lacking C-mycoside glycopeptidolipid antigens. *J*
736		*Bacteriol.* **150**, 381-384. (1982).
737	16	Howard, S. T. *et al.* Spontaneous reversion of *Mycobacterium abscessus* from a
738		smooth to a rough morphotype is associated with reduced expression of
739		glycopeptidolipid and reacquisition of an invasive phenotype. *Microbiology* **152**,
740		1581-1590, (2006).
741	17	Pawlik, A. *et al.* Identification and characterization of the genetic changes responsible
742		for the characteristic smooth-to-rough morphotype alterations of clinically persistent
743		*Mycobacterium abscessus*. *Mol.Microbiol.***90**, 612-629. (2013).
744	18	Burguiere, A. *et al.* LosA, a key glycosyltransferase involved in the biosynthesis of a
745		novel family of glycosylated acyltrehalose lipooligosaccharides from *Mycobacterium*
746		*marinum*. *J Biol Chem.* **280**, 42124-42133. (2005).
747	19	Ren, H. *et al.* Identification of the lipooligosaccharide biosynthetic gene cluster from
748		*Mycobacterium marinum*. *Mol Microbiol* **63**, 1345-1359 (2007).
749	20	Lemassu, A., Levy-Frebault, V. V., Laneelle, M. A. & Daffe, M. Lack of correlation
750		between colony morphology and lipooligosaccharide content in the *Mycobacterium*
751		*tuberculosis* complex. *J Gen Microbiol.* **138**, 1535-1541. (1992).
752	21	Etienne, G. *et al.* Identification of the polyketide synthase involved in the biosynthesis
753		of the surface-exposed lipooligosaccharides in mycobacteria. *J Bacteriol* **191**, 2613-
754		2621 (2009).
755	22	Nataraj, V. *et al.* MKAN27435 is required for the biosynthesis of higher subclasses of
756		lipooligosaccharides in *Mycobacterium kansasii*. *PLoS One.* **10**, e0122804. (2015).

757   23   Minnikin, D. E. *et al.* in *Tuberculosis - Expanding Knowledge* (ed W. Ribon) Ch. 7,
758   (InTech, 2015).
759   24   Stinear, T. P. *et al.* Insights from the complete genome sequence of *Mycobacterium*
760   *marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Res* **18**, 729-741.
761   (2008).
762   25   Wang, J. *et al.* Insights on the Emergence of *Mycobacterium tuberculosis* from the
763   Analysis of *Mycobacterium kansasii*. *Genome Biol Evol.* **7**, 856-870. doi:
764   810.1093/gbe/evv1035. (2015).
765   26   Quadri, L. E. Biosynthesis of mycobacterial lipids by polyketide synthases and
766   beyond. *Crit Rev Biochem Mol Biol.* **49**, 179-211. (2014).
767   27   Rousseau, C. *et al.* Virulence attenuation of two Mas-like polyketide synthase mutants
768   of *Mycobacterium tuberculosis*. *Microbiology.* **149**, 1837-1847. (2003).
769   28   Bange, F. C., Collins, F. M. & Jacobs, W. R., Jr. Survival of mice infected with
770   *Mycobacterium smegmatis* containing large DNA fragments from *Mycobacterium*
771   *tuberculosis*. *Tuber Lung Dis* **79**, 171-180. (1999).
772   29   Daffe, M., McNeil, M. & Brennan, P. J. Novel type-specific lipooligosaccharides from
773   *Mycobacterium tuberculosis*. *Biochemistry.* **30**, 378-388. (1991).
774   30   Angala, S. K., Belardinelli, J. M., Huc-Claustre, E., Wheat, W. H. & Jackson, M. The
775   cell envelope glycoconjugates of *Mycobacterium tuberculosis*. *Crit Rev Biochem Mol*
776   *Biol.* **49**, 361-399. (2014).
777   31   Onwueme, K. C., Ferreras, J. A., Buglino, J., Lima, C. D. & Quadri, L. E.
778   Mycobacterial polyketide-associated proteins are acyltransferases: proof of principle
779   with *Mycobacterium tuberculosis* PapA5. *Proc Natl Acad Sci U S A.* **101**, 4608-4613.
780   (2004).
781   32   Bhatt, K., Gurcha, S. S., Bhatt, A., Besra, G. S. & Jacobs, W. R., Jr. Two polyketide-
782   synthase-associated acyltransferases are required for sulfolipid biosynthesis in
783   *Mycobacterium tuberculosis*. *Microbiology.* **153**, 513-520. (2007).
784   33   Bottai, D. *et al.* Increased protective efficacy of recombinant BCG strains expressing
785   virulence-neutral proteins of the ESX-1 secretion system. *Vaccine.* **33**, 2710-2718.
786   (2015).
787   34   Rhoades, E. R. *et al. Mycobacterium abscessus* Glycopeptidolipids mask underlying
788   cell wall phosphatidyl-myo-inositol mannosides blocking induction of human
789   macrophage TNF-alpha by preventing interaction with TLR2. *J Immunol* **183**, 1997-
790   2007, (2009).
791   35   Roux, A. L. *et al.* Overexpression of proinflammatory TLR-2-signalling lipoproteins
792   in hypervirulent mycobacterial variants. *Cell Microbiol* **13**, 692-704, (2011).
793   36   Robinson, R. T., Orme, I. M. & Cooper, A. M. The onset of adaptive immunity in the
794   mouse model of tuberculosis and the factors that compromise its expression. *Immunol*
795   *Rev.* **264**, 46-59. (2015).
796   37   Achtman, M. Insights from genomic comparisons of genetically monomorphic
797   bacterial pathogens. *Philos Trans R Soc Lond B Biol Sci* **367**, 860-867 (2012).
798   38   Portevin, D. *et al.* A polyketide synthase catalyzes the last condensation step of
799   mycolic acid biosynthesis in mycobacteria and related organisms. *Proc Natl Acad Sci*
800   *U S A* **101**, 314-319. (2004).
801   39   Constant, P. *et al.* Role of the *pks15/1* gene in the biosynthesis of phenolglycolipids in
802   the *Mycobacterium tuberculosis* complex. Evidence that all strains synthesize
803   glycosylated *p*-hydroxybenzoic methyl esters and that strains devoid of
804   phenolglycolipids harbor a frameshift mutation in the *pks15/1* gene. *J Biol Chem* **277**,
805   38148-38158, (2002).

806   40   Matsunaga, I. *et al. Mycobacterium tuberculosis pks12* produces a novel polyketide
807   presented by CD1c to T cells. *J Exp Med.* **200**, 1559-1569. (2004).
808   41   Rombouts, Y. *et al.* Fatty acyl chains of *Mycobacterium marinum*
809   lipooligosaccharides: structure, localization and acylation by PapA4 (MMAR_2343)
810   protein. *J Biol Chem.* **286**, 33678-33688. (2011).
811   42   Alibaud, L. *et al.* Increased phagocytosis of *Mycobacterium marinum* mutants
812   defective in lipooligosaccharide production: a structure-activity relationship study. *J*
813   *Biol Chem.* **289**, 215-228. (2014).
814   43   Eckstein, T. M., Inamine, J. M., Lambert, M. L. & Belisle, J. T. A genetic mechanism
815   for deletion of the *ser2* gene cluster and formation of rough morphological variants of
816   *Mycobacterium avium. J Bacteriol* **182**, 6177-6182 (2000).
817   44   Ortalo-Magne, A. *et al.* Identification of the surface-exposed lipids on the cell
818   envelopes of *Mycobacterium tuberculosis* and other mycobacterial species. *J Bacteriol*
819   **178**, 456-461 (1996).
820   45   Mortaz, E. *et al.* Interaction of Pattern Recognition Receptors with *Mycobacterium*
821   *tuberculosis. J Clin Immunol* **14**, 14 (2014).
822   46   Cambier, C. J. *et al.* Mycobacteria manipulate macrophage recruitment through
823   coordinated use of membrane lipids. *Nature.* **505**, 218-222. (2014).
824   47   Gopinath, K., Moosa, A., Mizrahi, V. & Warner, D. F. Vitamin B(12) metabolism in
825   *Mycobacterium tuberculosis. Future Microbiol* **8**, 1405-1418 (2013).
826   48   Young, D. B., Comas, I. & de Carvalho, L. P. Phylogenetic analysis of vitamin B12-
827   related metabolism in Mycobacterium tuberculosis. *Front Mol Biosci.* **2**, 6. (2015)
828   49   Danilchanka, O. *et al.* An outer membrane channel protein of *Mycobacterium*
829   *tuberculosis* with exotoxin activity. *Proc Natl Acad Sci U S A.* **111**, 6750-6755.
830   (2014).
831   50   Delahay *et al.* The coiled-coil domain of EspA is essential for the assembly of the type
832   III secretion translocon on the surface of enteropathogenic *Escherichia coli. J Biol*
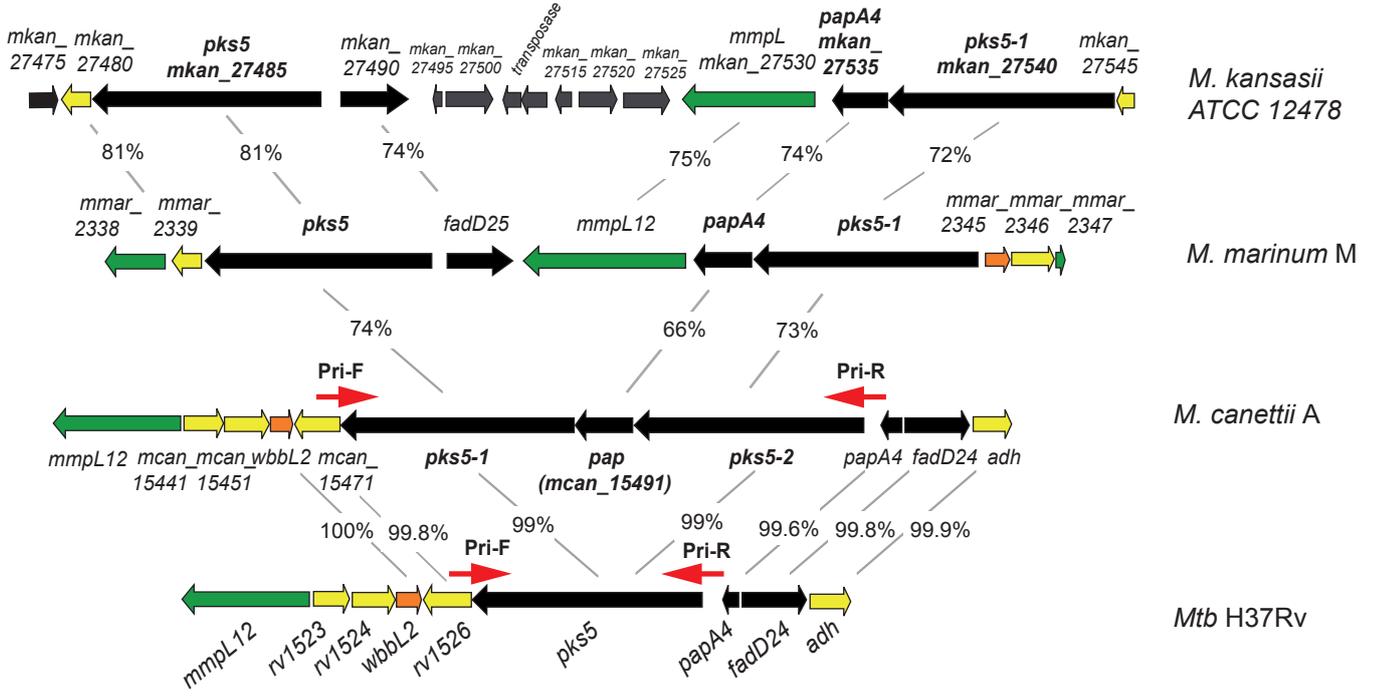833   *Chem.* **274**, 35969-35974.
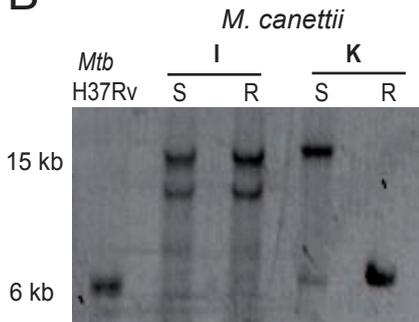834

835 **References Methods**

836
837   51   David, M., Dzamba, M., Lister, D., Ilie, L. & Brudno, M. SHRiMP2: sensitive yet
838   practical SHort Read Mapping. *Bioinformatics.* **27**, 1011-1012. (2011).
839   52   Milne, I. *et al.* Using Tablet for visual exploration of second-generation sequencing
840   data. *Brief Bioinform.* **14**, 193-202. (2013).
841   53   Pouseele, H. & Supply, P. Accurate whole-genome sequencing-based epidemiological
842   surveillance of mycobacterium tuberculosis. *Methods in Microbiology.*
843   doi:10.1016/bs.mim.2015.04.001 (2015).
844   54   Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary
845   studies. *Mol Biol Evol* **23**, 254-267 (2006).
846   55   Brosch, R. *et al.* Comparative genomics uncovers large tandem chromosomal
847   duplications in Mycobacterium bovis BCG Pasteur. *Yeast* **17**, 111-123. (2000).
848   56   Brosch, R. *et al.* Use of a Mycobacterium tuberculosis H37Rv bacterial artificial
849   chromosome library for genome mapping, sequencing, and comparative genomics.
850   *Infect Immun* **66**, 2221-2229 (1998).
851   57   Pham, T. T., Jacobs-Sera, D., Pedulla, M. L., Hendrix, R. W. & Hatfull, G. F.
852   Comparative genomic analysis of mycobacteriophage Tweety: evolutionary insights
853   and construction of compatible site-specific integration vectors for mycobacteria.
854   *Microbiology.* **153**, 2711-2723 (2007).

855   58   Cascioferro, A. *et al.* Xer site-specific recombination, an efficient tool to introduce
856   unmarked deletions into mycobacteria. *Appl Environ Microbiol* **76**, 5312-5316 (2010).
857   59   Delogu, G. *et al.* Rv1818c-encoded PE_PGRS protein of Mycobacterium tuberculosis
858   is surface exposed and influences bacterial cell structure. *Mol Microbiol* **52**, 725-733
859   (2004).
860   60   Majlessi, L. *et al.* Influence of ESAT-6 Secretion System 1 (RD1) of *Mycobacterium*
861   *tuberculosis* on the Interaction between Mycobacteria and the Host Immune System. *J*
862   *Immunol* **174**, 3570-3579 (2005).

863

A

M. canettii

K_S | K_R | $K_R::C9$

I_S | I_R | $I_R::C9$

Mtb H37Rv | *Mtb* H37Rv::C9

B

K_S | K_R | $K_R::C9$

C

K_S | K_R | $K_R::C9$

A

B

C

D

Mycobacterium tuberculosis complex

Mycobacterium canettii strain pool

pe_pgrs33
rv3902/03c
pks8/17 split
Δ*cobF*

Lineage 1

Lineage 4

Lineage 2

Lineage 3

Lineages 5 & 6
& animal strains

recombination of *pks5*
and deletion of *pap*

0.01