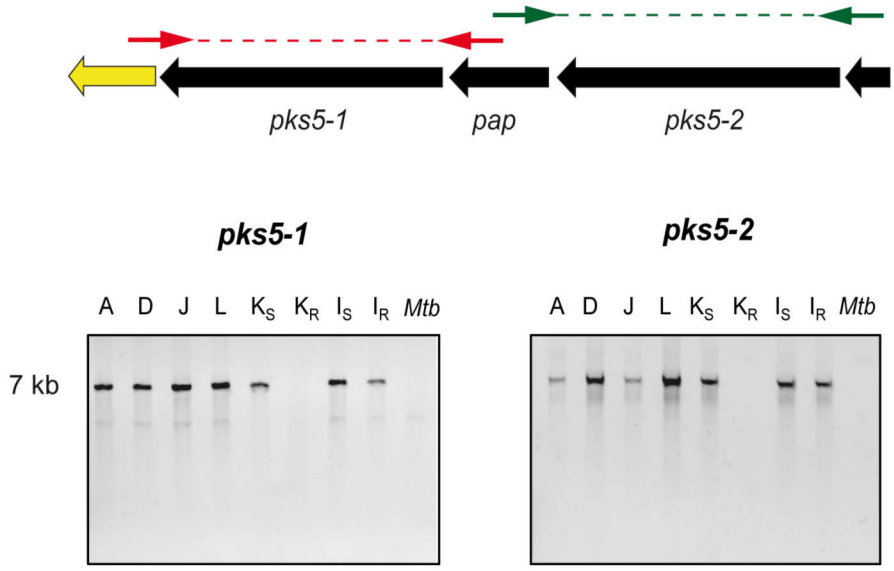
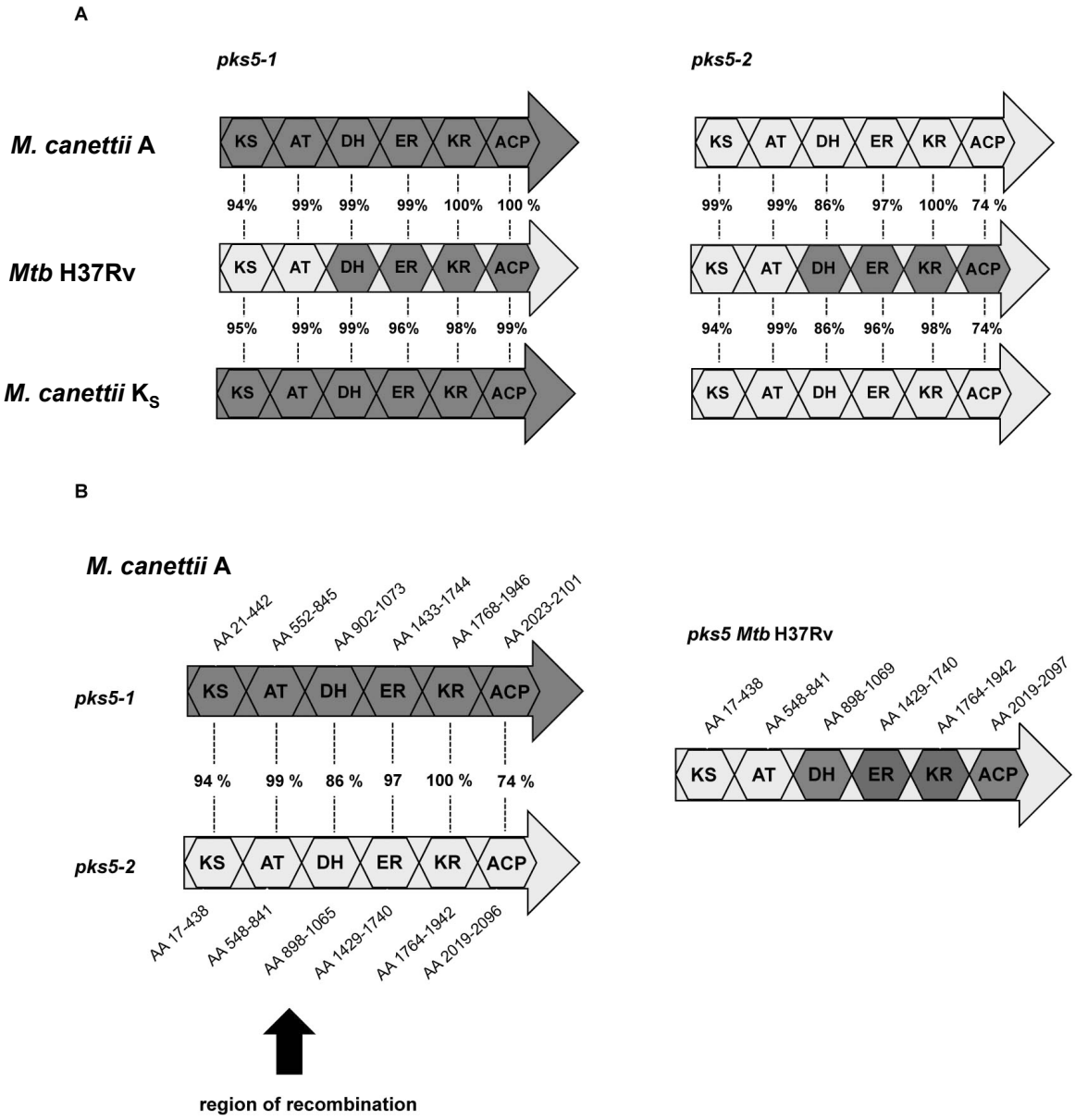
Supplementary Information:

~~Key role of~~ Pks5 recombination-mediated surface remodelling in *Mycobacterium tuberculosis* emergence

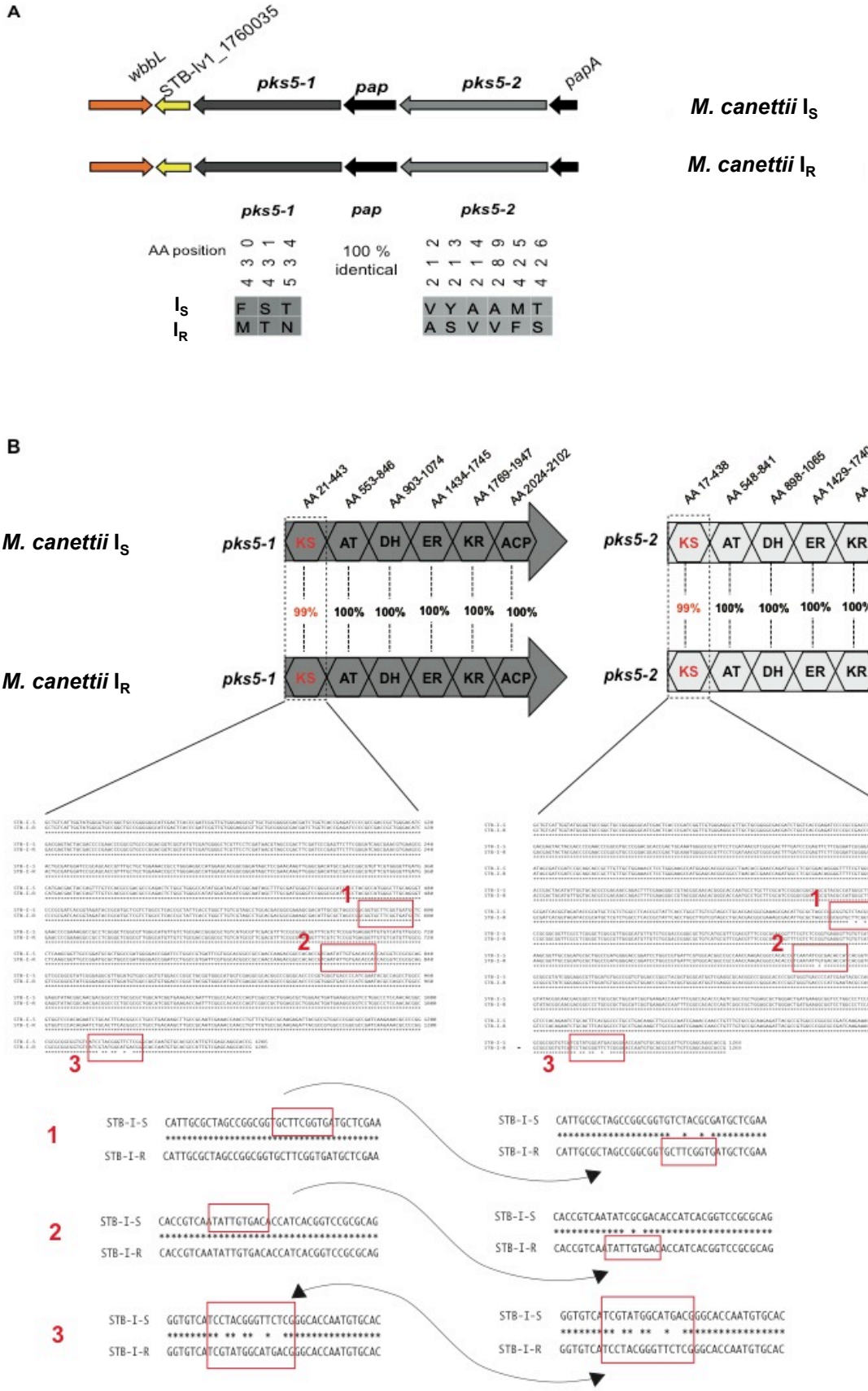
Eva C. Boritsch, Wafa Frigui, Alessandro Cascioferro, Wladimir Malaga, Gilles Etienne, Françoise Laval, Alexandre Pawlik, Fabien Le Chevalier, Mickael Orgeur, Laurence Ma, Christiane Bouchier, Timothy P. Stinear, Philip Supply, Laleh Majlessi, Mamadou Daffé, Christophe Guilhot and Roland Brosch.



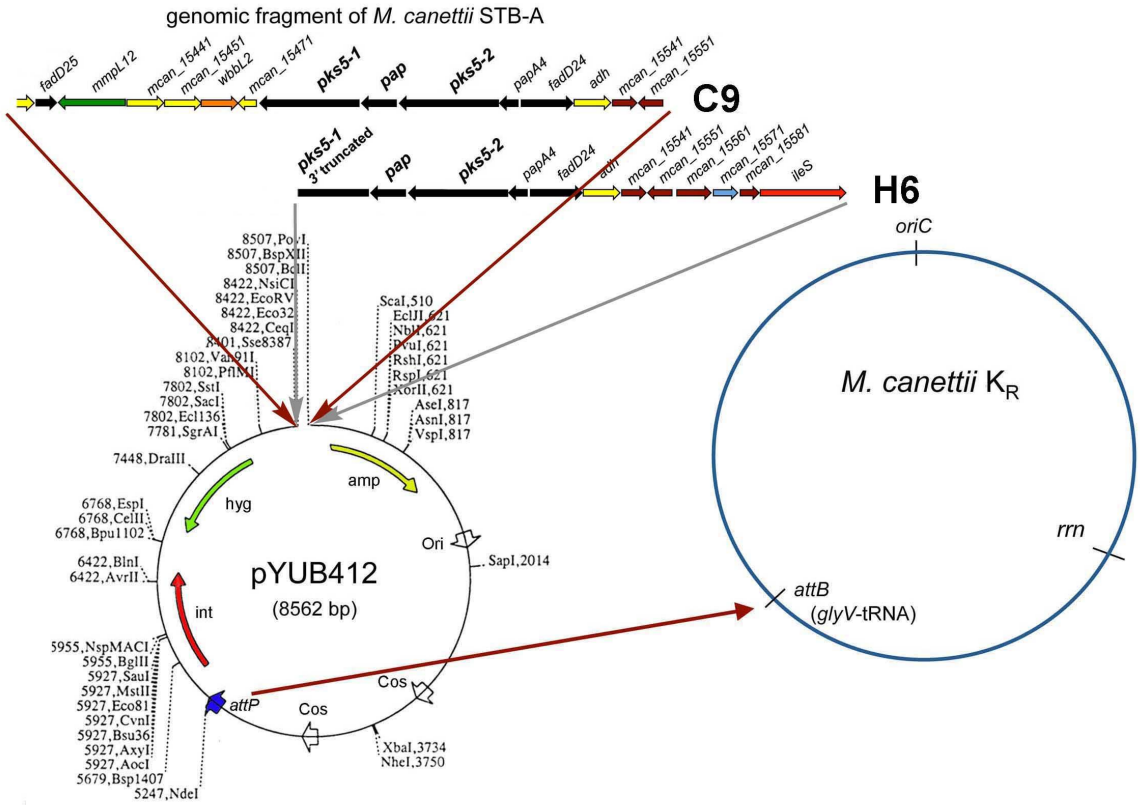
Supplementary Figure 1. Smooth *M. canettii* strains contain both *pks5* genes and *pap*. Long range PCRs were performed using oligos that bind either downstream *pks5-1* and inside *pap* to amplify *pks5-1* (red dotted line) or upstream *pks5-2* as well as inside *pap* to amplify *pks5-2* (green dotted line). *M. canettii strains* A, D, J, L, K_S, K_R, I_S, I_R and *M. tuberculosis* H37Rv (*Mtb*).



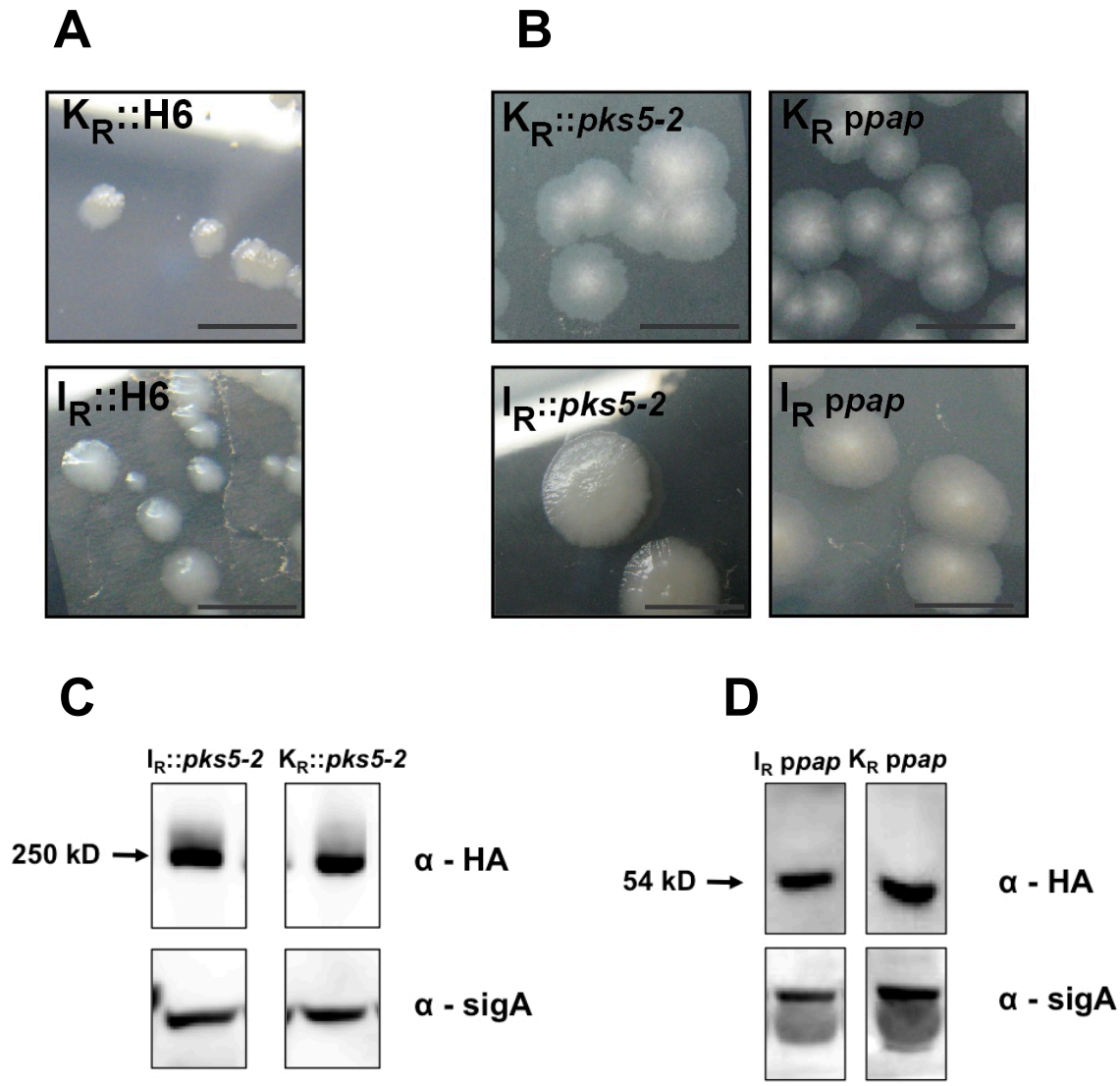
Supplementary Figure 2. Domain organization of the different *pks5* genes.
(**A**) Comparison of the different domains of *pks5-1* (left panel) and *pks5-2* (right panel) between *M. tuberculosis* H37Rv (middle), *M. canettii* strains A (CIPT 140010059) (top) and K_S (bottom).
(**B**) Potential recombination of *pks5* of *Mtb* H37Rv as compared to the two *pks5* genes of *M. canettii* strain A. Domains of *M. canettii* strain A were predicted according to the organization *of pks5* of *M. tuberculosis* H37Rv (Quadri et al., 2014) and the respective amino acid positions (AA) are depicted either above or below the particular domains. Sequences of individual domains of *pks5-1* and *pks5-2* were aligned using ClustalW2 and identity values are shown as percentage. Origin of the domains of the potentially recombined *pks5* of *Mtb* H37Rv is represented in different nuances of gray (light gray originating from *pks5-2* and dark gray from *pks5-1*). Domain abbreviations: ketosynthase (KS), acyltransferase (AT), dehydratase (DH), enoylreductase (ER), ketoreductase (KR) and acyl-carrier protein (ACP).


A
wbbL
STB-lv1_1760035
pks5-1
pap
pks5-2
papA
M. canettii I_S
M. canettii I_R
pks5-1
pap
pks5-2
AA position
430 431 534
100 % identical
212 213 214 289 425 426
I_S F S T V Y A A M T
I_R M T N A S V V F S
B
M. canettii I_S
M. canettii I_R
AA 21-443
AA 553-846
AA 903-1074
AA 1434-1745
AA 1769-1947
AA 2024-2102
AA 17-438
AA 548-841
AA 898-1065
AA 1429-1740
AA 1764-1942
AA 2019-2096
pks5-1
pks5-2
KS
AT
DH
ER
KR
ACP
99%
100%
1
2
3
STB-I-S CATTGCGCTAGCCGGCGGTGCTTCGGTGATGCTCGAA
STB-I-R CATTGCGCTAGCCGGCGGTGCTTCGGTGATGCTCGAA
STB-I-S CACCGTCAATATTGTGACACCATCACGGTCCGCGCAG
STB-I-R CACCGTCAATATTGTGACACCATCACGGTCCGCGCAG
STB-I-S GGTGTCATCCTACGGGTTCTCGGGCACCAATGTGCAC
STB-I-R GGTGTCATCGTATGGCATGACGGGCACCAATGTGCAC

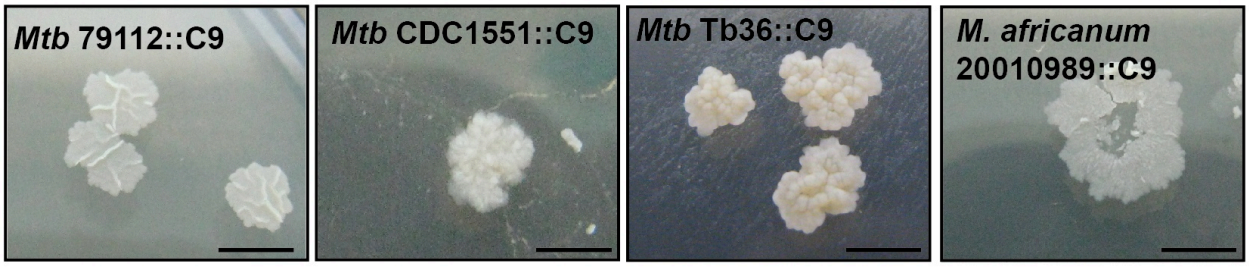
Supplementary Figure 3. (**A**) Genetic locus of *pks5* genes in *M. canettii* strains I_S and I_R and amino acid (AA) positions of non-synonymous SNPs occurring between the smooth and the rough morphotype in the two genes *pks5-1* and *pks5-2*. No mutations were found in the *pap* gene. Amino acid sequences were compared by using MEGA5 software (Tamura, et al., 2011).
(**B**) Comparison of *pks5-1* and *pks5-2* domains of *M. canettii*-I_S and *M. canettii*-I_R. Note that mutations were only located in the ketosynthase (KS) domains and were attributed to probable recombination events of sub-fractions of the *pks5* genes. The three mutated regions are highlighted in red and probable genetic exchange is indicated by black arrows.



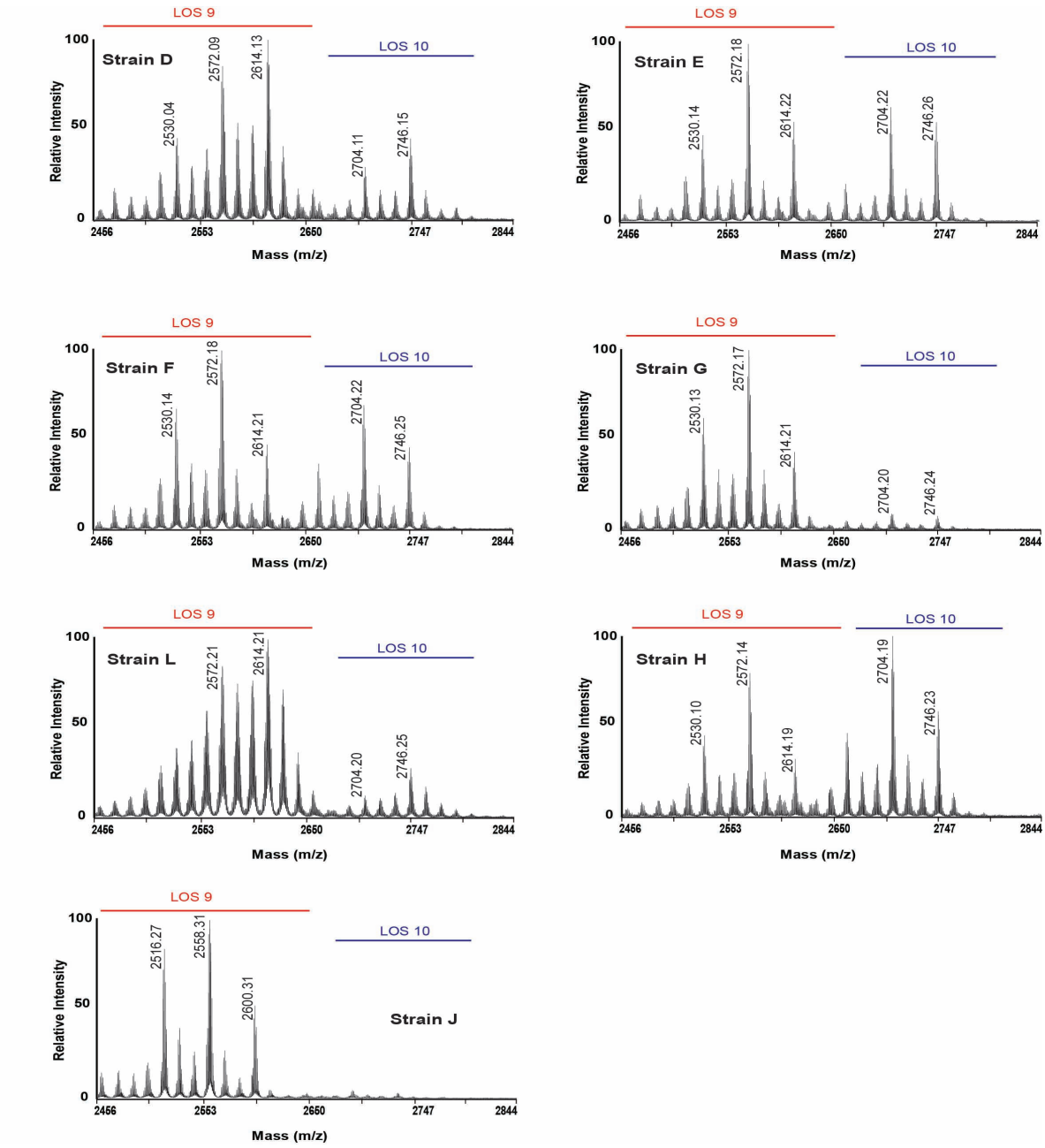
Supplementary Figure 4. Map of the pYUB412-based cosmids C9 and H6 used to complement *M. canettii* K_R or other tubercle bacilli with rough morphotypes. The genetic regions of *M. canettii* strain A (CIPT 140010059), cloned into the backbone of the integrating cosmid pYUB412 is shown on top. Arrows indicate insert position on the cosmid. The right panel shows the site of integration (*attB*) of the cosmid in mycobacterial genomes (within glyV-tRNA gene) (Bange et al., 1999).



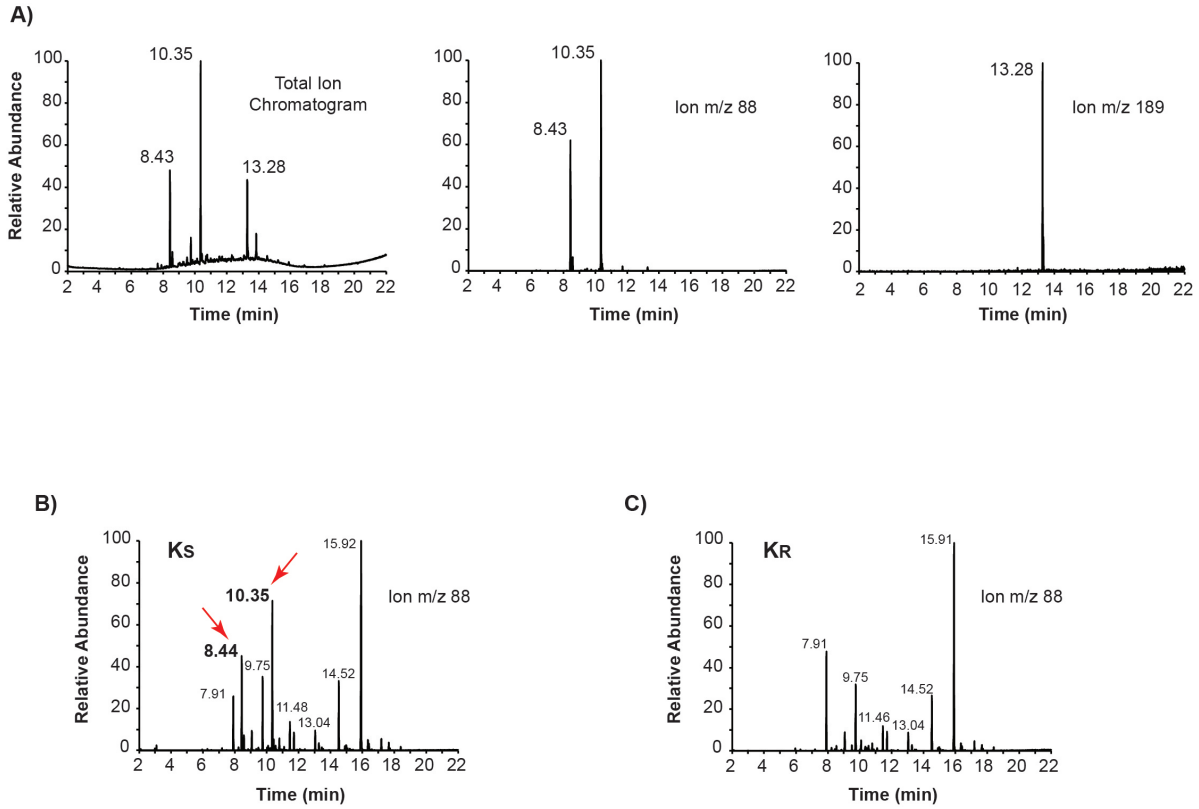
Supplementary Figure 5. Complementation of *M. canettii* strains I_R and K_R with cosmid H6, or HA-tagged *pks5-2* and *pap*. (**A**) Colony morphologies of *M. canettii* I_R::H6 and K_R::H6. Scale bar = 5 mm. (**B**) Colony morphologies of strains K_R and I_R complemented with either *pks5-2* or *pap.* Scale bar = 5 mm. (**C**) Expression of HA-tagged *pks5-2* or *pap* in *M. canettii* I_R and K_R was confirmed by Western Blot analysis using anti-HA antibodies. Expression of SigA was used as loading control.



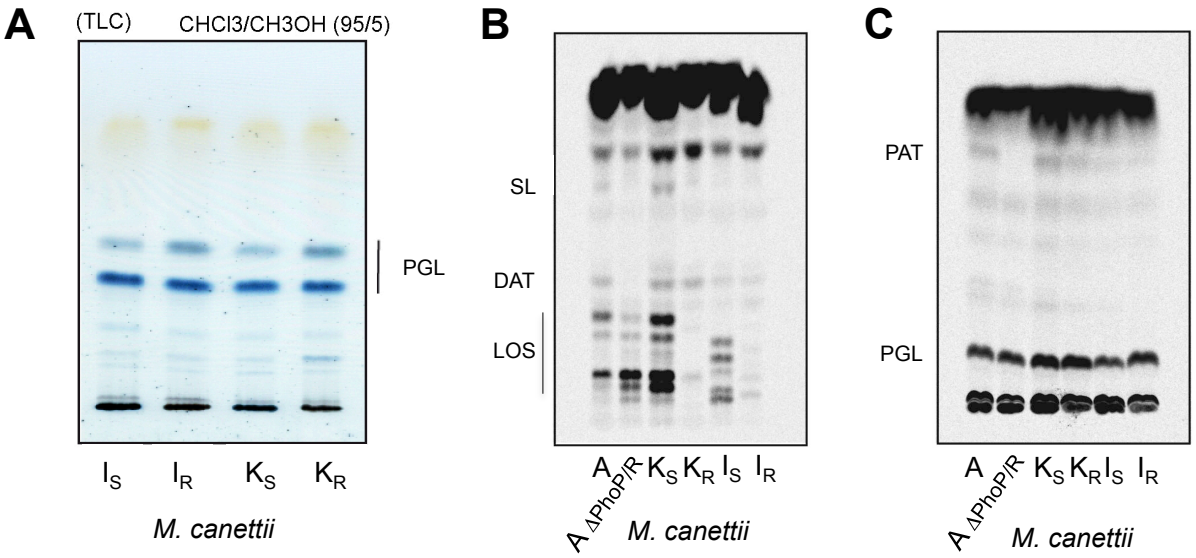
Supplementary Figure 6. Colony morphotypes observed after complementation attempts with integrating cosmid C9 in different members of the *M. tuberculosis* complex, including *M. tuberculosis* (*Mtb*) 79112 (lineage 1), *M. tuberculosis* CDC1551 (lineage 4), *M. tuberculosis* Tb36 (lineage 1) and *M. africanum* 20010989 (lineage 5/6). Scale bar = 5 mm.



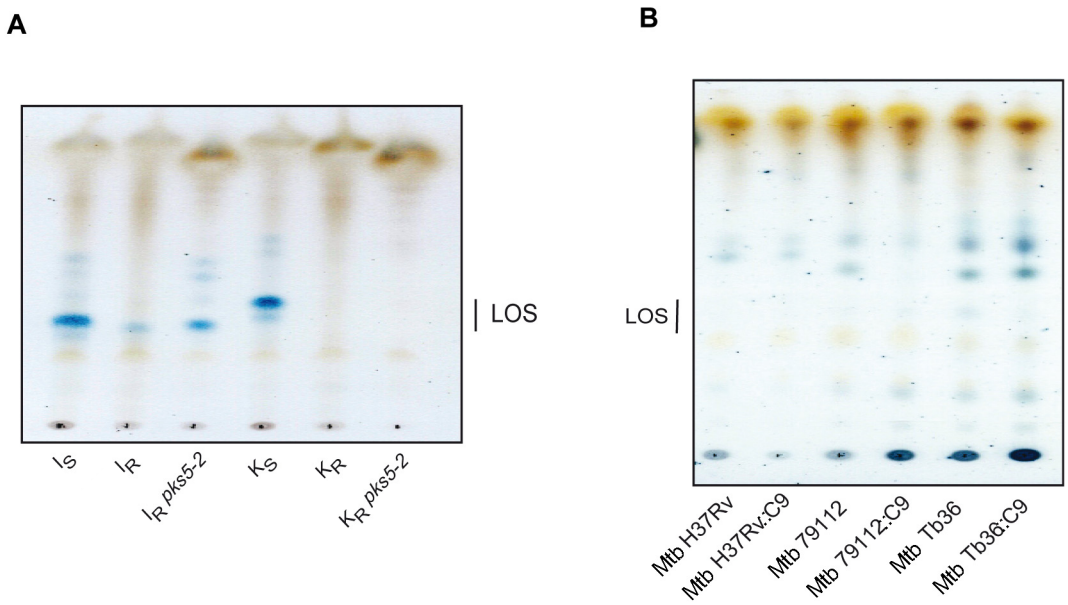
Supplementary Figure 7. MALDI-TOF mass spectra of purified LOS from *M. canettii* strains D, E, F, G, L, H, and J. The analysis of the mass spectra of the LOS-like compounds purified from the various strains shows a series of pseudo-molecular ions peaks [M+Na]+, between m/z 2450 and 2800, similar to that of the previously characterized LOS from *M. canettii* strain A (CIPT 140010059). The mass values of the major forms of the glycoconjugates from *M. canettii* strains G, J, K, and L correspond to that of the LOS 9 from *M. canettii* A. In strains D, E, F, and H, additional mass peaks were observed. Based on their mass values, these compounds may correspond to LOS containing 10 sugar residues (LOS 10), *e.g* the nine sugar units of LOS 9 plus an additional pentosyl residue. Note that LOS from strain J shows the same 14 mass-units shift as that observed for LOS from strain I (Fig. 4). This difference is predicted to be caused in these strains by the absence of a methyl residue on one 2-O-Me-rhamnosyl unit, probably due to a missing or non-functional methyltransferase.



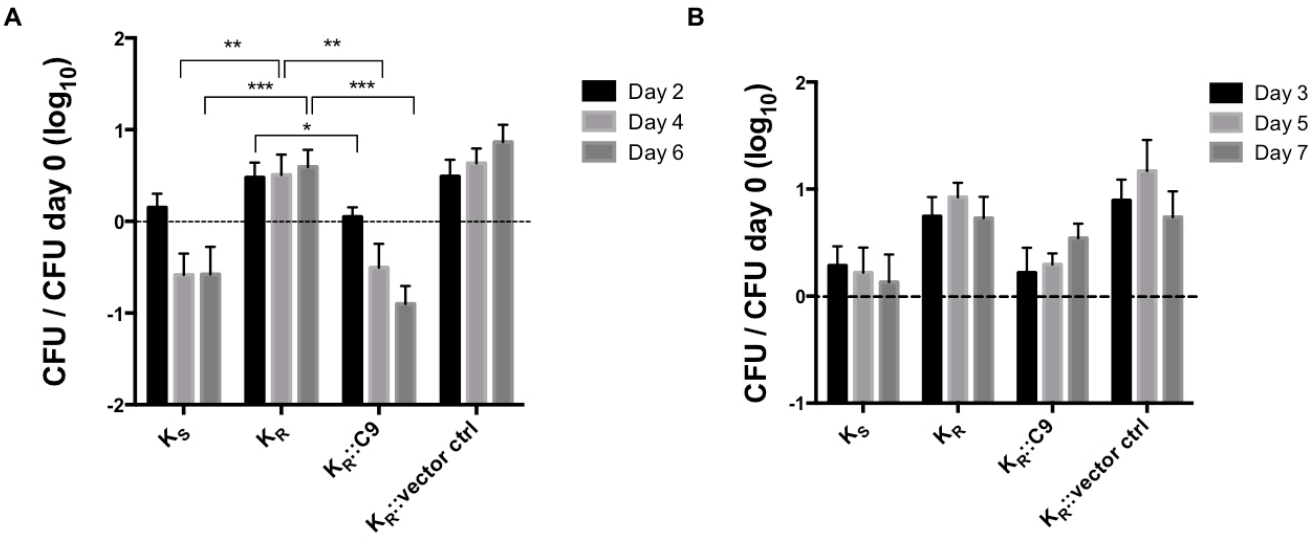
Supplementary Figure 8. GC-MS analysis of the LOS-specific fatty acids produced by strains K_S and K_R. (**A**) Fatty acid composition of LOS purified from strain K_S by GC-MS. The chromatogram of the released fatty acid methyl esters shows the presence of three major peaks with retention times (t_R) of 8.43, 10.35,and 13.28 min. Electron impact EI-MS and the McLafferty rearranged ions of these peaks are consistent with the 2L-,4L-dimethylhexadecanoate (t_R=8.43min), 2L-,4L-,6L-,8L tetramethyloctadecanoate (t_R=10.35min) and 2-methyl-3-hydroxyeicosanoate (t_R=13.28min), the acyl substituents of the LOS as reported earlier (Daffé et al., 1991). Note that fragment ions at *m/z* 88 and *m/z* 189, which indicate the presence of 2-Me and 3-OH-2-Me/TMS derivatives in the purified LOS, were selected as base peaks to detect the presence or not of fatty acid methyl esters released from the total extractable lipids from the S- and R-variants.
(**B**) and (**C**) Identification of LOS-specific fatty acids from crude lipid extracts from strains K_S and K_R that were first released from complex lipids by alkaline methanolysis and converted to trimethylsilyl derivatives before being subjected to GC-MS analysis. The chromatograms show fatty acid methyl esters from strain K_S (**B**) and K_R (**C**) that gave fragmentation ion at *m/z* 88. Note that 2L-,4L-dimethylhexadecanoate (t_R=8.43min) and 2L-,4L-,6L-,8L-tetramethyloctadecanoate (t_R=10.35min) were detected in strain K_S, but not in strain K_R.



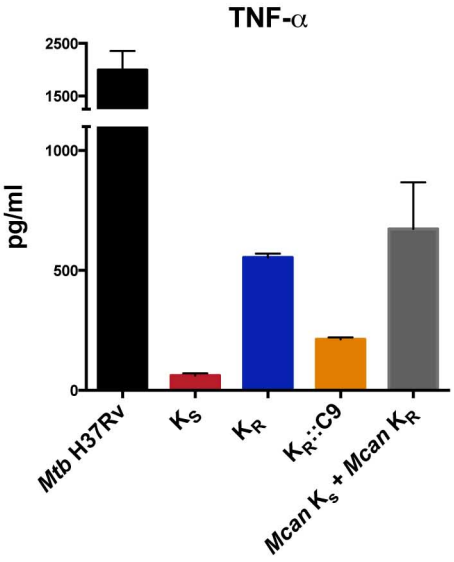
Supplementary Figure 9. TLC analyses of lipid extracts from S- and R-variants of *M. canettii* strains. (**A**) TLC analysis of lipid extracts from *M. canettii* strains I_R, I_S, K_R and K_S, which were spotted on silica gel G60 plates, separated with $CHCl_3/CH_3OH$ (95/5) and visualized by spraying with 0.2% anthrone in concentrated H_2SO_4, followed by heating. Phenolic glycolipids (PGL) are indicated. This experiment was performed once.
(**B**) and (**C**) TLC analysis of various lipids, including acyl-trehaloses such diacyltrehaloses (DAT), polyacyltrehaloses (PAT), sulfolipids (SL), phenolic glycolipids (PGL) or lipooligosaccharides (LOS) from different *M. canettii* strains using metabolic labelling with ^{14}C propionate. Note that under the growth conditions used, very low production of SL and DAT/PAT was observed for K_S and I_S strains and no accumulation of these compounds is seen in the K_R and I_R variants. In contrast, clear differences between S- and R-variants can be seen for production of LOS. Strain *M. canettii* A ΔPhoP/R was used as control as this strain does not produce DAT/PAT and SL due to PhoP/R-mediated downregulation of *pks3/4* and *pks2* genes (Gonzalo-Asensio J et al., 2014). This experiment was performed once.



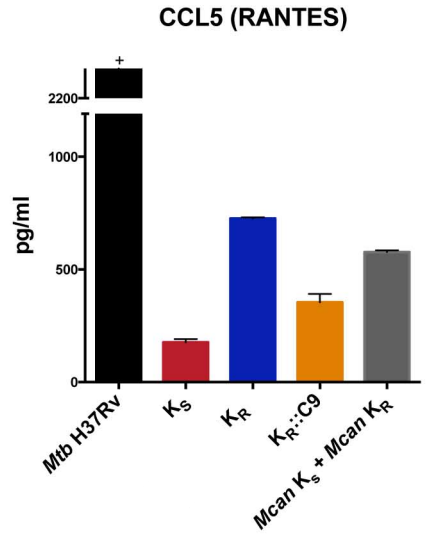
Supplementary Figure 10. Complementation of LOS production in various strains. (**A**) TLC analysis ($CHCl_3$:CH_3OH:H_2O, 60:24:2) of lipid extracts of *M. canettii* strains K_R and I_R complemented with the *pks5-2* gene. (**B**) TLC analysis ($CHCl_3$:CH_3OH:H_2O, 60:24:2) of lipid extracts of various MTBC strains complemented with the C9 cosmid. LOS: lipooligosaccharides. Glycolipids were visualized by spraying with anthrone, followed by charring. This experiment was performed once.

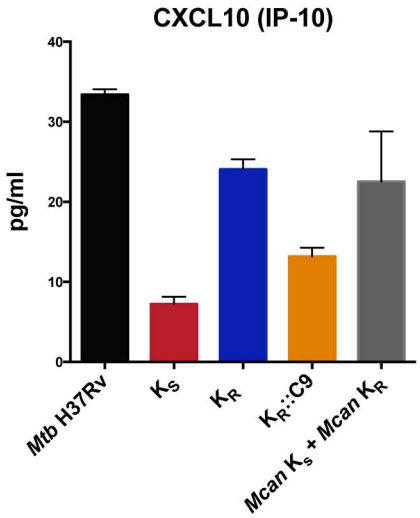


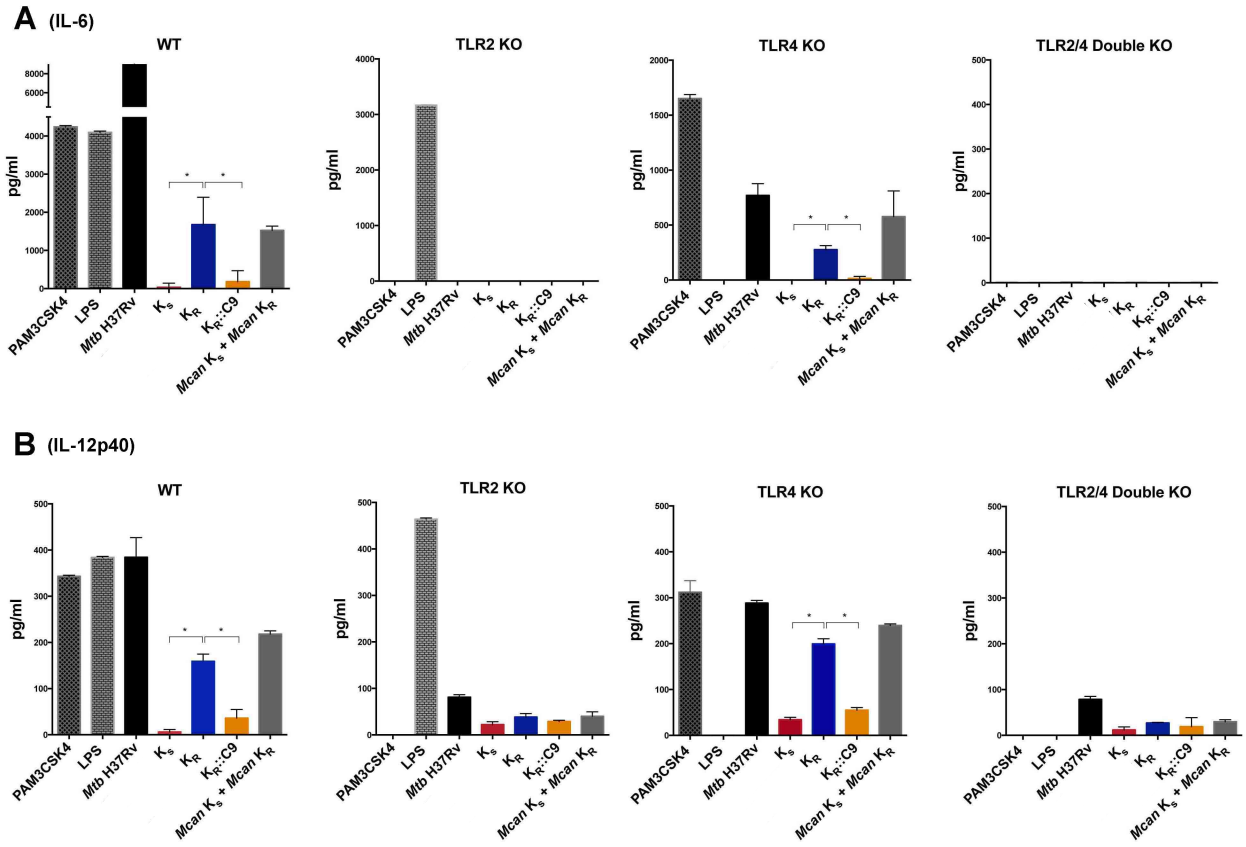
Supplementary Figure 11. Intra-macrophagal behaviour of *M. canettii* S- and R-variants in (**A**) human monocyte-derived macrophages and (**B**) in murine Raw macrophages. Results shown correspond to experiments in which macrophages (7.5 x 10^4 cells per well) were infected with various strains at an MOI of 0.05 (~ 1 bacterium per 20 cells). The Figure shows fold growth rates of intracellular bacteria determined at 3 h and at the indicated timepoints post infection. Data are represented as means and standard deviation of three independent experiments. Significance in difference between strains K_R and K_S as well as K_R and K_R::C9 was determined using Two-way *ANOVA* (*$p < 0.05$, **$p < 0.001$, ***$p < 0.0001$).



Supplementary Figure 12. Cytokine and chemokine production in BM-DCs upon infection with different morphotypes. BM-DCs of C57BL/6 WT mice were infected with Sauton-grown *M. tuberculosis (Mtb)* H37Rv, *M. canettii strains* K_S, K_R or K_R::C9 at an MOI of 1. 24 h post infection the level of indicated cytokines and chemokines in the cell supernatants was determined by Luminex Muliplex Assay. Data represent mean and standard deviation of 2 independent biological replicates each done in duplicate.
+ indicates that values were above the detectable range limit. *Mcan = M. canettii*







Supplementary Figure 13. IL-6 production and IL-12p40 in BM-DCs of WT and TLR-KO mice upon infection with different morphotypes. BM-DCs of C57BL/6 WT, TLR2 KO, TLR4 KO or TLR2/4 KO mice were infected with Sauton-grown *M. tuberculosis* H37Rv, *M. can*ettii K_S, K_R or K_R::C9 at an MOI of 1. 24 h post infection, the level of IL-6 or IL-12p40 in the cell supernatants was determined by ELISA. PAM3CSK4 (10 µg/ml) was used as a positive control for TLR2 and LPS (100 ng/ml) for TLR4 stimulation. Viability of TLR2/4 double KO cells was comparable to the other cell lines as determined by observations by microscope. Data are represented as means and standard deviation of three biological replicates. Significance in difference was determined using Mann Whitney test (*$p < 0.05$).

Supplementary Table 1: Putative SNPs suggested by mapping of Illumina-generated sequence reads relative to the *M. canettii* strain I (CIPT 140070007 or STB-I)(Supply et al., 2013) using the parameters of an inhouse SNP analysis pipeline described in Material and Methods.

Please note that results obtained by complementary mapping and SNP analyses, using the same set of reads but pre-calibrated parameters for re-sequencing data of reference genomes (Pouseele & Supply,, 2015), some of the listed SNPs were not confirmed and might correspond to be read-mapping artefacts caused by the high GC content and the repetitive nature of the genes involved.

Locus Tag	Ref Base	Base Position	Variant	Mutation	Ref Amino-Acid	Amino-Acid Position	Substitution	Gene Product	mutation in S, R or both
STB-Iv1_800020	A	984468	C	non-synonymous	*	77	Y	fragment of PROBABLE CONSERVED INTEGRAL MEMBRANE TRANSPORT PROTEIN (part 1)	both
STB-Iv1_860045	G	1094121	C	synonymous				Uncharacterized PE-PGRS family protein PE_PGRS33 (fragment)	both
STB-Iv1_890003	G	1108961	C	synonymous				fragment of PROBABLE CONSERVED INTEGRAL MEMBRANE PROTEIN (part 1)	both
STB-Iv1_1170007	G	1253492	C	non-synonymous	G	427	A	Uncharacterized PE-PGRS family protein PE_PGRS20	both
STB-Iv1_1340012	C	1369460	T	synonymous				PROBABLE PYRROLINE-5-CARBOXYLATE DEHYDROGENASE ROCA	S
STB-Iv1_1900001	A	1834669	G	non-synonymous	V	146	A	polyketide synthase Pks5	R
STB-Iv1_1900001	G	1834668	A	non-synonymous	V	146	A	polyketide synthase Pks5	R
STB-Iv1_1900001	T	1834666	G	non-synonymous	Y	147	S	polyketide synthase Pks5	R
STB-Iv1_1900001	G	1834665	C	non-synonymous	Y	147	S	polyketide synthase Pks5	R
STB-Iv1_1900001	G	1834663	A	non-synonymous	A	148	V	polyketide synthase Pks5	R
STB-Iv1_1950022	G	1903561	C	non-synonymous	G	239	A	fragment of CONSERVED HYPOTHETICAL PROTEIN (part 1)	both
STB-Iv1_1950022	G	1903567	A	non-synonymous	G	241	D	fragment of CONSERVED HYPOTHETICAL PROTEIN (part 1)	both
STB-Iv1_2000016	G	1968979	A	synonymous				CONSERVED HYPOTHETICAL PROTEIN	R
STB-Iv1_2420001	G	2324496	T	non-synonymous	T	303	N	transposase	both
STB-Iv1_2970010	G	2747759	A	non-synonymous	A	132	V	PROBABLE OXYGEN-INDEPENDENT COPROPORPHYRINOGEN III OXIDASE HEMN (COPROPORPHYRINOGENASE) (COPROGEN OXIDASE)	R
STB-Iv1_3040002	G	2863651	C	non-synonymous	R	52	S	Ice-structuring protein 4 (fragment)	both
STB-Iv1_3050001	G	2864527	C	non-synonymous	P	4	A	PE-PGRS FAMILY PROTEIN (fragment)	both
STB-Iv1_4880028	G	4387574	C	non-synonymous	G	250	A	fragment of POSSIBLE CONSERVED MEMBRANE PROTEIN (part 1)	both
STB-Iv1_4920001	C	4404314	G	non-synonymous	V	443	L	POSSIBLE CONSERVED MEMBRANE PROTEIN	both
STB-Iv1_4930001	C	4427879	A	synonymous				CONSERVED HYPOTHETICAL ALANINE RICH PROTEIN	both

Supplementary Table 2: Putative SNPs suggested by mapping of Illumina-generated sequence reads relative to the *M. canettii* strain K (CIPT 140070010 or STB-K, accession number NC_019951.1) (Supply et al., 2013) using the parameters of an inhouse SNP analysis pipeline described in Material and Methods.

Please note that results obtained by complementary mapping and SNP analyses, using the same set of reads but pre-calibrated parameters for re-sequencing data of reference genomes (Pouseele & Supply, 2015), most of the listed SNPs were not confirmed and might correspond to be read-mapping artefacts caused by the high GC content and the repetitive nature of the genes involved.

Locus Tag	Ref Base	Base Position	Variant	Mutation	Ref Amino-Acid	Amino-Acid Position	Substitution	Gene Product	mutation in S, R or both
BN42v3_CDS248358R	T	249658	G	non-synonymous	N	606	T	Putative zinc metalloprotease (part 1)	both
BN42v3_CDS248358R	T	249655	G	non-synonymous	Q	607	P	Putative zinc metalloprotease (part 1)	both
BN42v3_10242	T	249658	G	non-synonymous	K	3	N	Putative zinc metalloprotease (part 2))	both
BN42v3_10242	T	249655	G	synonymous				Putative zinc metalloprotease (part 2))	both
BN42v3_10330	G	353091	C	non-synonymous	P	28	A	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	C	353065	A	non-synonymous	R	36	S	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	T	353056	C	synonymous				Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	T	353055	C	non-synonymous	K	40	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	T	353054	C	non-synonymous	K	40	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	T	353053	C	non-synonymous	K	40	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	T	353052	C	non-synonymous	T	41	A	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	G	353039	C	non-synonymous	A	45	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_10330	G	353035	C	synonymous				Conserved protein of unknown function, PE-PGRS family protein PE_PGRS3	both
BN42v3_20003	A	354699	C	synonymous				Uncharacterized PE-PGRS family protein PE_PGRS3	both
BN42v3_20003	A	354693	C	synonymous				Uncharacterized PE-PGRS family protein PE_PGRS3	both
BN42v3_20003	C	354686	G	non-synonymous	A	162	P	Uncharacterized PE-PGRS family protein PE_PGRS3	both
BN42v3_20005	T	355948	G	non-synonymous	E	390	A	WAG22 antigen (fragment)	R
BN42v3_20005	C	355947	G	non-synonymous	E	390	A	WAG22 antigen (fragment)	R
BN42v3_20005	T	355873	G	non-synonymous	K	415	T	WAG22 antigen (fragment)	both
BN42v3_20005	C	355872	G	non-synonymous	K	415	T	WAG22 antigen (fragment)	both
BN42v3_20079	A	449413	G	synonymous				Conserved protein of unknown function, PPE6/PPE7 family protein	S
BN42v3_20079	G	449411	A	non-synonymous	S	868	F	Conserved protein of unknown function, PPE6/PPE7 family protein	S
BN42v3_20270	A	639536	G	synonymous				Conserved protein of unknown function, uncharacterized PE-PGRS family protein PE_PGRS33	both
BN42v3_20270	A	639537	G	non-synonymous	T	542	A	Conserved protein of unknown function, uncharacterized PE-PGRS family protein PE_PGRS33	R
BN42v3_20324	G	694345	T	non-synonymous	T	862	N	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS54	R
BN42v3_20602	C	942403	G	synonymous				Conserved protein of unknown function, PE-PGRS family protein (part2)	R
BN42v3_20603	A	942574	C	non-synonymous	L	490	V	Conserved protein of unknown function, PE-PGRS family protein (part1)	R
BN42v3_20603	C	942572	A	non-synonymous	L	490	V	Conserved protein of unknown function, PE-PGRS family protein (part1)	R
BN42v3_20603	C	942403	G	non-synonymous	E	547	Q	Conserved protein of unknown function, PE-PGRS family protein (part1)	R
BN42v3_20695	T	1034001	C	non-synonymous	M	481	T	Conserved protein of unknown function, PPE family (fragment)	R
BN42v3_20695	T	1034010	C	non-synonymous	M	484	T	Conserved protein of unknown function, PPE family (fragment)	R
BN42v3_20914	G	1252332	A	non-synonymous	L	222	F	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	G	1252309	A	synonymous				Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	A	1252307	G	non-synonymous	V	230	A	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	A	1252300	C	synonymous				Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	C	1252299	A	non-synonymous	D	233	S	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	T	1252298	G	non-synonymous	D	233	S	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	T	1252281	C	non-synonymous	N	239	A	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	T	1252280	G	non-synonymous	N	239	A	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	A	1252276	G	synonymous				Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	C	1252275	T	non-synonymous	A	241	I	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	G	1252274	A	non-synonymous	A	241	I	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	G	1252268	C	non-synonymous	A	243	G	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	G	1252267	C	non-synonymous	A	243	G	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	T	1252261	C	synonymous				Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_20914	C	1252260	T	non-synonymous	A	246	T	Conserved protein of unknown function, PE-PGRS family protein (PE-PGRS 20)	R
BN42v3_21367	G	1708360	C	synonymous				Conserved protein of unknown function, PE-PGRS family protein PE_PGRS54 (part1)	both
BN42v3_21367	A	1708348	G	synonymous				Conserved protein of unknown function, PE-PGRS family protein PE_PGRS54 (part1)	R
BN42v3_21371	G	1710804	C	non-synonymous	A	250	G	Conserved protein of unknown function, PE_PGRS family protein (part3)	both
BN42v3_21371	C	1710803	G	non-synonymous	A	250	G	Conserved protein of unknown function, PE_PGRS family protein (part3)	S
BN42v3_21372	T	1711684	A	synonymous				Conserved protein of unknown function, PE-PGRS family protein PE_PGRS54 (part2)	both
BN42v3_21372	C	1711682	A	non-synonymous	G	153	V	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS54 (part2)	both
BN42v3_21457	A	1806180	G	synonymous				Putative polyketide synthase Pks5	R
BN42v3_21457	A	1806178	G	non-synonymous	V	268	A	Putative polyketide synthase Pks5	R
BN42v3_40444	A	2925449	G	synonymous				Conserved protein of unknown function, PE-PGRS family protein PE_PGRS54	both
BN42v3_40840	G	3286119	A	non-synonymous	D	461	N	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS48 (part1)	both
BN42v3_40840	C	3286129	G	non-synonymous	A	464	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS48 (part1)	both
BN42v3_40840	G	3286130	C	non-synonymous	A	464	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS48 (part1)	both
BN42v3_40841	G	3286119	A	synonymous				Conserved protein of unknown function, PE-PGRS family protein PE_PGRS48 (part2)	both
BN42v3_40841	C	3286129	G	non-synonymous	R	17	A	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS48 (part2)	both
BN42v3_40841	G	3286130	C	non-synonymous	R	17	A	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS48 (part2)	both
BN42v3_40893	T	3332127	C	synonymous				Conserved protein of unknown function (part1)	both
BN42v3_40893	T	3332125	C	non-synonymous	E	96	G	Conserved protein of unknown function (part1)	both
BN42v3_41048	A	3526871	C	non-synonymous	F	6	L	Putative cytochrome C oxidase polypeptide I CtaD (cytochrome AA3 subunit 1) Putative cytochrome C oxidase polypeptide I CtaD (cytochrome AA3 subunit 1) (part2)	both
BN42v3_41048	A	3526868	C	synonymous				Putative cytochrome C oxidase polypeptide I CtaD (cytochrome AA3 subunit 1) Putative cytochrome C oxidase polypeptide I CtaD (cytochrome AA3 subunit 1) (part2)	both
BN42v3_41049	A	3526871	C	non-synonymous	L	260	W	Putative cytochrome C oxidase polypeptide I CtaD (cytochrome AA3 subunit 1) (part1)	both
BN42v3_41049	A	3526868	C	non-synonymous	V	261	G	Putative cytochrome C oxidase polypeptide I CtaD (cytochrome AA3 subunit 1) (part1)	both
BN42v3_50005	T	3881715	G	synonymous				Conserved protein of unknown function, PPE family protein PPE54 (fragment)	R
BN42v3_90104	T	4159167	A	synonymous				Conserved protein of unknown function, PE-PGRS family protein	both
BN42v3_90104	A	4159170	C	non-synonymous	E	518	D	Conserved protein of unknown function, PE-PGRS family protein	both
BN42v3_90104	A	4159171	G	non-synonymous	R	519	G	Conserved protein of unknown function, PE-PGRS family protein	both
BN42v3_90104	G	4159173	C	non-synonymous	R	519	G	Conserved protein of unknown function, PE-PGRS family protein	both
BN42v3_90153	A	4205115	C	synonymous				Conserved protein of unknown function, PE-PGRS family protein	both
BN42v3_90159	T	4210087	G	non-synonymous	*	143	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS33 (fragment)	S
BN42v3_90159	A	4210088	G	non-synonymous	*	143	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS33 (fragment)	S
BN42v3_90159	A	4210089	G	non-synonymous	*	143	G	Conserved protein of unknown function, PE-PGRS family protein PE_PGRS33 (fragment)	S
BN42v3_90160	T	4210086	G	non-synonymous	L	8	R	Conserved protein of unknown function, PE-PGRS family protein (fragment)	S
BN42v3_90160	T	4210087	G	non-synonymous	L	8	R	Conserved protein of unknown function, PE-PGRS family protein (fragment)	S
BN42v3_90160	A	4210088	G	non-synonymous	K	9	G	Conserved protein of unknown function, PE-PGRS family protein (fragment)	S
BN42v3_90160	A	4210089	G	non-synonymous	K	9	G	Conserved protein of unknown function, PE-PGRS family protein (fragment)	S
BN42v3_90160	A	4210090	G	non-synonymous	K	9	G	Conserved protein of unknown function, PE-PGRS family protein (fragment)	S

Supplementary Table 3: Comparison of genes present in the LOS locus in different mycobacterial species

	M. marinum	*M. tuberculosis*	*M. canettii* A	*M. canettii* K	*M. kansasii*	*M. smegmatis*	gene function	phenotype of mutants
	mmar_2307	-	-	-	-	-	hypothetical transmembrane protein (van der Woude et al. 2012)	LOS-I accumulation in *M. marinum*
	mmar_2309	-	-	-	mkan_27380	-	*udgL*; UDP-glucose/GDP-mannose dehydrogenase family (Ren et al. 2007)	
	mmar_2310	-	-	-	mkan_27385	-	putative UDP-glucuronate decarboxylase	
	mmar_2311	-	-	-	mkan_27390	-	glycosyl transferase/ methyl transferase	
	mmar_2312	-	-	-	-	-	putative sugar epimerase	
	mmar_2313	Rv1500	MCAN_15191	BN42v3_21427	-		*losA*; glycosyl transferase (Burguière et al. 2005)	LOS-III accumulation in *M. marinum*
	mmar_2314	Rv1501	MCAN_15201	BN42v3_21428	-	-	hypothetical protein (Rombouts et al. 2009)	
M. marinum specific cluster	mmar_2315	-	-	-	-	-	hypothetical methyltransferase (Rombouts et al. 2009)	
	mmar_2316	-	-	-	-	-	transcriptional regulator (Rombouts et al. 2009)	
	mmar_2317	-	-	-	-	-	hypothetical O-methyltransferase (Rombouts et al. 2009)	
	mmar_2319	-	-	-	-	-	hypothetical transmembrane protein (van der Woude et al. 2012)	
	mmar_2320	Rv1504c/Rv1503c	MCAN_15221/ MCAN_15231	BN42v3_21430	-	-	*wecE*; pyridoxal phosphate-dependent enzyme (van der Woude et al. 2012)	
	mmar_2321	Rv1505c	MCAN_15241	BN42v3_21431	-	-	hypothetical acyltransferase (Alibaud et al. 2013)	
	mmar_2322	Rv1506c	MCAN_15251	BN42v3_21432	-	-	hypothetical protein	
	mmar_2325	Rv1507c	MCAN_15261	BN42v3_21433	-	-	hypothetical protein	
	mmar_2327	Rv1508c	MCAN_15281	BN42v3_21436	-	-	hypothetical di- and tri-carboxylate transporter (van der Woude et al. 2012)	LOS-II and LOS-II* accumulation in *M. marinum*
	mmar_2328	-	-	-	-	-	carbohydrate kinase	
	mmar_2330	-	-	-	-	-	short-chain dehydrogenase	
	mmar_2331	-	-	-	-	-	hypothetical protein, potentially involved in synthesis of caryophyllose (Alibaud et al. 2013)	
	mmar_2332	-	-	-	-	-	*ilvB1_3*; acetolactate synthase (Ren et al. 2007)	
M. marinum specific cluster	mmar_2333	-	-	-	-	-	*wcaA*; glycosyl transferase (Sarkar et al. 2011)	
	mmar_2334	-	-	-	-	-	nucleotidyltransferase	
	mmar_2336	-	-	-	-	-	*galE6*; UDP-glucose 4-epimerase (van der Woude et al. 2012)	
	mmar_2337	-	-	-	-	-	GtrA-like membrane protein; putative role in synthesis and assembly of LOS	
	-	-	-	-	mkan_27425	-	glycosyl transferase	
	-	-	-	-	mkan_27430	-	NAD dependent epimerase/dehydratase family	
M. kansasii specific cluster	-	-	-	-	mkan_27435	-	glycosyl transferase (Nataraj et al. 2015)	no higher LOS (LOS IV-VII) in *M. kansasii*
	-	-	-	-	mkan_27440	-	Glucose-1-phosphate cytidiltransferase	
	-	-	-	-	mkan_27445	-	NAD dependent epimerase/dehydratase family	
	-	-	-	-	mkan_27450	-	Rhamnose epimerase	
	-	-	-	-	mkan_27475	-	methyltransferase	
	-	-	-	-	mkan_27695	-	glycosyl transferase	
	mmar_2339	-	-	-	mkan_27480	-	methyltransferase	
	-	Rv1511	MCAN_15321	BN42v3_21441	MKAN_23850 MKAN_23855 (outside LOS cluster)	-	*amdA*: GDP-D-mannose dehydratase	
	-	Rv1512	MCAN_15331	BN42v3_21442		-	*epiA*; putative nucleotide-sugar epimerase	
	-	Rv1513	MCAN_15341	BN42v3_21443	-	-	hypothetical protein	
MTBC specific cluster	-	Rv1514c	MCAN_15351	BN42v3_21444	-	-	putative glycosyl transferase	
	-	Rv1515c	MCAN_15361	BN42v3_21445	-	-	putative methyl transferase	
	-	Rv1516c	MCAN_15371	BN42v3_21446	-	-	putative sugar transferase	
	-	Rv1518	MCAN_15391	BN42v3_21448	-	-	putative glycosyl transferase	
	-	Rv1520	MCAN_15411	BN42v3_21450	-	-	putative sugar transferase	
	-	Rv1523	MCAN_15441	BN42v3_21453	-	-	putative methyltransferase	
	mmar_2340	Rv1527c	MCAN_15481	**pks5-1 (BN42v3_21457)**	mkan_27485	msmeg_4727	*pks5*; polyketide synthase (van der Woude et al. 2012)	LOS-deficient or accumulation of LOS-0 in *M. marinum*
	mmar_2341[1]	Rv1521	MCAN_15421	BN42v3_21451	mkan_27490	msmeg_4731	*fadD25*; fatty acyl AMP ligase (van der Woude et al. 2012)	
	mmar_2341[1]	Rv1529	MCAN_15521	BN42v3_21460	mkan_27490	msmeg_4731	*fadD24*; fatty-acid-AMP ligase	
	mmar_2342	Rv1522c	MCAN_15431	BN42v3_21452	mkan_27530	msmeg_4741	MmpL family transport protein	
	mmar_2343	-	MCAN_15491	**pap**	mkan_27535	msmeg_4728	*papA*; polyketide synthase-associated protein (*papA4* in *M. marinum*) (Rombouts et al. 2011)	
	mmar_2344	-	MCAN_15501	**pks5-2**	mkan_27540	-	*pks5.1*; polyketide synthase	
	-	Rv1528c (truncated ?)	MCAN_15511	BN42v3_21458 + BN42v3_21459	-	-	*papA4*; MTBC version (truncated ?)	
conserved core cluster for first steps in LOS biosynthesis	-	Rv1530	MCAN_15531	BN42v3_21461	mkan_27545 (fragment)	-	*adh*; alcohol dehydrogenase	
	mmar_2345	Rv1531	MCAN_15541	BN42v3_21462	mkan_27550	-	hypothetical protein	
	-	Rv1532c	MCAN_15551	BN42v3_21463	mkan_27555	-	hypothetical protein	
	mmar_2349	Rv1525	MCAN_15461	BN42v3_21455	mkan_27575	-	*wbbL2*; Rhamnosyltransferase (Alibaud et al. 2013)	
	mmar_2350	Rv1531	MCAN_15541	BN42v3_21462	-	-	methylase	
	mmar_2351	-	-	-	mkan_27580	-	glycosyl transferase	
	mmar_2352	Rv1517	MCAN_15381	BN42v3_21447	-	msmeg_4733	hypothetical transmembrane protein	
	mmar_2353[2]	Rv1524	MCAN_15451	BN42v3_21454	mkan_27600	msmeg_4740	putative glycosyl transferase	
	mmar_2353[2]	Rv1526c	MCAN_15471	BN42v3_21456	mkan_27600	msmeg_4740	glycosyl transferase (van der Woude et al. 2012)	
	mmar_2354	-	-	-	-	msmeg_4729 / msmeg_4730	hypothetical protein	
	mmar_2355	-	-	-	mkan_27610	msmeg_4728 (see papA4)	*papA3*; acyl transferase (van der Woude et al. 2012)	
	mmar_2366	Rv1543	MCAN_15661	BN42v3_21474	mkan_27675	-	fatty acyl co-A reductase	
	mmar_2367	Rv1544	MCAN_15671	BN42v3_21475	mkan_27680	msmeg_4722	keto acyl reductase	
	mmar_2370	Rv1549 + Rv1550	MCAN_15721	BN42v3_21481	mkan_27700	msmeg_4772	*fadD11*; fatty-acid-CoA ligase	
	mmar_2371	Rv1551	MCAN_15731	BN42v3_21482	mkan_27705	msmeg_4703	*plsB1*; acyltransferase	
	mmar_2405	-	-	-	-	-	*cphB*; cyanophycinase (van der Woude et al. 2012)	
	-	-	-	-	-	msmeg_4732	glycosyl transferase	
	-	-	-	-	-	msmeg_4734	hypothetical protein, PE-PPE like	
M. smegmatis specific cluster	-	-	-	-	-	msmeg_4735	possible glycosyl transferase	
	-	-	-	-	-	msmeg_4736	pyruvyltransferase	
	-	-	-	-	-	msmeg_4737	pyruvyltransferase	
	-	-	-	-	-	msmeg_4738	hypothetical protein	
	-	-	-	-	-	msmeg_4739	possible methyl transferase	
	mmar_5170	Rv3681c	MCAN_37021	BN42v3_90194	mkan_24700	msmeg_6199	*whiB4*; transcriptional regulator protein (van der Woude et al. 2012)	LOS diminished in *M. marinum*
	mmar_5437	Rv3862c	MCAN_38841	BN42v3_90396	mkan_27515	msmeg_0051	putative transcriptional regulator; whiB-like	

[1] note that mmar_2341 shows homology to Rv1521 and Rv1529
[2] note that mmar_2353 shows homology to Rv1524 and Rv1526c

Supplementary Table 4: Comparison of gene similarities present in the orthologous LOS locus in *M. canettii* vs *M. tuberculosis*

gene name in *M. tuberculosis*	% amino acid identity to homolog in *M. canettii* A (AA *Mcan* A / AA *Mtb*)	% amino acid identity to homolog in *M. canettii* K (AA *Mcan* K / AA *Mtb*)	gene function	phenotype of mutants	
Rv1500	100%	147 AA longer	glycosyl transferase LosA	**LOS-III accumulation in *M. marinum***	
Rv1501	100%	99% (272/274)	hypothetical protein (Rombouts et al. 2009)		
Rv1504c/Rv1503c	MCAN_15231/ MCAN_15221; 36 AA shorter	one polypeptide (as in *M. marinum*)	pyridoxal phosphate-dependent enzyme WecE (van der Woude et al. 2012)		
Rv1505c	100%	100%	hypothetical acyltransferase		
Rv1506c	100%	98% (164/167)	hypothetical protein		
Rv1507c	100%	99% (231/232)	hypothetical protein		
Rv1508c	99% (598/600)	97% (583/600)	hypothetical di- and tri-carboxylate transporter (van der Woude et al. 2012)	**LOS-II and LOS-II* accumulation in *M. marinum***	
Rv1511	99% (1022/1023)	99% (1022/1023)	GDP-D-mannose dehydratase GmdA		
Rv1512	99% (968/969)	99% (965/969)	putative nucleotide-sugar epimerase EpiA		
Rv1513	1 AA more at beginning	1 AA more at beginning	hypothetical protein		
Rv1514c	100%	97% (255/263)	putative glycosyl transferase		
Rv1515c	99% (297/299)	97% (290/299)	putative methyl transferase		
Rv1516c	100%	96% (325/337)	putative sugar transferase		
Rv1518	100%	99% (311/314; 1AA shorter)	putative glycosyl transferase		
Rv1520	16 AA longer	16 AA longer	putative sugar transferase		
Rv1521	100%	99% (579/584)	fatty acyl AMP ligase FadD25 (van der Woude et al. 2012)	**LOS-deficient in *M. marinum***	**region present on C9**
Rv1522c	99% (1146/1147)	99% (1132/1147)	MmpL family transport protein		
Rv1523	100%	99% (344/348)	putative methyltransferase		
Rv1524	99% (414/415)	99% (410/415)	putative glycosyl transferase		
Rv1525	100%	99% (260/262)	Rhamnosyltransferase WbbL2		
Rv1526c	99% (424/427)	99% (424/427)	glycosyl transferase (van der Woude et al. 2012)		
Rv1527c	97% (2042/2108; 4 AA longer)	94% (115/2108; 9 AA longer)	polyketide synthase Pks5 (van der Woude et al. 2012)		
-	MCAN_15491	**pap**	*papA* polyketide synthase-associated protein (*papA4* in *M. marinum*) (Rombouts et al. 2011)		
-	MCAN_15501	**pks5-2**	polyketide synthase Pks5.1		
Rv1528c	99% (164/166)	frameshift	*papA4;* short MTBC version		
Rv1529	99% (584/585)	99% (580/585)	fatty-acid-AMP ligase FadD24		
Rv1531	100%	100%	methylase		
Rv1543	100%	99% (341/342)	fatty acyl co-A reductase		
Rv1544	100%	99%(267/268)	keto acyl reductase		
Rv1549 + Rv1550	one single gene (frameshift)	one single gene (frameshift; 32 AA shorter in the beginning[1])	fatty-acid-CoA ligase FadD11		
Rv1551	99% (620/622)	99% (614/622)	acyltransferase PlsB1		
Rv3681c	100%	100%	transcriptional regulator protein WhiB4 (van der Woude et al. 2012)	**LOS diminished in *M. marinum***	

[1]same as in *M. marinum* FadD11

Supplementary Table 5: List of primers (oligos) used in this study

Strain/ Gene	Use	Sequence (5'-3')
pks5 locus	Amplification of *pks5* locus (long range PCR)	TTTATTAATCAGGGAAAGGCGACATCGGA TTTTTATAACCGCCAAGACAAACTTCATC
pks5-2 + *pap* (5')	Amplification of *pks5-2* (long range PCR)	TTTCAGGGAAAGGCGACATCGGA TTTCGCTACCAACGACTAGTAGTTCGTC
pap (3') + *pks5-1*	Amplification of *pks5-1* (long range PCR)	GACGAACTACTAGTCGTTGGTAGCG TTTTTATAACCGCCAAGACAAACTTCATC
pks5-2	Sequencing of *pks5-2*	GTTGTGGGAGGCGTTGCT CGAAGAACTCGGGATCAAAG GAAACGTCGAACGCATGAC GTCATGCGTTCGACGTTTC GCCACACCCGGTATCGAC GGTGGTGGCCTCCCCGCAGT AACGAGGTCGCCGAGTAGTA GACTGATCAACGCACCACTG ACTGCGAGATGGCGTTGGC CTCATCCCGCCGTCCAGGGC CGACGTGCTGGTCACCTT CGACCTTGAGTTCGCTGAC TGGCAGGGCGAGGTCGGCAC CATCGAACTCGTCCGCGCGA TGAGTTCGTCGGTGATGTTG CACTCTTCCTGTGTCAGCTC CTTGACTCACTGGGCAACCT GACCCTGCTGCGCCACAACC
pks5-1	Sequencing of *pks5-1*	GTTGTGGGAGGCGTTGCT CGAAGAACTCGGGATCAAAG GAAACGTCGAACGCATGAC GTCATGCGTTCGACGTTTC GCCACACCCGGTATCGAC GGTGGTGGCCTCCCCGCAGT AACGAGGTCGCCGAGTAGTA GACTGATCAACGCACCACTG TGTCAAACATGTGGTGCGTA ACTGCGAGATGGCGTTGGC GCTCGCAGGTCAAAGCTTAC CGACGTGCTGGTCACCTT CGACCTTGAGTTCGCTGAC TGAGTTCGTCGGTGATGTTG TAGACCTGGGGGTTGATGTCG TGGCAGGGCGAGGTCGGCAC ACGAACGGTGGTGTTGATTT GACCCTGCTGCGCCACAACC
pap	Sequencing of *pap*	ACCAGCCGTGAATAATCGAG AGCACAAAGTCTCGCCATTC CGTATAGCCCGGTGATCAAC CAGAACACCCGATGAGTACA GGCACATTTGCGAGGTCTAT
hygromycin	amplification of hygromycin cassette	ACAGGCCTGTCGTCGAGGTCCACCAA ACAGGCCTGGATGCCAGGGCCTTTCA
hsp60	amplification of *hsp60* promoter	aaaGCTAGCAAGCTTggtgaccacaacgacgcgcccgctttgatc aaaTCTAGAgatatcACTAGTtgtcttggccattgcgaagtgattcctcc
pks5-2-HA	HA-tagged *pks5-2*	aaaACTAGTGTGGGTAAGGAGAGAACAAAG aaaTTATAATtaAGCATAATCAGGAACATCATACGGATATGAAGGTGCTGCAATG TCGG
pks5-1-HA	HA- tagged *pks5-1*	aaaACTAGTGATGGCTGGGCTCCCGTGG aaaTTATAATtaAGCATAATCAGGAACATCATACGGATAggcgggtgccggtgcgtcc
pap-HA	HA-tagged *pap*	aaaACTAGTGTGATCATTGGCGGGGGC aaaTTATAATtaAGCATAATCAGGAACATCATACGGATAGCTAGATACGCGAACT GCTG.
pks5	probe for Southern Blot	GTTGTGGGAGGCGTTGCT GAAACGTCGAACGCATGAC
pap	probe for Southern Blot	CTCGATTATTCACGGCTGGT CGTATAGCCCGGTGATCAAC
C9 cosmid	verification of complemented strains (T7 side)	AGGCATGCAAGCTCAGGATA GGATCGGTCCAGTAATCGT
C9 cosmid	verification of complemented strains (T3 side)	GCAGAAGCACTAGACGATCC GCCGCAATTAACCCTCACTA

Supplementary Note

Whole genome sequence (WGS) analysis of *M. canettii* $I_{S/R}$ and $K_{S/R}$

We compared whole genome sequences and found R-specific differences that mapped to two genes of strain I, corresponding to a codon change in *hemN* and to 5 SNPs in the polyketide-synthase-encoding gene *pks5* (Supplementary Table 1). For strains $K_{S/R}$, we also found a non-synonymous SNP mapping to *pks5* of K_R and noted a two-fold higher depth of read coverage for *pks5* of K_S relative to its flanking regions, whereas this was not seen for strain K_R. Other putative SNPs mapped to genes encoding polymorphic GC-rich PE_PGRS proteins but these could not be confirmed and are likely to be read-mapping artefacts caused by the high GC content and the repetitive nature of these sequences (Cole et al., 1998), (Supplementary Table 2).

Polyketide synthase domain comparison

Sequences of *pks5-1* from *M. canettii* strains K_S and A (CIPT 140010059) showed 99% identity (ClustalW2), and sequences of *pks5-2* genes 97% identity, respectively. As shown in Fig. 1C, alignment of *pks5* of strain K_R with the two *pks5* of strain K_S revealed that the 5' region of *pks5* from the R variant was 100 % identical to *pks5-2* (bp 1 – 6056), while the 3' region was identical to *pks5-1* (bp 3533 – 6327), which suggested a recombination between the *pks5* genes in the R variant. Since Pks5 belongs to the family of polyketide synthases with multiple functional domains on one large polypeptide chain (Rousseau, C., et al., 2003), we investigated whether the recombination of the two *pks5* genes in *M. canettii* K_R possibly affected the function of one of the domains or whether the recombination happened in a non-functional region in between, leaving the domains potentially operational. The overall organization of the domains of *pks5* of strain K was the same as for *M. tuberculosis* H37Rv, as determined by sequence alignment with the predicted domains of *pks5* of *M. tuberculosis* and consists of a ketosynthase (KS), an acyltransferase (AT), a dehydratase (DH), an enoylreductase (ER), a ketoreductase (KR) and an acyl-carrier protein (ACP) (Fig. 1). The recombination in the rough strain took place in a sequence stretch of about 2523 bp, between basepair 3533 and 6056, comprising the ER and KR domains which were completely identical between the two *pks5* genes of STB-K_S. Consequently, the recombined *pks5* in the R variant of *M. canettii* K consisted of the KS, AT and DH domain of *pks5-2*, the ER and KR domain which were identical in both genes and the ACP domain of *pks5-1* (Fig. 1D). A similar event seemed to have happened in *M. tuberculosis* H37Rv, when comparing the *M. tuberculosis pks5* with the two copies in *M. canettii* STB-A (CIPT 140010059) whose independent domains generally showed higher sequence identity scores to those of *M. tuberculosis* than the domains of *M. canettii* strain K (Supplementary Fig. 2A). Recombination in *M. tuberculosis* possibly happened in a region of about 960 bp in between the AT and DH domains, resulting in a recombined *pks5* with KS and AT from *pks5-2* and the remaining four domains (DH, ER, KR and ACP) from *pks5-1* (Supplementary Fig. 2B). A BLAST comparison of *pks5* of *M. tuberculosis* H37Rv with available *pks5* sequences from various MTBC members showed more than 99% identity within the MTBC, suggesting that recombination in this particular locus might have happened in the last common ancestor of the MTBC strains after their separation from the *M. canettii* strains.

Supplementary Note (cont.)

Homologous recombination events in genomes of *M. tuberculosis* complex (MTBC) members

Homologous recombination in general plays a significant role in the evolution of the MTBC, due to abundant repetitive sequences and highly conserved gene paralogues. One such an example is the serine/threonine protein kinase-encoding *pknH* gene (*rv1266c)* of *M. tuberculosis,* which evolved from the recombination of *pknH1* and *pknH2* that are still present in most other MTBC members, such as *M. africanum* (Bentley et al., 2012) or *M. suricattae* (Parsons et al., 2013), as well as in *M. canettii* strains (Supply et al. 2013). This genetic lesion named region of difference 900 (RD900), which is characteristic for "modern" (TbD1 region deleted; Brosch et al., 2002) *M. tuberculosis* strains (Bentley et al., 2012) also involved the deletion of an intervening gene originally located between the *pknH1 and pknH2* genes. The *pknH* gene of *M. tuberculosis* was described as playing a role in regulating bacillary load in mouse organs to facilitate adaptation to the host environment (Papavinasasundaram et al., 2005), which suggests that the recombined *pknH* gene might have retained or gained some signalling function.

Similarly, recombination of two *pks5* genes might have resulted in the generation of a potentially functional new gene. Earlier studies reported that a *pks5* knockout-mutant of *M. tuberculosis* H37Rv was attenuated for virulence in mice, which suggests a putative function for the single, recombined *pks5* of *M. tuberculosis* (Rousseau et al., 2003), although some uncertainty prevails as the mutant was not complemented (Rousseau et al., 2003). Moreover, a non-synonymous SNP in *pks5* was recently suggested as a mutation that might have contributed, among others, to the expansion of the highly successful European-Russian Beijing lineage of *M. tuberculosis* (Merker et al., 2015), but this finding also needs experimental confirmation.

In conclusion, the recombination-derived Pks5 present in *M. tuberculosis* and the other members of the clonal MTBC might have retained some biological function, although the original function in LOS biosynthesis seems to have been lost by the recombination event during the evolution from *M. canettii*-like tubercle bacilli towards the MTBC, as we have shown in this study by using different strains and complementation constructs. At present it is unclear however, to which biological process the recombined Pks5 of *M. tuberculosis* might contribute, as indistinguishable lipid profiles between the parental strain and the *pks5* knock-out strain were found in previous studies (Rousseau et al., 2003). Further research is warranted to clarify this point.

REFERENCES

Bange *et al*. Survival of mice infected with *Mycobacterium smegmatis* containing large DNA fragments from *Mycobacterium tuberculosis*. Tubercle and Lung Disease **79**, 171–180 (1999)

References (cont.)

Bentley, *et al.* The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. *PLoS Negl Trop Dis* **6**, e1552 (2012).

Brosch *et al*. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci USA **99**, 3684-3689 (2002).

Cole, *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544 (1998).

Daffe, *et al*. Novel type-specific lipooligosaccharides from *Mycobacterium tuberculosis*. Biochemistry. **30**, 378-388 (1991).

Gonzalo-Asensio, *et al*. Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. Proc Natl Acad Sci U S A. **111**, 11491-6 (2014).

Merker, *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat Genet.* **47**, 242-249 (2015).

Papavinasasundaram et al. Deletion of the *Mycobacterium tuberculosis* pknH gene confers a higher bacillary load during the chronic phase of infection in BALB/c mice. J Bacteriol. **187**, 5751-60 (2005).

Parsons et al., Novel cause of tuberculosis in meerkats, South Africa. *Emerg Infect Dis*. **19,** 2004–07 *(*2013).

Pouseele & Supply Accurate whole-genome sequencing-based epidemiological surveillance of *Mycobacterium tuberculosis.* Methods in Microbiology. doi:10.1016/ bs.mim.2015.04.001 (2015).

Quadri, L.E. Biosynthesis of mycobacterial lipids by polyketide synthases and beyond. Crit Rev Biochem Mol Biol,. **49**, 179-211 (2014).

Rousseau, C. *et al.* Virulence attenuation of two Mas-like polyketide synthase mutants of *Mycobacterium tuberculosis.* Microbiology **149**, 1837-47 (2003).

Supply, P., *et al*. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis.* Nat Genet. **45**,172-9 (2013).

Tamura, K., *et al. MEGA5:* molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. **28,** 2731-9 (2011).