

In the format provided by the authors and unedited.

Mutations in *ppe38* block PE_PGRS secretion and increase virulence of *Mycobacterium tuberculosis*

Louis S. Ates^{1,2*}, Anzaan Dippenaar³, Roy Ummels¹, Sander R. Piersma⁴, Aniek D. van der Woude¹, Kim van der Kuij¹, Fabien Le Chevalier², Dulce Mata-Espinosa⁵, Jorge Barrios-Payán⁵, Brenda Marquina-Castillo⁵, Carolina Guapillo⁵, Connie R. Jiménez⁴, Arnab Pain⁶, Edith N. G. Houben⁷, Robin M. Warren³, Roland Brosch², Rogelio Hernández-Pando⁵ and Wilbert Bitter^{1,7*}

¹Department of Medical Microbiology and Infection Prevention, VU University Medical Center, Amsterdam, The Netherlands. ²Unit for Integrated Mycobacterial Pathogenomics, Institut Pasteur, Paris, France. ³DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa. ⁴Department of Medical Oncology, OncoProteomics Laboratory, VU University Medical Center, Amsterdam, Netherlands.

⁵Experimental Pathology Section, Department of Pathology, National Institute of Medical Sciences and Nutrition “Salvador Zubirán”, México City, Mexico.

⁶Pathogen Genomics Laboratory, BESE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. ⁷Section Molecular Microbiology, Amsterdam Institute of Molecules, Medicine & Systems, Vrije Universiteit, Amsterdam, The Netherlands. *e-mail: Louis.Ates@gmail.com; W.Bitter@vumc.nl

Supplementary information to:

Mutations in *ppe38* block PE_PGRS secretion and increase virulence of *Mycobacterium tuberculosis*

Authors:

Louis S Ates, Anzaan Dippenaar, Roy Ummels, Sander R Piersma, Aniek D van der Woude, Kim van der Kuij, Fabien le Chevalier, Dulce Mata-Espinosa, Jorge Barrios-Payán, Brenda Marquina-Castillo, Carolina Guapillo, Connie R Jiménez, Arnab Pain, Edith NG Houben, Rob M Warren, Roland Brosch, Rogelio Hernández-Pando, Wilbert Bitter

Contents:

Supplementary Tables 1-8

Supplementary Figures 1-12

Supplementary Discussion

Supplementary Methods

Supplementary References

Supplementary Tables

Supplementary Table 1

Protein name (H37Rv)	MW (kDa)	Cover age %	SAWC 2088	SAWC 2135	SAWC 2701	SAWC 2135-C	SAWC 2701-C
esxB	10.8	99	884.4	627.4	531.9	423.1	541.0
esxA	9.9	98.9	426.1	272.4	256.4	288.3	451.3
espB	47.6	55.7	144.6	39.4	92.1	58.6	84.5
mpt64	24.9	65.8	102.0	102.6	107.6	77.1	128.5
espA	39.9	31.4	99.4	133.5	133.7	142.0	141.9
esxK	11.0	95.9	84.6	100.4	125.7	78.8	143.0
espK	73.9	20.5	81.0	28.8	55.0	38.5	43.4
espE	42.1	19.4	61.1	17.5	34.4	25.4	32.8
espF	10.6	58.3	53.0	6.2	35.4	21.4	34.3
fbpC	36.8	38.2	48.8	28.7	52.7	35.9	61.1
mpt63	16.5	57.2	44.4	26.4	47.1	21.3	29.0
espC	10.8	28.2	39.5	17.5	31.7	28.8	31.0
esxN	9.9	71.3	39.4	40.0	63.2	43.8	64.5
PPE18	39.2	38.4	38.8	53.7	71.6	85.6	82.3
mpt53	18.4	39.9	30.8	22.0	20.6	16.2	32.0
espJ	27.4	29.3	28.5	3.0	18.8	12.3	19.7
fbpA	35.7	32.8	24.1	27.7	38.0	29.0	39.4
Apa	32.7	32	21.6	38.1	44.7	35.9	43.7
PPE41	25.2	15.2	18.5	3.0	5.3	4.1	7.9
hbhA	21.5	29.6	17.9	13.5	7.0	10.9	1.2
mpt51	31.1	35.5	17.9	19.6	27.1	23.2	54.7
fbpB	34.6	24.9	16.7	26.0	25.1	23.2	30.5
PPE40	58.4	6	11.8	0.0	0.0	4.1	10.5
PPE52	39.1	15.2	9.3	0.0	0.0	6.1	9.3
PPE10	47.2	14.2	9.3	0.0	0.0	3.0	9.2
esxL	9.9	71.3	9.3	6.5	9.5	5.2	9.9
PE25	10.7	36.4	6.8	0.0	0.0	0.0	1.9
PE15	9.9	33.3	6.2	1.0	2.6	6.9	4.0
PPE68	37.3	7.6	6.2	1.1	3.5	0.9	2.7
PPE53	57.7	3.8	5.0	0.0	0.0	2.0	4.0
PPE20	55.4	5.6	5.0	2.1	0.0	3.0	1.2
PPE19	39.7	19.8	4.3	3.2	8.6	3.9	8.6
PPE32	39.9	8.8	4.3	2.1	2.6	3.9	4.0
PE35	9.8	24.5	3.1	0.0	3.5	0.0	1.3
espL	12.2	20	3.1	3.0	1.8	5.0	0.0
PPE16	60.2	3.7	2.4	0.0	0.0	0.0	3.4
PE4	53.1	2.6	1.9	0.0	0.0	0.0	1.3
MT3449	127	0.7	1.9	0.0	0.0	0.0	1.3
PE_PGRS38	44.5	5.5	1.2	0.0	0.0	2.0	4.0
PE_PGRS24	49.8	2.7	1.2	0.0	0.0	2.0	2.7
PE_PGRS35	53.7	3.2	1.2	0.0	0.0	0.0	1.9
EsxH	10.4	19.8	1.2	0.0	0.0	0.0	1.8
EsxG	9.8	52.6	1.2	3.2	1.8	5.2	5.2
PPE36	27.5	21	0.0	22.8	0.0	21.6	0.0
PPE51	38.0	12.6	0.0	14.3	0.0	5.0	0.0
PPE65	40.7	4.8	0.0	3.0	0.0	0.0	0.0
PPE26	40.6	11.4	0.0	4.1	0.9	3.0	1.9
EspD	19.8	6.5	0.0	0.0	0.9	0.0	0.0
PPE38	37.4	9	0.0	0.0	0.0	4.1	15.3

PE_PGRS39	38.7	10.4	0.0	0.0	0.0	4.1	7.3
PE31	9.8	17.3	0.0	0.0	0.0	0.0	3.0
PE_PGRS29	32.9	9.1	0.0	0.0	0.0	0.9	2.5
PPE33	47.2	2.8	0.0	0.0	0.0	2.0	1.3
PE_PGRS59	37.0	5.7	0.0	0.0	0.0	0.0	1.9
PE_PGRS42	57.2	1.4	0.0	0.0	0.0	0.0	1.3
PE12	31.0	7.5	0.0	0.0	0.0	0.0	1.3
PE5	9.6	27.5	0.0	0.0	0.0	0.0	1.5
PPE64	61.6	4.5	0.0	0.0	0.0	0.0	1.2
PE22	10.4	12.2	0.0	0.0	0.0	1.1	0.0

Supplementary Table 1: LC-MS/MS detected secreted proteins of *M. tuberculosis* isolates.

Concentrated culture filtrate of the indicated strains was separated by SDS-PAGE and was analysed by LC-MS/MS. Samples were prepared and analysed in biological replicates and are depicted as the sum of spectral counts after normalization. –C = Indicated strain expressing plasmid pMV-*ppe38-71*. Only predicted secreted proteins were analysed. Data are available via ProteomeXchange with identifier PXD003842 ¹.

Supplementary Table 2

	SAWC_2135	SAWC_2135- pMV:: <i>ppe38-71</i>	SAWC_2701	SAWC_2701-- pMV:: <i>ppe38-71</i>
PPE40	0.0	4.1	0.0	10.5
PPE38	0.0	4.1	0.0	15.3
PE_PGRS39	0.0	4.1	0.0	7.3
PPE10	0.0	3.0	0.0	9.2
PPE53	0.0	2.0	0.0	4.0
PE_PGRS38	0.0	2.0	0.0	4.0
PE_PGRS24	0.0	2.0	0.0	2.7
PPE33	0.0	2.0	0.0	1.3
PE_PGRS29	0.0	0.9	0.0	2.5
PPE52	0.0	6.1	0.0	9.3

Supplementary Table 2: Proteins restored in secretion upon PPE38 complementation.

Depicted are the spectral counts of proteins detected in LC-MS/MS analysis of the culture filtrate of two clinical isolates of *M. tuberculosis* with mutations in the *ppe38* locus (SAWC_2135 and SAWC_2701) and these isolates complemented with the *ppe38* locus on a plasmid (pMV::*ppe38-71*). Only the PE, PPE, Esx, ESX-1 associated (Esp) and nine major secreted substrates of *M. tuberculosis* were analysed and proteins were depicted when a two-fold increase in detected spectral counts occurred in both complemented strains compared to the original isolates. Data are available via ProteomeXchange with identifier PXD003842¹.

Supplementary Tables three and four can be downloaded in the online version of this paper.

Supplementary Table 3: Small insertions and deletions (indels) identified in strains SAWC_1945, SAWC_2135 and SAWC_2701. Sheet 1 contains method summary and legend for the file. Sheet 2-4 lists the identified indels for the corresponding strain. Identified indels were filtered to keep only those variants identified in all three alignments (BWA, Novoalign and SMALT). Sheet 5 summarizes indels present in all strains (green), in SAWC_1945, but not in SAWC_2135 and SAWC_2701 (blue), or vice versa (red).

Supplementary Table 4: Single nucleotide polymorphisms identified in strains SAWC_1945, SAWC_2135 and SAWC_2701. Sheet 1 contains method summary and legend for the file. Sheet 2-4 list Single-nucleotide variants for each strain. Sheet 5 contains a between strain comparison of the three sequenced isolates. Identified SNPs were filtered to keep only those variants identified in all three alignments (BWA, Novoalign and SMALT).

Supplementary Table 5: Strain information

Strain name	Species	Original Strain	Relevant strain characteristics
M^{USA}	<i>M. marinum</i>	M ^{USA}	Clinical isolate
<i>ppe38::tn</i>	<i>M. marinum</i>	M ^{USA}	Himar transposon in <i>mmar_3661</i> (position 4508719). Kan resistance cassette on transposon has same orientation as ORF <i>mmar_3661</i>
LA10	<i>M. marinum</i>	M ^{USA} ::pSMT3- <i>mspA</i>	Transposon mutant in <i>mmar_3661</i> (4508474). Orientation reversed to ORF <i>mmar_3661</i>
<i>M. tuberculosis</i> CDC1551	<i>M. tuberculosis</i>	CDC1551	Wild-type strain
<i>Δppe38-71</i>	<i>M. tuberculosis</i>	CDC1551	<i>ppe38</i> locus deleted, see figure 2A
<i>ΔesxXY</i>	<i>M. tuberculosis</i>	CDC1551	<i>ppe38</i> locus partially deleted, See figure 2A
<i>ΔBeijing</i>	<i>M. tuberculosis</i>	CDC1551	See figure 2A
SAWC_2088	<i>M. tuberculosis</i>	Clinical isolate	Strain isolated in Cape Town cohort (Hanekom et al 2007). Strain 1 in Aguilar et al 2010.
SAWC_1945	<i>M. tuberculosis</i>	Clinical isolate	Strain isolated in Cape Town cohort (Hanekom et al 2007). Isolated from the same patient as SAWC_2088 and shown to be identical by spoligotyping.
SAWC_2135	<i>M. tuberculosis</i>	Clinical isolate	Strain isolated in Cape Town cohort (Hanekom et al 2007).
SAWC_2701	<i>M. tuberculosis</i>	Clinical isolate	Strain isolated in Cape Town cohort (Hanekom et al 2007).

Supplementary Table 6: Primers used in this study

Primer name	Sequence 5' → 3'	
<i>Ppe38-Fw</i>	CGGGATCCAGGAGGGGTTGTGATGATTTTGGACTTTG	BamHI
<i>Ppe38-Rv</i>	TTTAAGCTTTCCGACACCGATCCGCGGCACCA	HindIII
<i>Ppe38-Rv-HA</i>	CCCAAGCTTTCACGCGTAGTCCGGCACGTCGTACGGGTA TCCGACACCGATCCGCGGCACCA	HindIII
<i>esxPN-Fw</i>	AAAGGATCCGCGGGGGGTGATTGGTCGGG	BamHI
<i>esxPN-Rv</i>	AAGCTTATGGGCCACATCTGACTAGG	HindIII
<i>3581-Fw</i>	CCCTCTAGAGTGCCCGTAGACGTTGTCCC	XbaI
<i>3581-HA-Rv</i>	CCCAGATCT TCACGCGTAGTCCGGCACGTCGTACG GGTACATAACCTCCGGTGCCGCCG	BglII
Mt2422-Fw	TTTGAATTCTTTTCGGTGTGGATTGTCT	EcoRI
Mt2419-Rv	TTTAAGCTTGCCAGGGATTCCAACGAC	HindIII
PPE38 Del2 RR	TTTTTTTTCCATCTTTTGG GGTCGAATGACCGGTTCTCC	Van91I
PPE38 Del2 RF	TTTTTTTTCCATAGATTGG GGCCTTGGTGCCGTGATTGT	Van91I
PPE38 Del2 LR	TTTTTTTTCCATTTCTTGG ATTCGGCATGAGCCGTTAC	Van91I
PPE38 Del2 LF	TTTTTTTTCCATAAATTGG ACTCCGGTACTGCGAACACT	Van91I
PPE38 (51bp) KO LR	TTTTTTTTCCATTTCTTGG CTATGCCGGTGCGGGGTC	Van91I
PPE38 (51bp) KO LF	TTTTTTTTCCATAAATTGG CGCTTCGGTGCACTTCATTT	Van91I
PPE38 (72bp) KO RF	TTTTTTTTCCATAGATTGG CCCCACCCCGCACCGGCATA	Van91I
PPE38 (72bp) KO RR	TTTTTTTTCCATCTTTTGG GATGTCAACACTGGCTTAGG	Van91I
PPE38 Del LF	TTTTTTTTCCATAAATTGG CCGCTTCGGTGCACTTCATT	Van91I
PPE38 Del LR	TTTTTTTTCCATTTCTTGG TTGTGTCTGAGCCGTTGAGG	Van91I
PPE38 Del RF	TTTTTTTTCCATAGATTGG TGCCGCAATGTCATCAACGC	Van91I
PPE38 Del RR	TTTTTTTTCCATCTTTTGG GTTGCCGCCGGAGATCAACT	Van91I

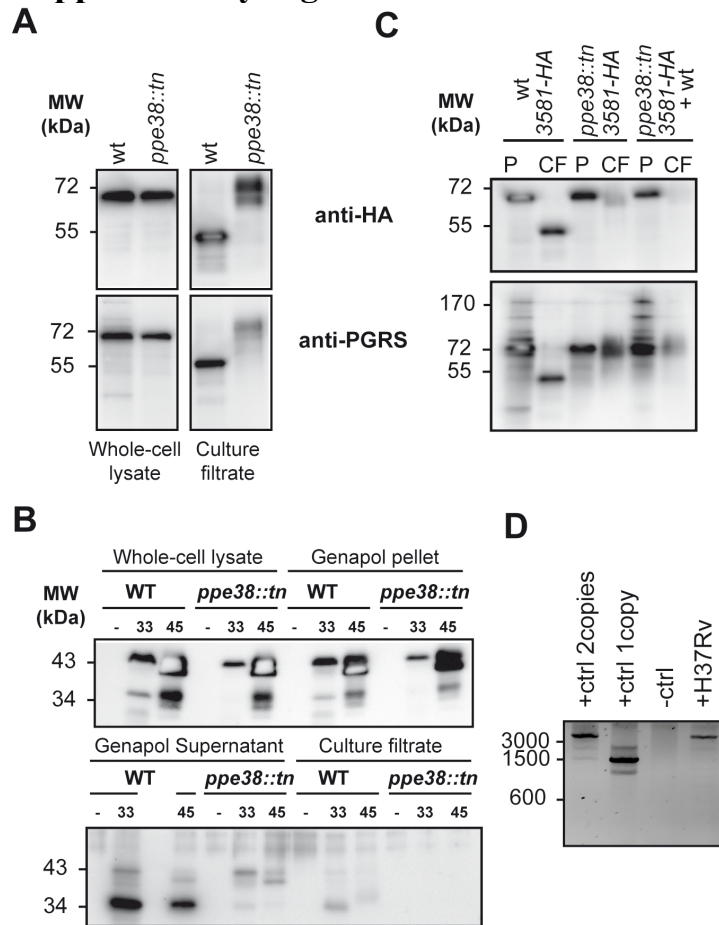
Supplementary Table 7: Plasmid information

Plasmid name	Plasmid backbone <i>promotor</i>, selection marker	Gene	Reference
pPPE38	pSMT3, <i>hsp60</i> , hygromycin	<i>Mmar_3661</i>	This study
pPPE38-PN	pSMT3, <i>hsp60</i> , hygromycin	<i>Mmar_3661-59</i>	This study
pPN	pSMT3, <i>hsp60</i> , hygromycin	<i>Mmar_3660-59</i>	This study
pPPE38-HA	pSMT3, <i>hsp60</i> , hygromycin	<i>Mmar_3661</i>	This study
pPPE38-HA-PN	pSMT3, <i>hsp60</i> , hygromycin	<i>Mmar_3661-59</i>	This study
p3581-HA	pSMT3, <i>hsp60</i> , hygromycin	<i>Mmar_3581</i>	This study
p19kPro:: PE_PGRS45	p19kPro, <i>19kDa</i> <i>lipoprotein</i> , hygromycin	<i>rv2615c</i>	Abdallah et al. 2009 ⁴⁴
p19kPro:: PE_PGRS33	p19kPro, <i>19kDa</i> <i>lipoprotein</i> , hygromycin	<i>rv1818c</i>	Abdallah et al. 2009 ⁴⁴
pSMT3::<i>mmar_0761- HA</i>	pSMT3, <i>hsp60</i> , hygromycin	<i>Mmar_0761</i>	Ates et al. 2016 ⁴⁵
pMV::ppe38-71	pMV, <i>hsp60</i> , kanamycin	<i>mt2419-22</i>	This study

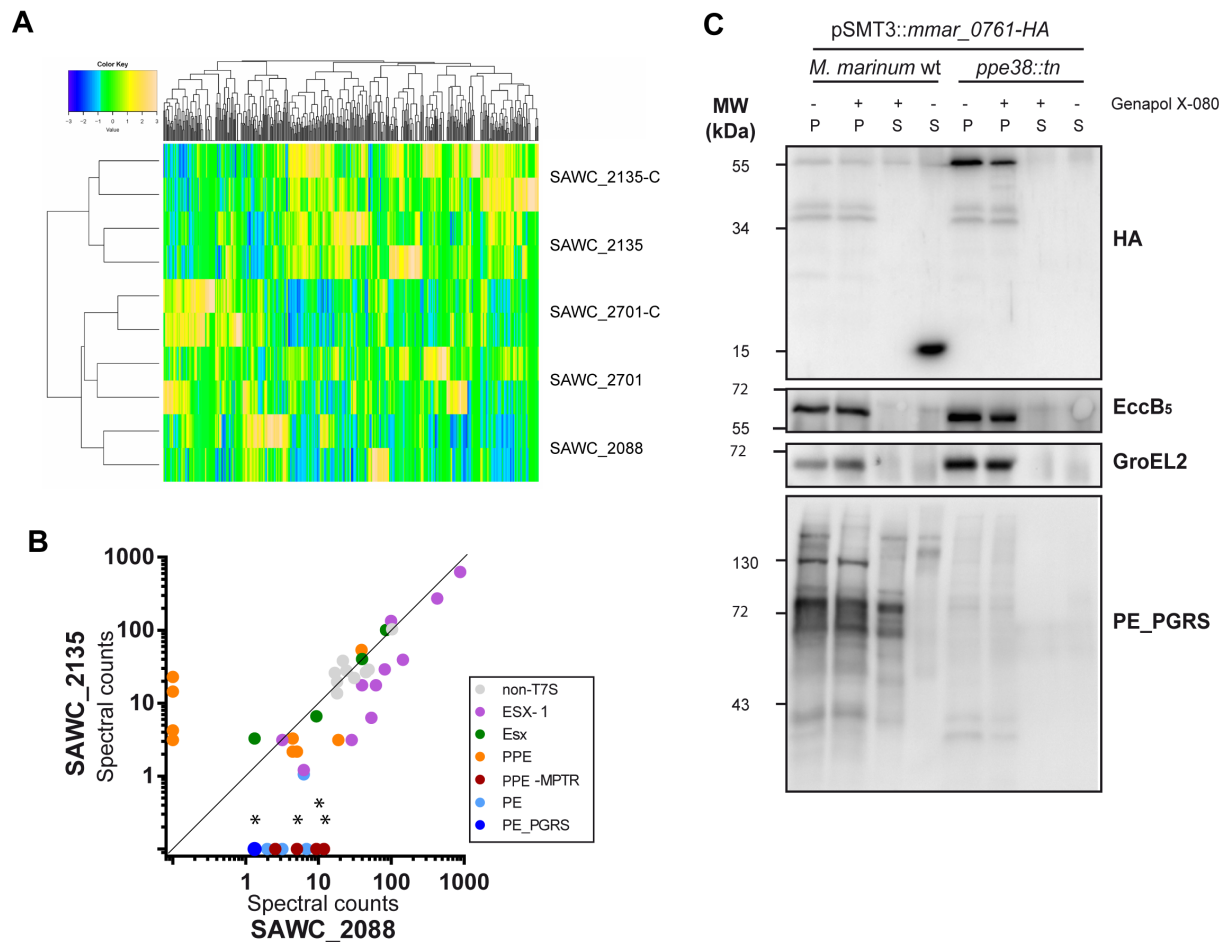
Supplementary Table 8: Sequences of *Mycobacterium. sp.* isolates included in the phylogenetic analysis

Isolate name	Original study	Accession
<i>M. tuberculosis</i> T92	Comas <i>et al.</i> 2010 ⁴⁶	SRX003589
<i>M. tuberculosis</i> T17	Comas <i>et al.</i> 2010 ⁴⁶	SRX005394
<i>M. tuberculosis</i> 95_0545	Comas <i>et al.</i> 2010 ⁴⁶	SRX007721
<i>M. tuberculosis</i> K21	Comas <i>et al.</i> 2010 ⁴⁶	SRX002001
<i>M. tuberculosis</i> K67	Comas <i>et al.</i> 2010 ⁴⁶	SRX002004
<i>M. tuberculosis</i> K93	Comas <i>et al.</i> 2010 ⁴⁶	SRX002005
<i>M. tuberculosis</i> T67	Comas <i>et al.</i> 2010 ⁴⁶	SRX007715
<i>M. tuberculosis</i> T85	Comas <i>et al.</i> 2010 ⁴⁶	SRX003590
<i>M. tuberculosis</i> 00_1695	Comas <i>et al.</i> 2010 ⁴⁶	SRX007716
<i>M. tuberculosis</i> 98_1833	Comas <i>et al.</i> 2010 ⁴⁶	SRX007718
<i>M. tuberculosis</i> M4100A	Comas <i>et al.</i> 2010 ⁴⁶	SRX007719
<i>M. tuberculosis</i> 91_0079	Comas <i>et al.</i> 2010 ⁴⁶	SRX007720
<i>M. tuberculosis</i> K49	Comas <i>et al.</i> 2010 ⁴⁶	SRX002002
<i>M. tuberculosis</i> GM_1503	Comas <i>et al.</i> 2010 ⁴⁶	SRX012272
<i>M. tuberculosis</i> 4783_04	Comas <i>et al.</i> 2010 ⁴⁶	SRX007723
<i>M. tuberculosis</i> K37	Comas <i>et al.</i> 2010 ⁴⁶	SRX002003
<i>M. africanum</i> 544404	Comas <i>et al.</i> 2010 ⁴⁶	SRX007725
<i>M. africanum</i> 11821_03	Comas <i>et al.</i> 2010 ⁴⁶	SRX007724
<i>M. africanum</i> 4141_04	Comas <i>et al.</i> 2010 ⁴⁶	SRX007726
<i>M. canettii</i>	Comas <i>et al.</i> 2010 ⁴⁶	SRX002429
<i>M. tuberculosis</i> H37Rv	Coscolla <i>et al.</i> 2013 ⁴⁷	ERS153830
<i>M. tuberculosis</i> ERR552912	Merker <i>et al.</i> 2015 ¹⁵	ERR552912
<i>M. tuberculosis</i> ERR550670	Merker <i>et al.</i> 2015 ¹⁵	ERR550670
<i>M. tuberculosis</i> ERR551412	Merker <i>et al.</i> 2015 ¹⁵	ERR551412
<i>M. tuberculosis</i> ERR553313	Merker <i>et al.</i> 2015 ¹⁵	ERR553313
<i>M. tuberculosis</i> ERR553116	Merker <i>et al.</i> 2015 ¹⁵	ERR553116
<i>M. tuberculosis</i> ERR551520	Merker <i>et al.</i> 2015 ¹⁵	ERR551520
<i>M. tuberculosis</i> MT5488	Lee <i>et al.</i> 2015 ⁴⁸	SRR1640751
<i>M. tuberculosis</i> MT578	Lee <i>et al.</i> 2015 ⁴⁸	SRR1640752
<i>M. tuberculosis</i> MT4466	Lee <i>et al.</i> 2015 ⁴⁸	SRR1640750
<i>M. tuberculosis</i> SAWC 1945	This study	ERS1203569
<i>M. tuberculosis</i> SAWC 2701	This study	ERS1203570
<i>M. tuberculosis</i> SAWC 1453	This study	ERS1203571
<i>M. tuberculosis</i> SAWC 1116	This study	ERS1203572
<i>M. tuberculosis</i> SAWC 4570	This study	ERS1203573
<i>M. tuberculosis</i> SAWC 1125	This study	ERS1203574
<i>M. tuberculosis</i> SAWC 2133	This study	ERS1203575
<i>M. tuberculosis</i> SAWC 2135	This study	ERS1203576
<i>M. tuberculosis</i> SAWC 4437	This study	ERS1203577

Supplementary Figures



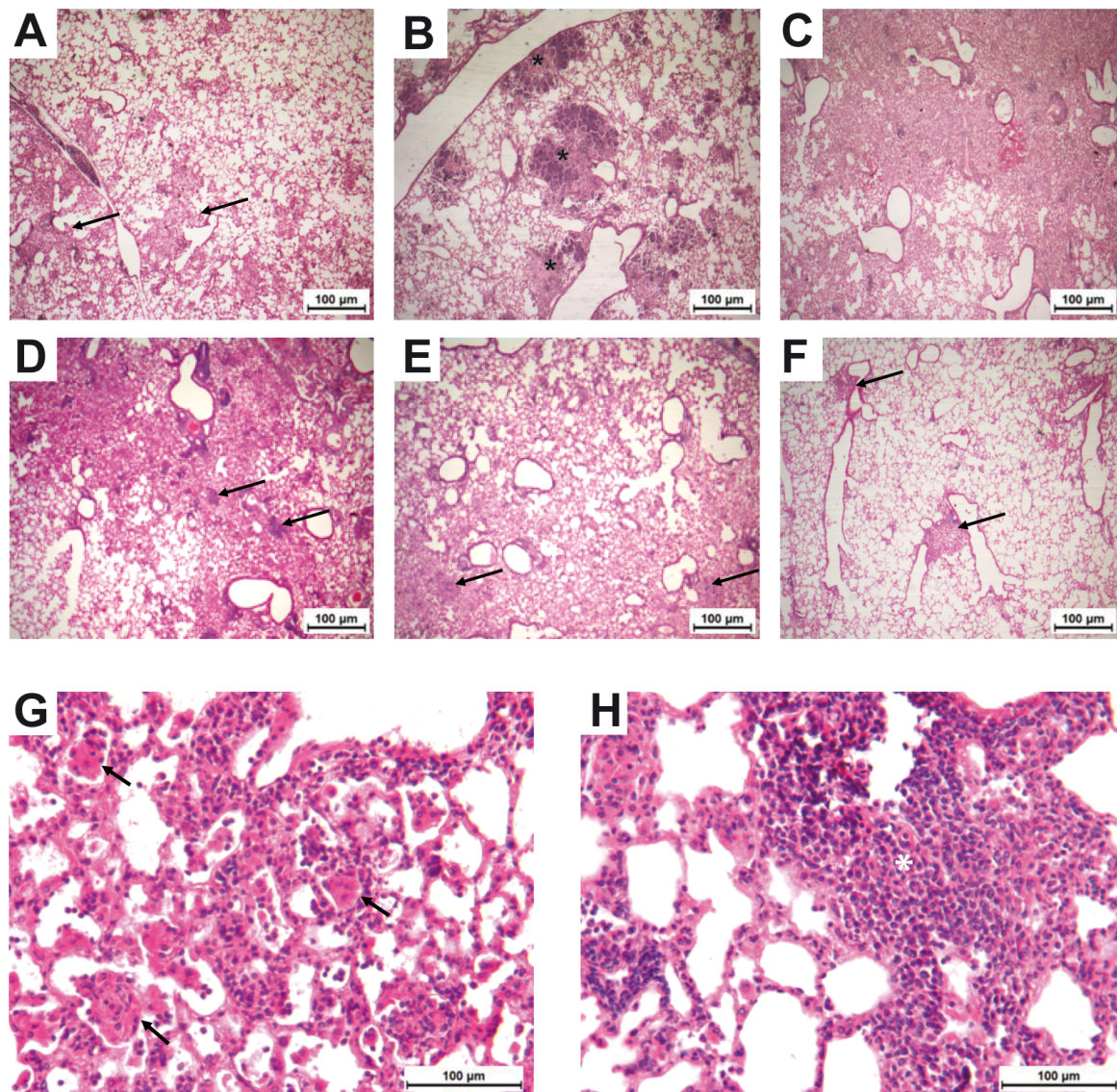
Supplementary Figure 1: Overexpressed HA-tagged PE_PGRS proteins are dependent on PPE38 for surface localisation or secretion. A) Whole cell lysate of *M. marinum* wild-type (wt) or the *ppe38::tn* mutant expressing pSMT3::mmar3581-HA. Immunoblots were stained with anti-HA or anti-PGRS antibodies. Representative of 2 replicates. B) Secretion analysis of *M. marinum* wild-type or *ppe38::tn* without plasmid (-) or expressing HA-tagged PE_PGRS33 (*rv1818c* (33)) or PE_PGRS45 (*rv2615c* (45)). Genapol supernatant contains surface enriched proteins obtained by Genapol-X080 extraction, while genapol pellet contains residual cellular material of this extraction. Proteins were detected by anti-HA staining. Represents a single experiment C) Co-culture of with wild-type *M. marinum* cannot rescue the phenotype of the *ppe38::tn*-mutant. Secretion analysis of wild-type *M. marinum* expressing 3581-HA, *ppe38::tn* expressing 3581-HA or a 1:1 mix of the latter with wild-type *M. marinum*. HA-tagged (upper figure) MMAR_3581 was detected in the culture filtrate (CF) in its processed form in wild-type *M. marinum* background (left two lanes), but not in the *ppe38::tn* background (middle lanes). Growth of the latter strain together with wild-type *M. marinum* (last 2 lanes) did not rescue this phenotype. PE_PGRS blots are shown as a control (bottom figure). Representative of technical duplicate. D) PCR verification of the *ppe38*-locus of *M. tuberculosis* H37Rv. Primers PPE38F and PPE38R from McEvoy *et al.* 2009 were used to amplify a PCR product from genomic DNA. A 3398bp product corresponds to a complete *ppe38*-locus (Figure 2A), whereas a 1335bp product corresponds to a single copy of *ppe38* and no *esxXY* genes ². PCR of *M. tuberculosis* H37Rv was performed three times in combination with different other isolates.



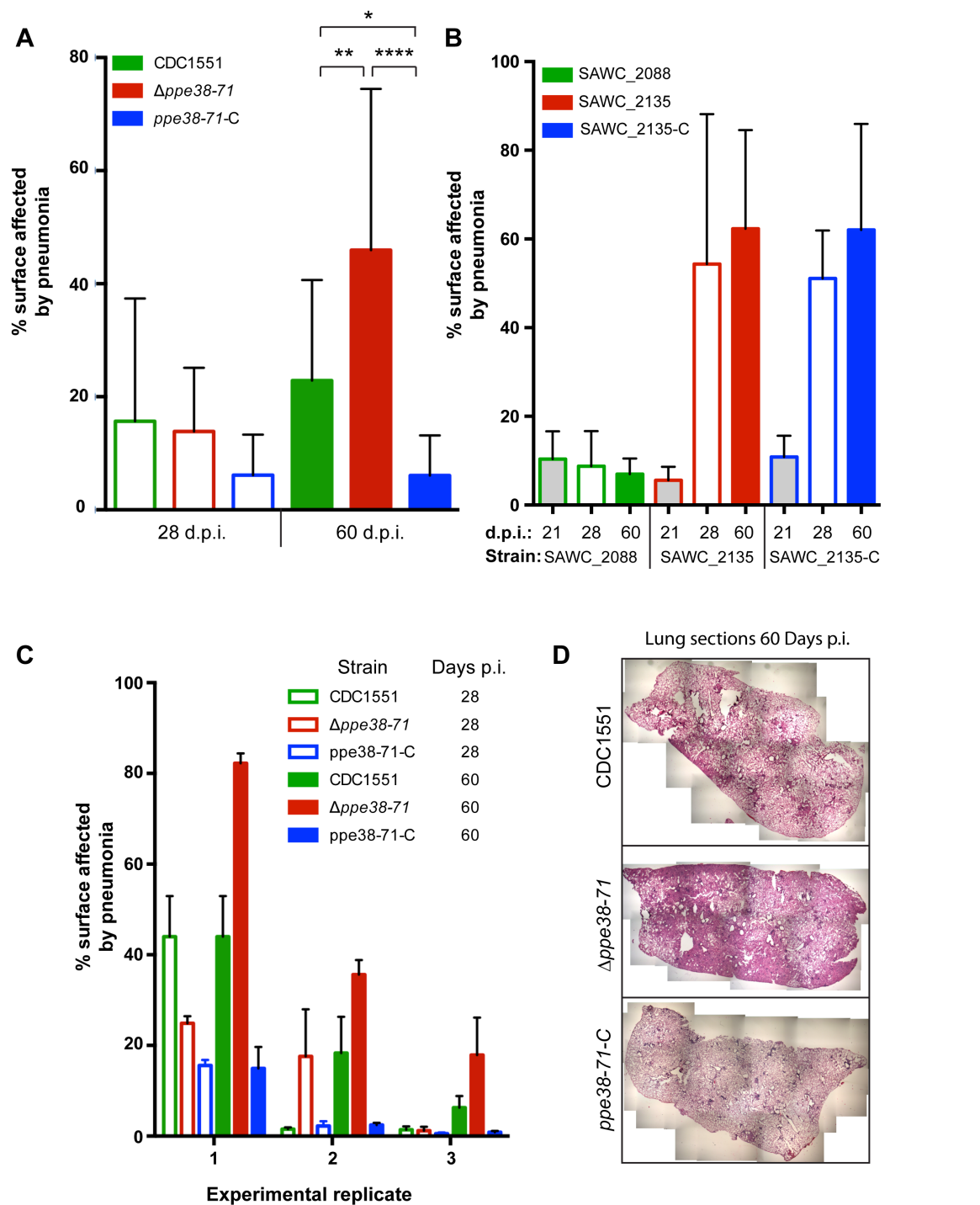
Supplementary Figure 2: LC-MS/MS analysis of the PPE38-dependent secretome of *M. tuberculosis* strains identifies PPE-MPTR and PE_PGRS as PPE38 dependent substrates. A) Unsupervised clustering of spectral counts detected in the concentrated culture filtrate of indicated strains. “-C” indicates complementation of the strain with pMV::ppe38-71. B) Visual representation of secretome analysis of PPE38 dependent substrates in SAWC_2135. Concentrated culture filtrates of the indicated strains were analysed by LC-MS/MS analysis and detected spectral counts of known secreted proteins were plotted. Legend colours indicate the different sub-groups of substrates. “*” Indicates proteins of which secretion is restored upon introduction of ppe38-71. Full data are available in Supplementary Table 1. Culture filtrate concentration and LC-MS/MS analysis depicted in S2A-B was performed in technical duplicate. C) Immunoblot secretion analysis of *M. marinum* or the *ppe38::tn* mutant strain expressing mmar_0761-HA. EccB5 and GroEL2 were used as loading and lysis controls. The processed form of MMAR_0761-HA (PPE10) was seen in the culture filtrate of wild-type *M. marinum*, but not in the *ppe38::tn* mutant. P= pellet, S= supernatant, -/+ indicates the extraction of proteins by 0.5% GenapolX-080. Inversely, full length MMAR_0761-HA accumulated in the pellet fractions of *ppe38::tn*, but not in wild-type *M. marinum*. Representative of two biological duplicates.



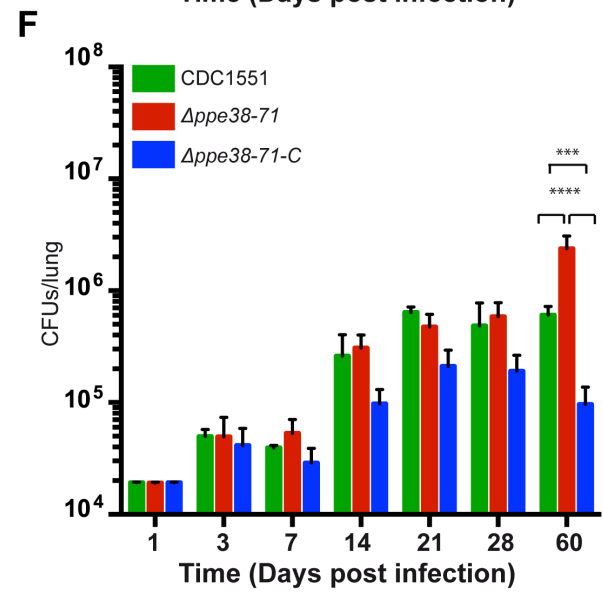
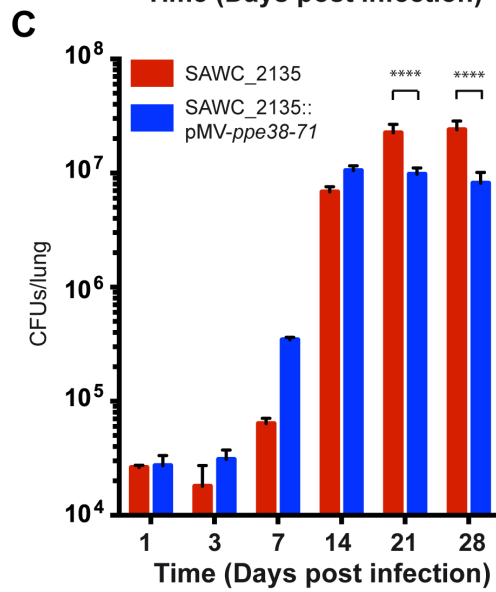
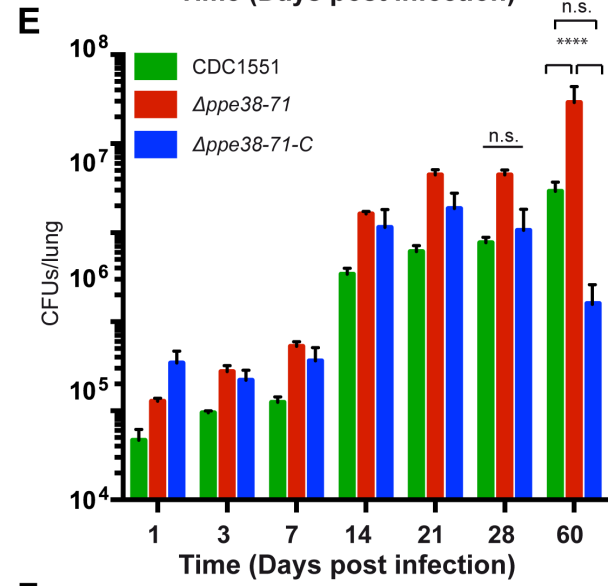
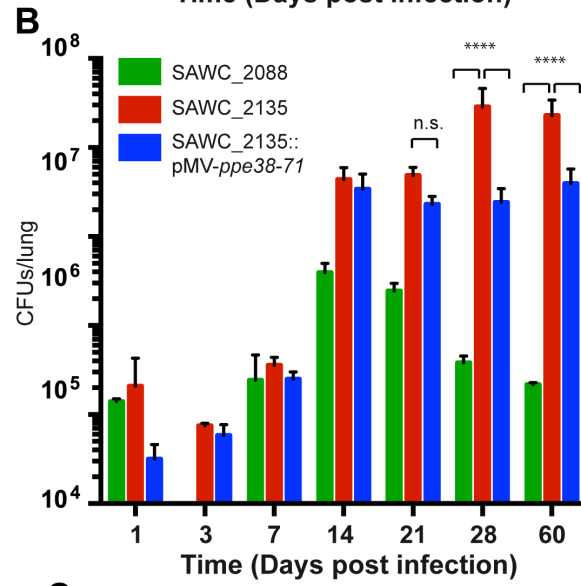
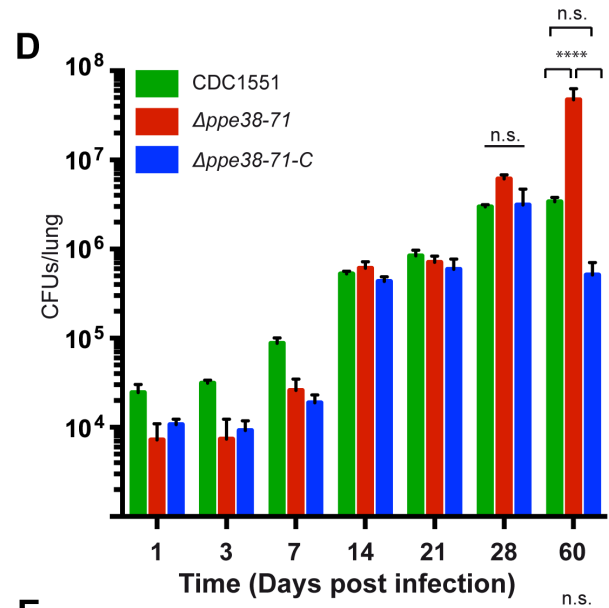
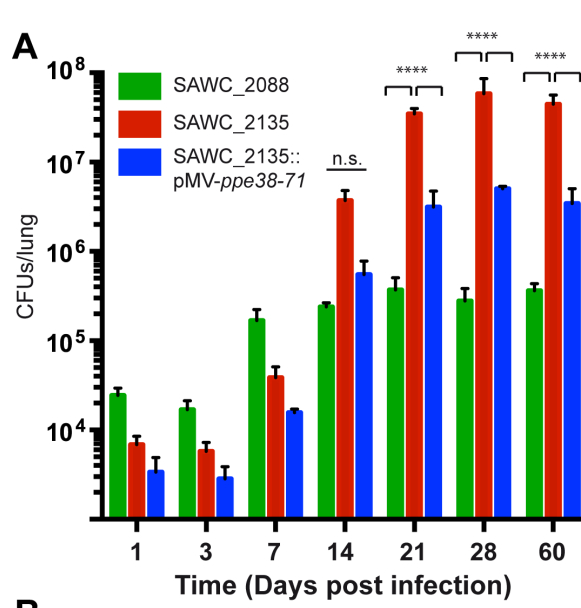
Supplementary Figure 3: Whole-genome-sequence analysis of Beijing lineage strains. A) Venn diagram summarising shared and differing small insertion and deletion events. Full data are available in Supplementary Information File 1. Numbers between brackets indicate (number of indels in coding sequences, number of indels causing a frameshift mutation) B) Venn diagram depicting shared and differing single nucleotide polymorphisms. Full data are available in Supplementary Information File 2. Numbers between brackets indicate (total number of SNPs/indels in coding sequences, number of non-synonymous SNPs/indels in coding sequences).



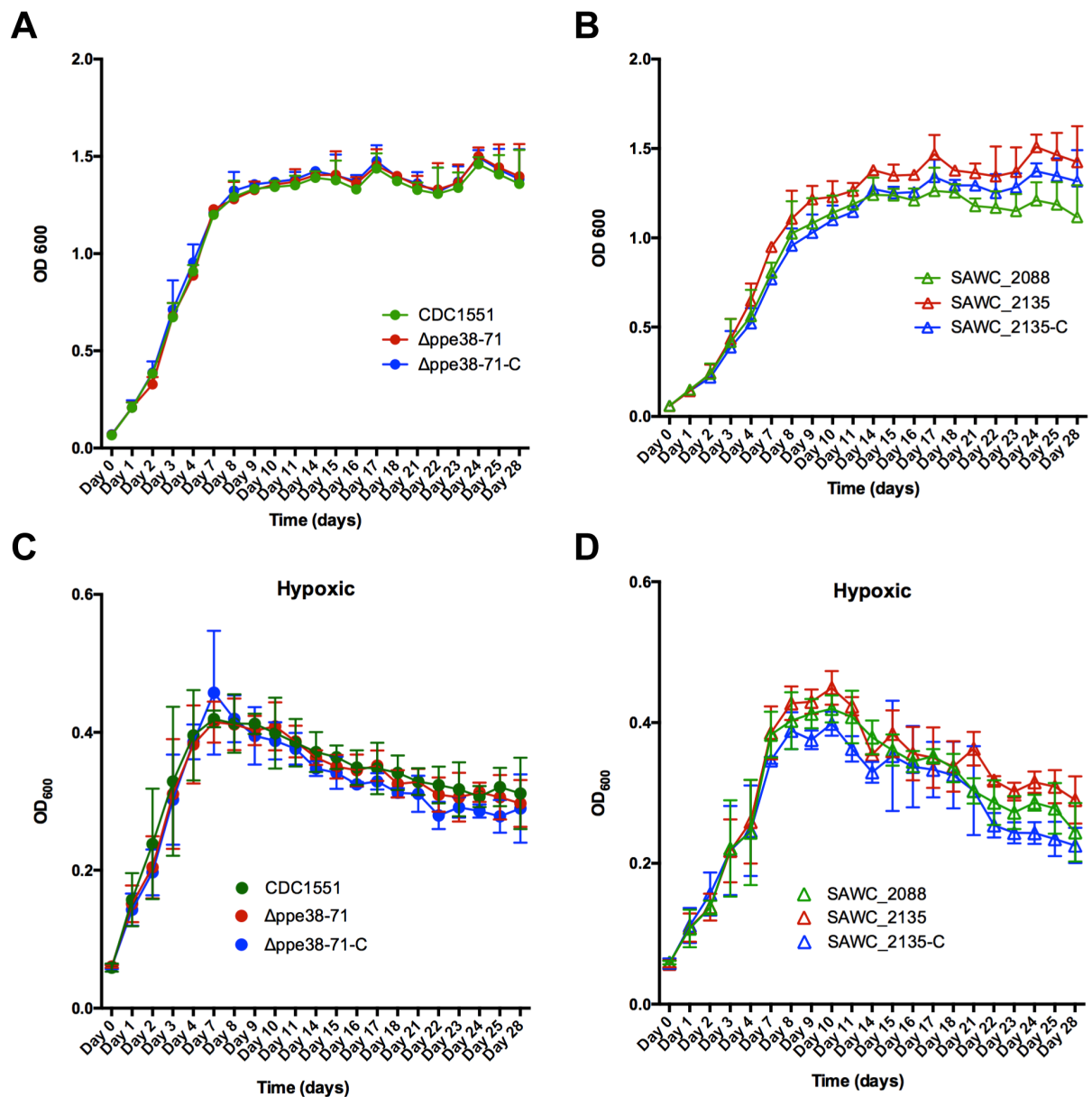
Supplementary Figure 4: Histology of lungs of infected mice. A-C, D-F) Representative low power micrographs of lungs during late infection. A) 60 d.p.i. lungs of mice infected with strain SAWC_2088 show mild size pneumonic patches (arrows). B) Large patches of pneumonia with extensive necrosis (asterisks) in the lung of mice after 28 days of infection with strain SAWC_2135. C) Extensive areas of compact pneumonia without necrosis are seen after 60 days of infection with strain the complemented SAWC2135-C strain. D) At 60 d.p.i. with strain CDC1551 lungs show extensive pneumonia with lymphocytes nodules (arrows). E) At 60 d.p.i. with strain $\Delta ppe38-71$, large areas of pneumonia with focal necrosis were observed (arrows). F) Infection with complemented strain $\Delta ppe38-71-C$ only induced small foci of pneumonia (arrows). All micrographs were stained by haematoxylin/eosin and photographed at 40x magnification. Quantification of lung surface affected by inflammation can be found in Supplementary Figure 5. G, H) Pneumonic patches induced by CDC1551 strain (G) infection are constituted by a mixture of lymphocytes with abundant macrophages preferentially located in alveolar lumina (arrows); while strain $\Delta ppe38-71$ (H) induced pneumonic patches constituted predominantly by lymphocytes (white asterisk). Hematoxylin/eosin staining, 200x magnification. Representative of three individual experiments except for A (two replicates).



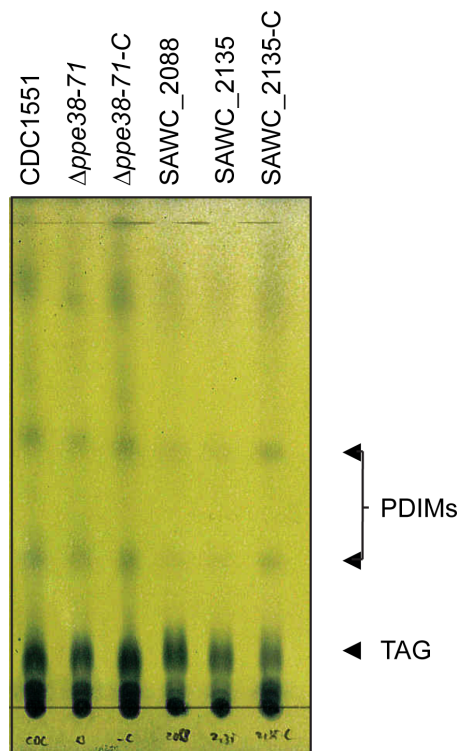
Supplementary Figure 5: Quantification of lung surface affected by pneumonia. Cryopreserved lung sections of infected animals from the indicated times post infection (d.p.i.) were photographed and analysed by automated lung morphometry. Data from A are averaged from three independent experiments depicted in C. Data in B are averaged from 2 independent experiments. All bars depict the mean level of percentage of lung surface affected by morphology, with the error bars depicting standard deviations. D) Representative (from three replications) image used for quantification of pneumonia from the second experiment depicted in C.



Supplementary Figure 6: Bacterial loads in the lungs of mice infected with *ppe38*-deficient strains; individual replications. The combined data presented in figure 2B and 2D were compiled by combining data from three independent experiments, presented here. All data were analysed by Two-way ANOVA followed by Tukey's test for multiple comparisons. For A and B, SAWC_2135 was significantly ($p<0.0001$) different from SAWC_2088 and for A, B and C SAWC_2135 was significantly higher than SAWC_2135-C at both day 28 and day 60. For D, E and F, only on day 60 were significant differences observed. In all three experiments, *Δppe38-71* was significantly ($p<0.0001$) higher than the wild-type or complemented strains. Each graph represents an experiment of 50 mice per strain, of which 5 mice were sacrificed at the indicated time-points. Experiments depicted in A and C were performed in parallel as were B and D. Infection inocula were plated at the time of infection and were: A) 2.86-, 2.40- & 2.33 *10⁵ B) 2.55-, 2.65- & 2.60 *10⁵ C) 2.60-, 2.43- & 2.46 *10⁵ D) 2.50-, 2.57- & 2.55 *10⁵ respectively in the order of strains in the graphs. All analyses were performed in GraphPad prism software version 6.0. Depicted are the means, with the error bars representing the standard deviations. **** = $p<0.0001$, *** = $p<0.001$, n.s. = not significant ($p>0.05$)

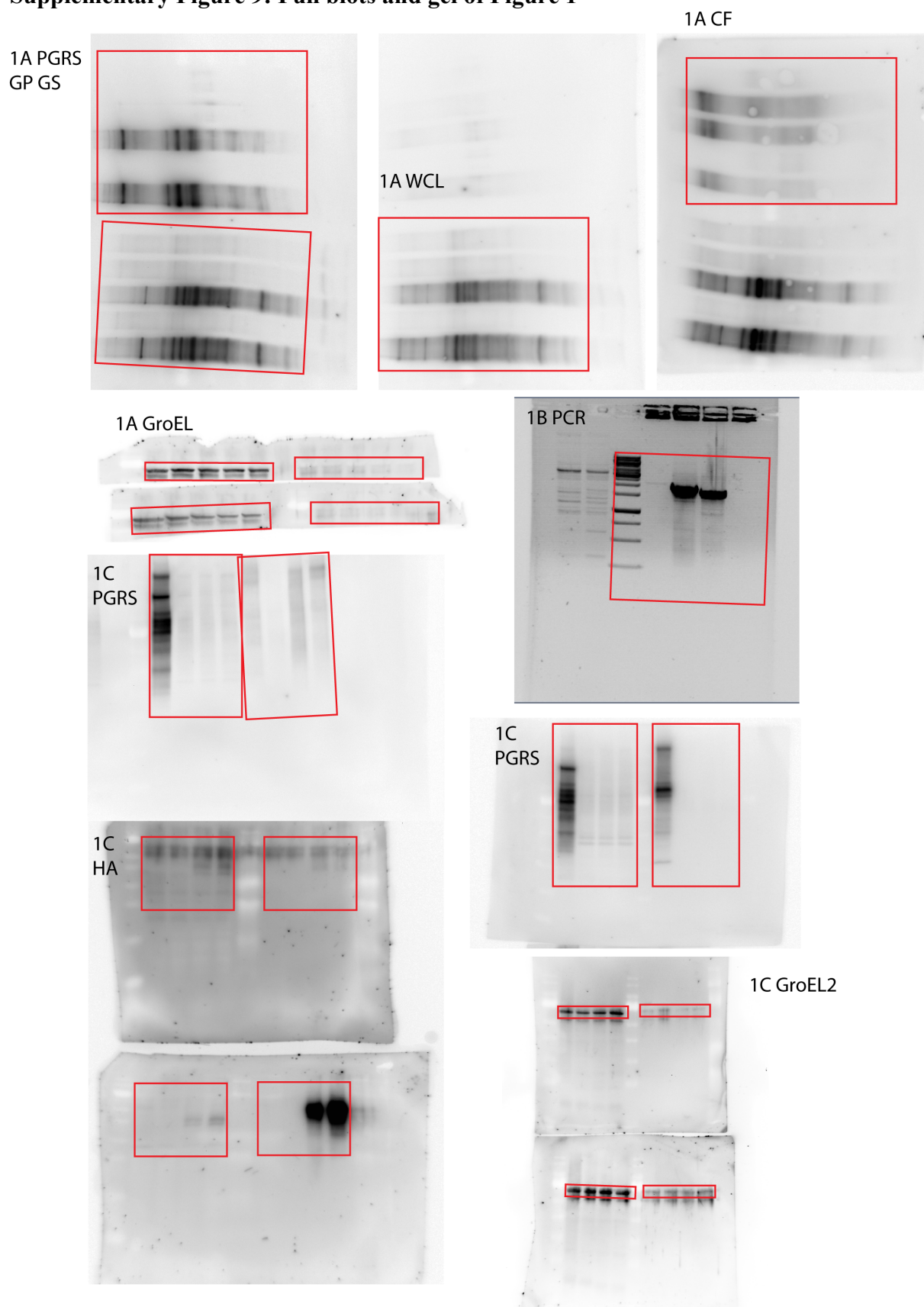


Supplementary Figure 7: *In vitro* growth curves of *in vivo* tested *M. tuberculosis* isolates. Indicated strains were inoculated at a starting density of 0.05 OD₆₀₀/ml in 5ml (A, B) or 20ml (C, D) 7H9 medium in 25ml glass tubes. OD₆₀₀ was measured inside the tubes at the indicated days. For A and B tubes were opened and aerated twice per week, while tubes remained closed for C and D and the latter were considered in hypoxia from day 7, when growth was fully arrested. Two-way ANOVA indicated that strain identity was no significant source of variation for A-C ($p > 0.05$), but was significant for D. Multiple comparisons per time point indicated no significant differences in A and C. B) SAWC_2088 and SAWC_2135 were significantly different from day 21 onward. SAWC_2088 and SAWC_2135C were significantly different from day 25 onward. D) SAWC_2135 and SAWC_2135C were significantly different in hypoxic conditions from day 21 onward. Data represent biological duplicates each consisting of technical duplicates and error bars represent the standard deviation. All statistical analyses were performed with GraphPad prism.

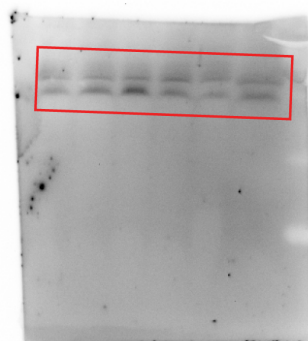
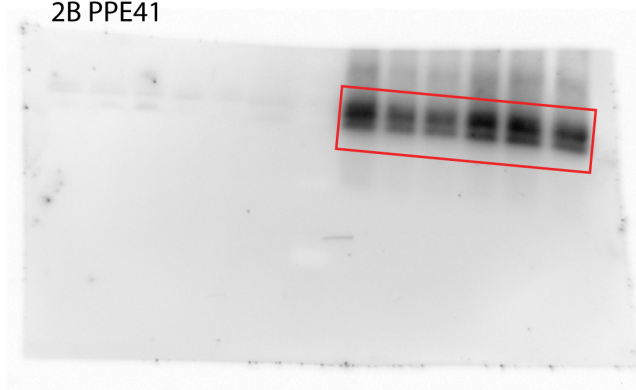


Supplementary Figure 8: PDIM analysis of *in vivo* tested *M. tuberculosis* isolates. Extracted lipids were dissolved in chloroform at 20mg/ml and 10 μ l was loaded on TLC plates. Separation was achieved by migration with a 9:1 Petroleum Ether (40-60°C):Di-ethylether solvent solution. Lipids were visualised by phosphomolybdic acid spraying and charring. TLC performed in duplicate.

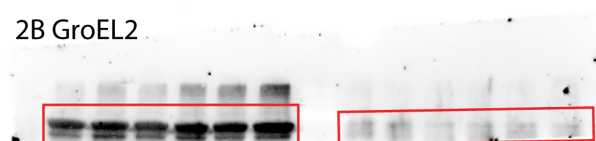
Supplementary Figure 9: Full blots and gel of Figure 1



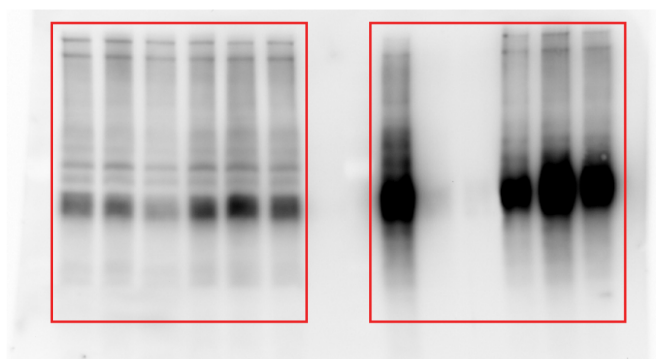
Supplementary Figure 10: Full blots of Figure 2
2B PPE41



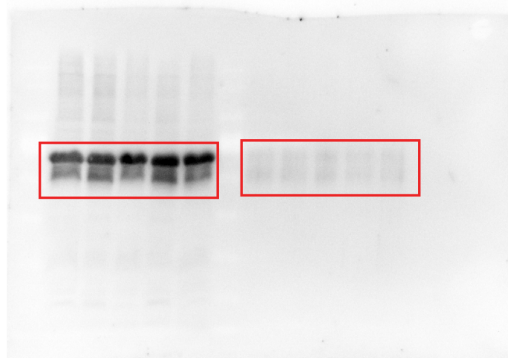
2B GroEL



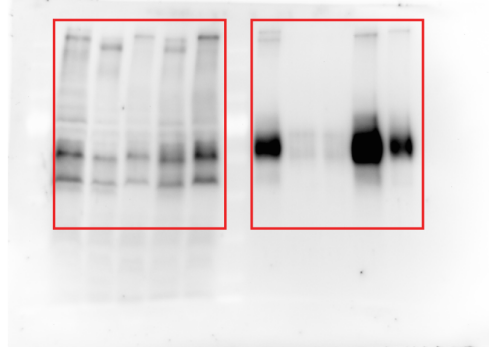
2B PGRS



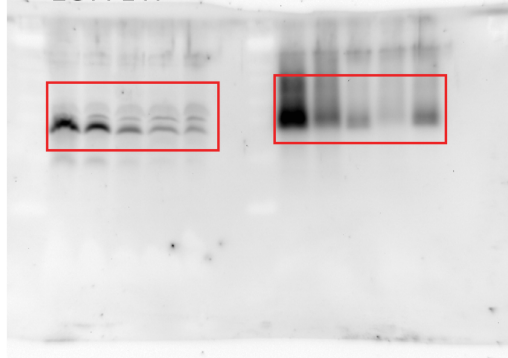
2C GroEL



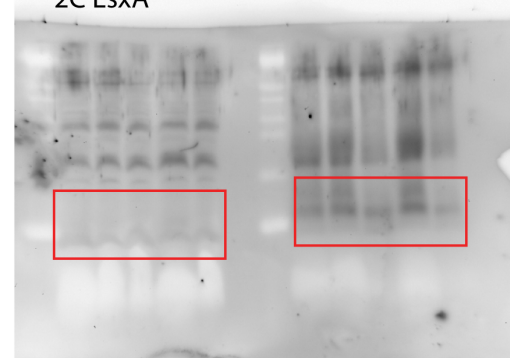
2C PGRS



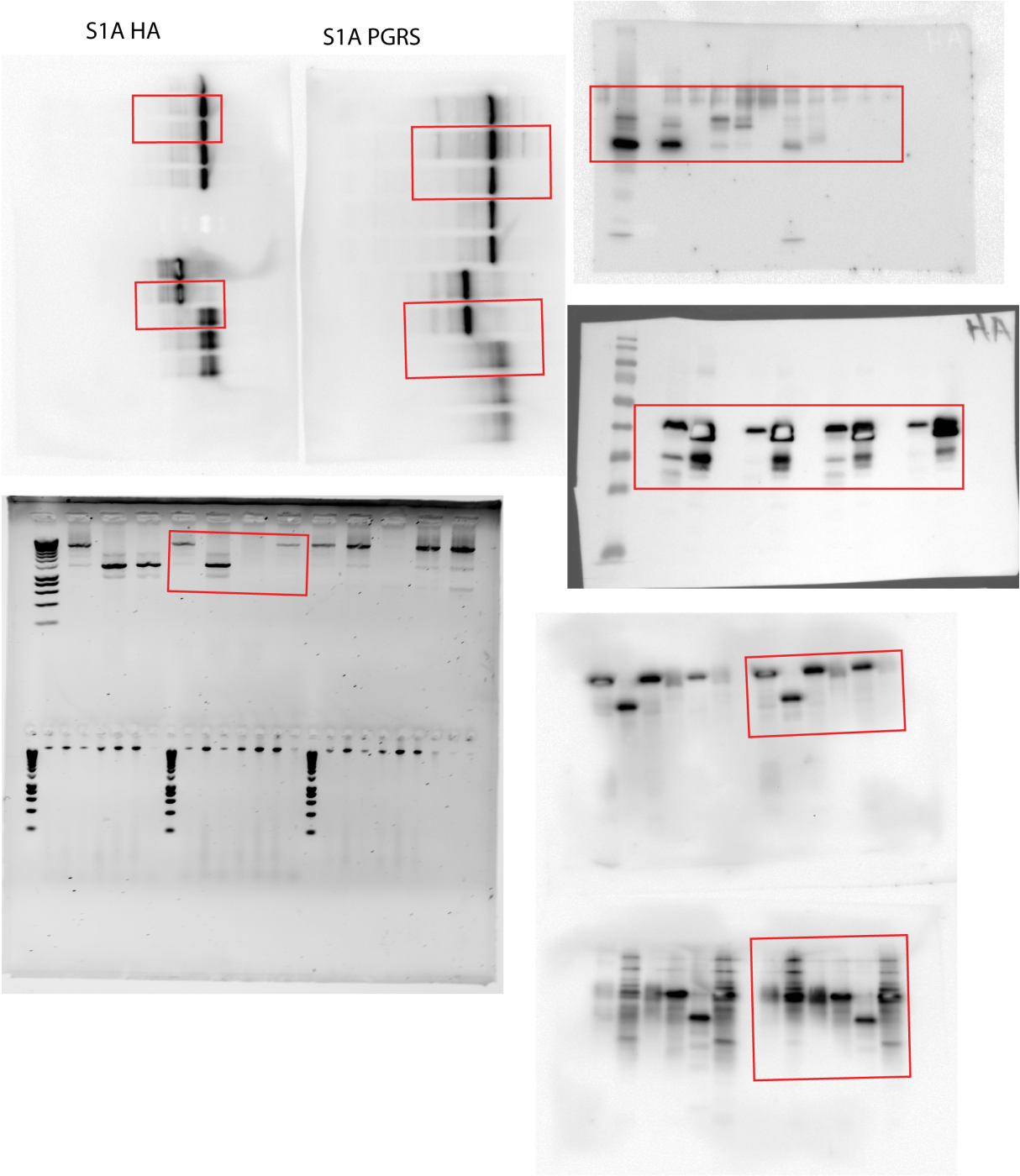
2C PPE41



2C EsxA

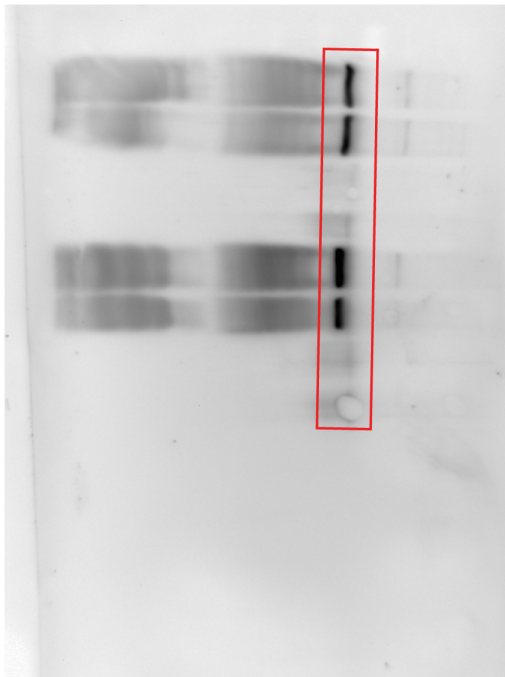


Supplementary Figure 11: Full blots and gel of Supplementary Figure 1

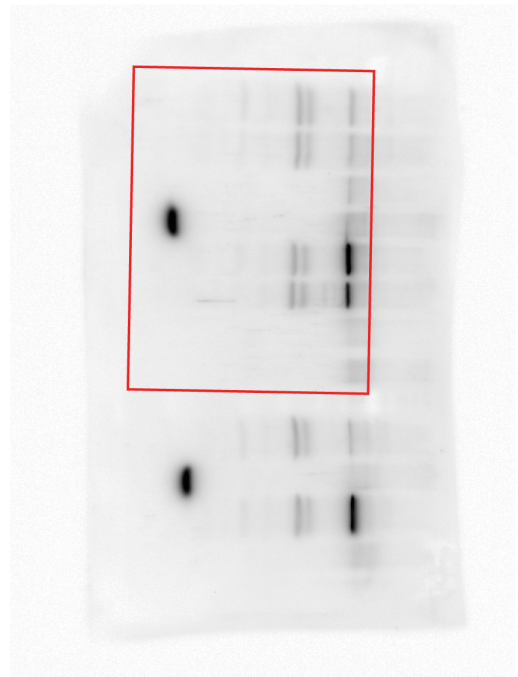


Supplementary Figure 12: Full blots of Supplementary Figure 2C

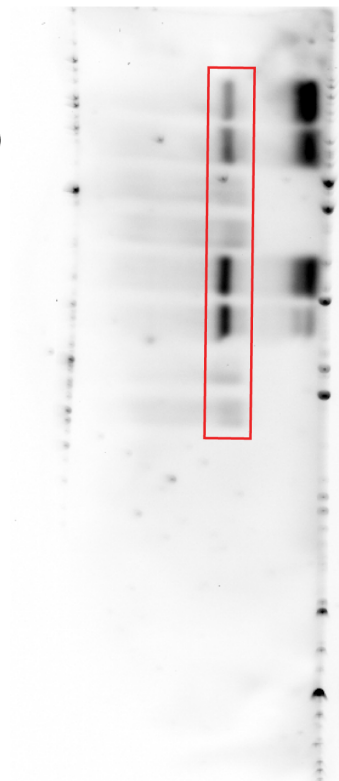
S2C EccB5



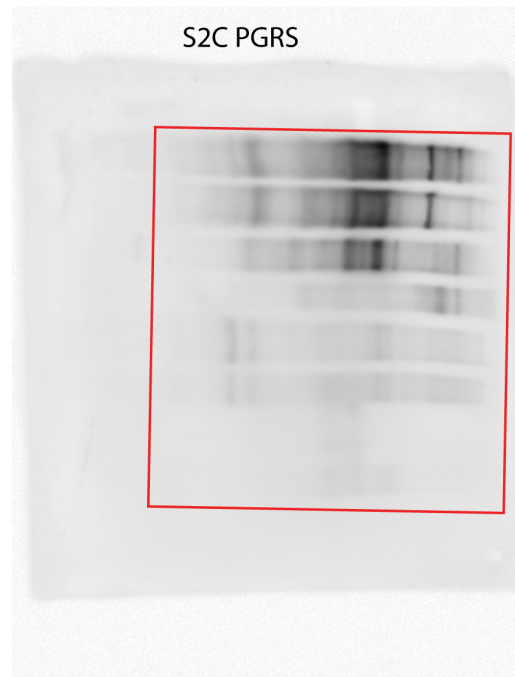
S2C HA



S2C GroEL (mirrored)



S2C PGRS



Supplementary Discussion

Nomenclature EsxX and EsxY

We propose to rename the gene *mt2420* to *esxX* and the gene *mt2421* to *esxY*. These genes correspond to the H37Ra genes *mra2374* and *mra2375* respectively ². This nomenclature is in concordance with the standardised nomenclature for TypeVII secretion substrates ³. The letters X and Y are the first available letters after the substrates EsxV and EsxW. Searches on pubmed (www.ncbi.nlm.nih.gov/pubmed) or google scholar (scholar.google.com) for EsxX or EsxY combined with Mycobacterium did not lead to any previous usages of the names, although the gene code *esxY* has been used in *Staphylococcus aureus* ⁴. We proposed that this nomenclature is assumed for all members of the *M. tuberculosis complex*, which contain orthologues of these genes.

M. marinum M^{USA} contains two copies of each of these *esx*-genes as well as many more *esx*-genes than *M. tuberculosis* ^{5,6}. Furthermore, the *ppe38* locus is differently organised in the other well-known *M. marinum* strain E11 ⁷. Therefore, we propose to maintain the current nomenclature for all *M. marinum* genes until more information about these genes might become available.

Comparison of whole-genome sequence polymorphisms of Beijing isolates

Previous studies suggested a difference in virulence between the Beijing strains SAWC_2088 and SAWC_2135 ^{8,9}. To gain insight in possible reasons of differing virulence and secretion between these strains, we investigated the whole genome sequences (WGSs) of three Beijing strains included in the present study and focused on insertions or deletions (indels) and single nucleotide polymorphisms (SNPs) in known virulence loci or differentially secreted substrates (Supplementary Figures 3A-B, Supplementary Tables 3-4). Because we were not able to culture SAWC_2088 from the original inoculum, the isolate SAWC_1945, isolated on a different timepoint from the same patient and shown to be the same strain by spoligotyping, was used for the WGS analysis instead. All three sequenced strains possessed the characteristic *pks15/1* frameshift mutation compared to H37Rv leading to the ability to produce phenolic glycolipids (PGLs) ¹⁰. Similarly, all three strains possessed the same intergenic 1bp deletion compared to H37Rv in the promoter region of *whiB7* described earlier ¹¹, which is known to be present in most clinical *M. tuberculosis* strains. No differences between indel events or SNPs were found in the ESX-1 region between the three different strains. All three strains possessed the same SNP in the ESX-1 associated gene *espA* and none of the strains had polymorphisms in *phoP*

that have been associated with virulence and ESX-1 secretion ¹². All three strains possessed the same frameshift variant in *mce1R*, which is part of the *mce1* cluster, previously shown to be involved in hypervirulence ¹³. No indel events were detected in *hspX*, *devR* or *devS*, which are implicated in *in vivo* hypervirulence ¹⁴. SNP events in these genes were also the same in all strains, or only lead to synonymous changes. From these data, we conclude that none of the analysed genetic differences are likely to result in hypervirulence of one of the strains over the other. Furthermore, no indel events were found in any of the PE or PPE proteins that were differentially secreted in a PPE38-dependant manner, excluding this alternative hypothesis to explain the secretion differences between the clinical strains.

Maximum likelihood phylogenetic analysis of the sequenced strains in conjunction to other strains from different lineages of *M. tuberculosis* and from within the Beijing lineage (Lineage 2) confirmed that SAWC_2088, SAWC_2135 and SAWC_2701 are all members of the Beijing lineage (Figure 4). Within this lineage, these strains are relatively distantly related, which is also in concordance with previous analysis based on large sequence polymorphisms (LSP) and genotype of the *ppe38-71* locus ^{2,9}. In accordance to the F31 genotype, SAWC_2088 clustered together with the early branching atypical Beijing strains. SAWC_2701 (F27 genotype, Beijing sublineage 2) branched of later, but this strain is still considered an atypical Beijing strain based on other genetic markers ^{2,9}. Strain SAWC_2135 (F29 genotype, Beijing sublineage 7) clustered with other known modern Beijing clonal clusters and strains such as those part of the W148 (Europe-Russia), Central Asia and Pacific RD150 outbreaks and belongs to the typical Beijing strains ^{2,9,15}.

BlastP and BlastN queries confirmed the F29 deletion in other Beijing strains that was suggested by our phylogenetic analysis and which was previously shown by PCR ². Most notable of these, is the known hypervirulent Beijing-W strain HN878 ^{10,16-18}. Although the presence of a full length PPE38 is predicted in this strain by automated annotation, this hypothetical protein lacks the typical PPE motif in the N-terminus and is in fact likely not produced, but a bioinformatic artefact. This is caused by the IS6110 sequence that truncates the N-terminus of PPE38 and introduces an *atg* codon 3 bp next to the original transcription start-site of *ppe38* in other *M. tuberculosis* strains. That this hypothetical protein is not produced, or at least not secreted is supported by our LC-MS/MS results, which detected no peptides of PPE38 or PPE71 in the F27 and F29 sub-family Beijing strains. This case highlights the difficulties associated with *in silico* analysis of the *ppe38-71* locus, that are due to difficulties in sequencing this region when an IS6110 element is present, as well as bio-informatic problems

caused by the use of the H37Rv genome sequence as a template which (incorrectly) only contains 1 copy of *ppe38* and no *esx*-genes or *IS6110*-element ^{2,5}.

Strain HN878 belongs to a group of hypervirulent Beijing strains, often referred to as the W-Beijing strains. These strains were previously shown to have an insertion in the pseudogenes *pks15/l* that restores these pseudogenes into two intact genes *pks15* and *pks1*, which allow these strains to produce phenolic glycolipids (PGL). The restored production of PGL was reported to lead to hypervirulence of this strain ^{10,18}, but was later shown to result in an increased mortality of infected animals due to hyperinflammatory immune responses, but not to the increases in bacterial burden in organs of animals infected with this strain ¹⁶. Since all three analysed strains possessed this *pks15/l*-restoring frameshift mutation and we show that the *ppe38*-mutation present in these strains leads to increased bacillary loads in an *in vivo* model, our data could explain this caveat.

Although deletion of the *ppe38*-locus increased virulence, multiple mutations might have occurred consecutively, exacerbating the phenotype of hypervirulence in different ways. This is supported by our finding that complementation of SAWC_2135 with *ppe38-71* resulted in a highly-reduced virulence, but not to levels of ancient-Beijing strain SAWC_2088, suggesting that further differences between these strains might also affect virulence. This could be either an attenuating mutation in SAWC_2088, or another mutation conferring increased virulence in SAWC_2135. However, the finding that the *ppe38-71* mutation, that increases virulence, occurred at the branching point of the ‘modern’ Beijing strains from the ‘ancient’ Beijing strains, suggests that this mutation contributed to the emergence of these strains. This is further suggested by the clinical data that show that all-but-one Beijing isolates in Hanekom *et al.* 2007 share *ppe38-71* truncating mutations and are therefore predicted to be hypervirulent and deficient in PE_PGRS and PPE_MPTR secretion ⁹.

Supplementary methods

Construction of plasmids

All plasmids are listed in Supplementary Table 7. Primers used for the construction of plasmids can be found in Supplementary Table 6. All plasmids containing *M. marinum* genes were constructed using *M. marinum* M^{USA} genomic DNA as template DNA. Plasmid pPPE38 was created by PCR amplification of *mmar_3661* with primers *ppe38-Fw* and *ppe38-Rv*. The PCR product was cloned into the pSMT3-LipY_{tub} plasmid replacing the *lipY* gene¹⁹, using restriction enzymes BamHI and HindIII. pPPE38-PN and pPN were constructed in the same way, using primers *ppe38-Fw* and *esxPN-Rv*, or *esxPN-Fw* and *esxPN-Rv*, respectively. The gene encoding PPE38-HA was created similarly by amplification with primers *ppe38-Fw* and *ppe38-Rv-HA* from pPPE38-PN template DNA. Since PCR amplification of *esxP_3* *esxN_4* and the intergenic region between *mmar_3661* and *esxP_3* was repeatedly unsuccessful, this product (Supplementary Methods) was synthesised (GeneArt, Life technologies). The synthesised product was cloned after PPE38-HA to create pPPE38-HA-PN. Plasmid p3581-HA was constructed by cloning the PCR-product obtained by primers *3581-Fw* and *3581-HA-Rv* into pSMT3 using XbaI and BglII. pMV::*ppe38-71* was constructed by amplification of CDC1551 genomic DNA by primers *mt2422-Fw* and *mt2419-Rv*. This PCR-product was cloned into the *pMV361::mycP₅* vector, replacing *mycP₅* by restriction enzymes EcoRI and HindIII²⁰. The relevant sequence of all constructs was checked by Sanger nucleotide sequence analysis.

Collection of clinical *M. tuberculosis* isolates

In earlier studies^{9,21}, sputum cultures were obtained from patients attending primary health care clinics in metropolitan Cape Town, South Africa. *M. tuberculosis* strains present in sputum culture were genotyped using IS6110 DNA fingerprinting²². DNA fingerprints were analysed with GelCompar software using the unweighted-pair group method, average linkages and Dice coefficients²³. Strains with an IS6110 similarity index of $\geq 65\%$ were grouped into strain lineages. Members of the Beijing lineage were identified by the absence of region of difference RD105²⁴ and were further divided into sublineages as previously described^{9,24}. The three selected strains were chosen because of their different genomic organizations of the *ppe38* genomic locus². Isolate SAWC_2088 (and SAWC_1945, which is the same strain isolated at a different timepoint from the same patient – see Supplementary Discussion) belonged to the Beijing sublineage 1 and therefore belongs to Beijing family F31; Isolate SAWC_2135 belongs to Beijing sublineage 7, corresponding to family F29; Isolate SAWC_2701 belongs to Beijing

sublineage 2, corresponding to family F27^{2,9}. Isolate SAWC_2088 was earlier described to be lowly virulent in a mouse model by Aguilar et al. 2010 (strain 1 in Aguilar et al., 2010⁸).

LC-MS/MS analysis of *M. tuberculosis* culture filtrate

M. tuberculosis strains SAWC_2088, SAWC_2135, SAWC_2701 and the SAWC_2135 and SAWC_2701 isolates complemented with pMV::*ppe38-71* were grown as duplicates in liquid culture. Cultures were grown until all strains reached exponential growth with comparable optical densities. Subsequently, cells were washed and inoculated at 0.8 OD₆₀₀/ml in 100 ml 7H9 medium without BSA and grown for another 48 hours. Culture supernatant was collected and filtered. Culture filtrate proteins were concentrated to 1-2 ml, by centrifugation in Amicon® Ultra 3K centrifugal filter devices (Millipore, IRL), followed by TCA precipitation to reach a final volume of 100 µl. Protein samples were separated by SDS-PAGE, stained with Coomassie Brilliant Blue, subjected to in gel digestion, and analyzed by LC-MS/MS as previously described²⁵. LC-MS/MS and protein identification and quantification were performed as described previously²⁵. Briefly, Peptides were separated with a 25 cm × 75 µm ID fused silica C18 column (Dr Maisch GMBH, Ammerbuch-Entringen, Germany). Peptides were trapped on a 10 mm × 100 µm ID trap column and separated at 300 nl/min in an 8–32% ACN in 0.5% HAc gradient in 90 min (120 min inject-to-inject). Eluting peptides were ionized at a potential of +2 kVa into a Q Exactive mass spectrometer (Thermo Fisher, Bremen, Germany). Intact masses were measured at resolution 70.000 (at m/z 200) in the orbitrap. MS/MS spectra (top-10 precursors) were acquired at resolution 17.500 (at m/z 200).

Data analysis was performed by searching MS/MS spectra against the UniprotMycobacterium_tuberculosis_CDC1551.FASTA file (release October 2015; 4200 entries) using MaxQuant 1.5.2.8.²⁶. Enzyme specificity was set to trypsin and up to two missed cleavages were allowed. Cysteine carboxamidomethylation was treated as fixed modification and methionine oxidation and N-terminal acetylation as variable modifications. Peptide precursor ions were searched with a maximum mass deviation of 4.5 ppm and fragment ions with a maximum mass deviation of 20 ppm. Peptide and protein identifications were filtered at an FDR of 1% using the decoy database strategy. Proteins were (label-free) quantified by spectral counting^{27,28}, *i.e.* the sum of all MS/MS spectra for each identified protein. For quantitative analysis across samples, spectral counts were normalized to the sum of the spectral counts per biological sample. Differential analysis of samples was performed using the beta-binominal test, which takes into account within- and between-sample variations, giving fold-

change values and associated p-values for all identified proteins²⁸. Protein cluster analysis was performed using hierarchical clustering in R. The protein abundances were normalized to zero mean and unit variance for each individual protein. Subsequently, the Euclidean distance measure was used for protein clustering. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE²⁹ partner repository with the dataset identifier PXD003842.

Whole genome sequencing

Whole-genome sequencing was done with Illumina technology (HiSeq2000, or MiSeq) at KAUST bioscience core laboratory. Four micrograms of DNA was used to prepare libraries for sequencing according to the manufacturer's instructions using Illumina Genomic DNA Sample preparation kits, (Illumina, Inc, San Diego, CA). Paired-end sequencing (2 x 100 bp), with ~500 bp fragment sizes which resulted in insert sizes between 350 and 550 bases. Sequence data have been deposited at the European Nucleotide Archive (<http://www.ebi.ac.uk/ena/data/view/PRJEB1440>), hosted by the EBI, under accession number: PRJEB14401, secondary accession ERP016042.

Alignment and variant detection

Data analysis was done as previously described³⁰. Briefly, low quality bases and reads were removed with Trimmomatic³¹, before aligning the reads to the *M. tuberculosis* H37Rv (NCBI Reference Sequence: NC_000962.3) and *M. tuberculosis* CDC1551 (GenBank: AE000516.2) reference genomes with three different alignment tools, namely Novoalign (Novocraft), Burrows-Wheeler Aligner (BWA)³², and SMALT³³. Alignments were viewed in Artemis^{34,35} to inspect the *ppe38* genomic region. The Genome Analysis Tool Kit (GATK) was used to detect single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels)³⁶. Variants identified in all three of the alignments were annotated and used in subsequent analysis.

Phylogenomic analysis

Concatenated sequences containing high-confidence variable sites (coding and non-coding SNPs) with respect to the *M. tuberculosis* H37Rv reference genome were used for phylogenomic analysis. Modelgenerator was used to determine the optimal nucleotide substitution model and subsequently, the general time reversal (GTR) model of substitution was applied to construct a maximum likelihood phylogeny of the isolates included in this analysis

with Randomized Axelerated Maximum Likelihood (RaxML) with 1000 bootstrap pseudo-replicates^{37–40}. Positions missing data (gaps, deletions) were not considered for the analysis. A description of the sequences used for the phylogenetic analysis can be found in Supplementary Table 8.

***De novo* assembly**

The Illumina HiSeq2000 reads for each isolate were assembled with the *de novo* assembly software package for short reads, Velvet, version 1.2.07⁴¹. Velvet was run with recommended parameters according to the documentation and the optimal kmer length for each strain was used as determined by velvetk (<http://bioinformatics.net.au/software/velvetk.shtml>). The *M. tuberculosis* H37Rv or CDC1551 genome annotation was transferred to the final assemblies using RATT⁴². Contigs containing the regions of interest were further analysed using BLAST.

***In vitro* growth curves**

Pre-cultures of the indicated strains were grown until an OD₆₀₀ of 0.5-1.0 was reached. At this point all strains were inoculated in 5ml medium in glass 25ml culture tubes at 0.05 OD₆₀₀/ml for strains grown under normoxia and in 20ml for hypoxic conditions. OD₆₀₀ was measured at the indicated timepoints in these tubes to avoid evaporation or dilution of bacteria, tubes were aerated twice per week for normoxic conditions. Four independent replicates were performed at two different starting timepoints. Two-way ANOVA was used to test overall significance (over 19 families) and Tukey's test of multiple comparisons was used to measure significance at each timepoint. Results were only deemed technically significant if $p < 0.05$ at least at two consecutive time points. Statistical analyses were performed with GraphPad Prism software (Version 6).

Lipid analysis

Lipid extraction was performed as described by Lan  elle et al 2015⁴³. Briefly, mycobacterial strains were pre-cultured in unperturbed conditions one week in 10ml 7H9 medium without Tween-80 and were subsequently inoculated in 100 ml of the same medium and grown for one more week. After gentle washing with PBS, lipids were extracted by different concentrations of 5ml Methanol:Chloroform for 24 hours each (Methanol:Chloroform, 2:1, 1:1, 1:2 respectively).

Extracted lipids were concentrated and resuspended in Chloroform at 20mg/ml. 10  l lipid concentrate was loaded on preheated silica TLC silica gel plates (TLC Silica gel 60 F₂₅₄, Merck,

Germany) and migrated by 9:1 petroleum ether 40-60°C:diethylether. Lipids were visualized by spraying with phosphomolybdic acid and subsequent heating for 20 minutes at 80°C.

Synthesized DNA product EsxP_3-N_4

Since PCR amplification of the region from the 3' of *mmar_3661* to the stop codon of *mmar_3559* was repeatedly unsuccessful, this DNA fragment was synthesised (Gene-Art, Life technologies).

Sequence:

**5'-CTACGCGTGAAAGCTGCGGGGGGTGATTGGTCGGGGTTTCGGGTTTGCGGTAT
TTGGTCGGGTCTGAATGTCTTTTGGGAAAGGGGAATAGGCAGTGACTACTCGGTT
TATGACTGACCCGCATGCGATGCGGGACATGGCGGGTCGTTTTGAGATGCATGCG
CAGACGGTCGAGGATGAGGCCCCGCAAGATGTGGGCGTCGTCGCAGAACATCGCC
GGCGCGGGCTGGAGCGGTATGGCGTCGGCGACTTCGCTGGACACCATGGGTCAG
ATGAACACGGCCTTCCGCAACATTGTGAACATGCTGCATGGGGTGCGTGACGGGT
TGGTTCGCGACGCGAACAACACTACGAGCAGCAAGAGCAGGCTTCGCAGCAGGTTT
TGGGTAGCTAGGAGGGGTTGGGGCGATGACGATCAATTACCAGTTCGGTGATGTG
GATGCCCCATGGGGCGCTGATTCGGGCTCAGGCGGGCGGCGTTGGAGGCTGAGCAT
CAGGCGATTGTGCGTGATGTGTTGGCCGCTGGGGATTTCTGGGGTGGCGCCGGGT
CGGTGGCGTGCCAGGAGTTCATCTCGCAGTTGGGCCGTAATTTCCAGGTGATCTA
CCAGCAGGCCAATGCCCACGGTCAGAAGGTGCAGGCTGCCGGTAACAACATGGC
CAGCACCGACTCCGCGGTGGGTTCGAGCTGGGCATAGAGCTTATCGATACCG-3'**

Supplementary References

1. Vizcaíno, J. A. *et al.* ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–6 (2014).
2. McEvoy, C. R. E., van Helden, P. D., Warren, R. M. & Gey van Pittius, N. C. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. *BMC Evol. Biol.* **9**, 237 (2009).
3. Bitter, W. *et al.* Systematic genetic nomenclature for type VII secretion systems. *PLoS Pathog.* **5**, e1000507 (2009).
4. Burts, M. L., Williams, W. A., DeBord, K. & Missiakas, D. M. EsxA and EsxB are secreted by an ESAT-6-like system that is required for the pathogenesis of *Staphylococcus aureus* infections. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1169–74 (2005).
5. Kapopoulou, A., Lew, J. M. & Cole, S. T. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)*. **91**, 8–13 (2011).
6. Stinear, T. P. *et al.* Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Res.* **18**, 729–41 (2008).
7. Weerdenburg, E. M. *et al.* Genome-wide transposon mutagenesis indicates that *Mycobacterium marinum* customizes its virulence mechanisms for survival and replication in different hosts. *Infect. Immun.* IAI.03050-14 (2015). doi:10.1128/IAI.03050-14
8. Aguilar, D. *et al.* *Mycobacterium tuberculosis* strains with the Beijing genotype demonstrate variability in virulence associated with transmission. *Tuberculosis (Edinb)*. **90**, 319–25 (2010).
9. Hanekom, M. *et al.* A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. *J. Clin. Microbiol.* **45**, 1483–90 (2007).
10. Reed, M. B. *et al.* A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* **431**, 84–7 (2004).
11. Solans, L. *et al.* A specific polymorphism in *Mycobacterium tuberculosis* H37Rv causes differential ESAT-6 expression and identifies WhiB6 as a novel ESX-1 component. *Infect. Immun.* **82**, 3446–56 (2014).
12. Frigui, W. *et al.* Control of *M. tuberculosis* ESAT-6 secretion and specific T cell recognition by PhoP. *PLoS Pathog.* **4**, e33 (2008).
13. Shimono, N. *et al.* Hypervirulent mutant of *Mycobacterium tuberculosis* resulting from disruption of the *mce1* operon. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15918–23 (2003).
14. Hu, Y., Movahedzadeh, F., Stoker, N. G. & Coates, A. R. M. Deletion of the *Mycobacterium tuberculosis* alpha-crystallin-like hspX gene causes increased bacterial growth in vivo. *Infect. Immun.* **74**, 861–8 (2006).
15. Merker, M. *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat. Genet.* **47**, 242–9 (2015).
16. Sinsimer, D. *et al.* The phenolic glycolipid of *Mycobacterium tuberculosis* differentially modulates the early host cytokine response but does not in itself confer hypervirulence. *Infect. Immun.* **76**, 3027–36 (2008).
17. Gopal, R. *et al.* Unexpected Role for IL-17 in Protective Immunity against Hypervirulent *Mycobacterium tuberculosis* HN878 Infection. *PLoS Pathog.* **10**, e1004099 (2014).
18. Tsenova, L. *et al.* Virulence of selected *Mycobacterium tuberculosis* clinical isolates in the rabbit model of meningitis is dependent on phenolic glycolipid produced by the bacilli. *J. Infect. Dis.* **192**, 98–106 (2005).

19. Daleke, M. H. *et al.* Conserved Pro-Glu (PE) and Pro-Pro-Glu (PPE) protein domains target LipY lipases of pathogenic mycobacteria to the cell surface via the ESX-5 pathway. *J. Biol. Chem.* **286**, 19024–34 (2011).
20. Van Winden, V. J. C. *et al.* Mycosins are required for the stabilization of the ESX-1 and ESX-5 type VII secretion membrane complexes. *MBio* **7**, 1–11 (2016).
21. van der Spuy, G. D. *et al.* Changing Mycobacterium tuberculosis population highlights clade-specific pathogenic characteristics. *Tuberculosis* **89**, 120–125 (2009).
22. van Embden, J. D. *et al.* Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.* **31**, 406–9 (1993).
23. Hermans, P. W. *et al.* Analysis of the population structure of Mycobacterium tuberculosis in Ethiopia, Tunisia, and The Netherlands: usefulness of DNA typing for global tuberculosis epidemiology. *J. Infect. Dis.* **171**, 1504–13 (1995).
24. Tsolaki, A. G. *et al.* Genomic Deletions Classify the Beijing/W Strains as a Distinct Genetic Lineage of Mycobacterium tuberculosis. *J. Clin. Microbiol.* **43**, 3185–3191 (2005).
25. Piersma, S. R. *et al.* Whole gel processing procedure for GeLC-MS/MS based proteomics. *Proteome Sci.* **11**, 17 (2013).
26. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–72 (2008).
27. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–201 (2004).
28. Pham, T. V., Piersma, S. R., Warmoes, M. & Jimenez, C. R. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* **26**, 363–9 (2010).
29. Vizcaino, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–56 (2016).
30. Dippenaar, A. *et al.* Whole genome sequence analysis of Mycobacterium suricattae. *Tuberculosis (Edinb)*. **95**, 682–8 (2015).
31. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–20 (2014).
32. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **3** (2013).
33. Ponstingl, H. & Ning, Z. <p>SMALT – A new mapper for DNA sequencing reads</p>. *F1000Research* **1**, (2010).
34. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–5 (2000).
35. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2011).
36. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
37. J., F. Notices. *Cladistics* **5**, 163–166 (1989).
38. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McInerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).
39. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–3 (2014).

40. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–90 (2006).
41. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–9 (2008).
42. Otto, T. D., Dillon, G. P., Degraeve, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* **39**, e57 (2011).
43. Lan  elle, M.-A., Nigou, J. & Daff  , M. in *Methods in molecular biology (Clifton, N.J.)* **1285**, 77–103 (2015).
44. Abdallah, A. M. *et al.* PPE and PE_PGRS proteins of *Mycobacterium marinum* are transported via the type VII secretion system ESX-5. *Mol. Microbiol.* **73**, 329–40 (2009).
45. Ates, L. S. *et al.* The ESX-5 System of Pathogenic *Mycobacteria* Is Involved In Capsule Integrity and Virulence through Its Substrate PPE10. *PLoS Pathog.* **12**, (2016).
46. Comas, I. *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503 (2010).
47. Coscolla, M. *et al.* Novel *Mycobacterium tuberculosis* complex isolate from a wild chimpanzee. *Emerg. Infect. Dis.* **19**, 969–76 (2013).
48. Lee, R. S. *et al.* Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proc. Natl. Acad. Sci.* **2000**, 201507071 (2015).