Supplementary Information

Content:

- **Supplementary Tables 1 - 9**

Supplementary Table 1. Pairwise F_{ST} values between MTBC Lineage 4 sublineages. Numbers in brackets are values with p>=0.05. All other values were statistically significant (p<0.05). P-values obtained by permutation (see Online Methods).

	L4.1.1	L4.1.3	L4.1.2	L4.2	L4.3	L4.4	L4.5	L4.6.1	L4.6.2	L4.10
L4.1.1	0									
L4.1.3	0.408	0								
L4.1.2	0.432	0.683	0							
L4.2	0.387	0.448	0.527	0						
L4.3	0.510	0.631	0.630	0.413	0					
L4.4	0.417	(0.57)	0.590	0.330	0.430	0				
L4.5	0.471	(0.63)	0.649	0.387	0.500	0.389	0			
L4.6.1	0.486	0.679	0.660	0.414	0.510	0.411	0.471	0		
L4.6.2	0.523	(0.67)	0.692	0.455	0.563	0.471	0.509	0.508	0	
L4.10	0.487	0.588	0.594	0.429	0.482	0.389	0.437	0.427	0.499	0

Supplementary Table 2. Sublineage-specific SNPs and oligonucleotides used for MOL-PCR.

MTBC Sublineage	Alternative Name	SNP	Reference Allele	Mutant Allele	Gene	Ess/Syn	Sense	Oligonucleotide	Bead Region	Sequence
L4.1.1	X	3798451	C	G	Rv3383c (idsB)	noness/syn	sense	LPO_ancestral	29	GGGTTCCCTAAGGGTTGGATACTACTTCTATAACTCACTTAAAAATGCTTTCCCAAGGTCAGCAGGGACACTC
								LPO_mutant	28	GGGTTCCCTAAGGGTTGGACACTTAATTCATTCTAAATCTATCAATGCTTTCCCAAGGTCAGCAGGGACACTG
								RPO	-	P-CCCAGAAAAGCCGCATCCAGAGTCAATATCTAGATTGGATCTTGCTGGCAC
L4.1.3	Ghana	4409231	T	G	Rv3921c	ess/nonsyn	sense	LPO_ancestral	27	GGGTTCCCTAAGGGTTGGATAACTTACACTTAACTATCATCTTCCTGCTTTTTGGCCTCCTCCT
								LPO_mutant	12	GGGTTCCCTAAGGGTTGGACATAATCAATTTCAACTTTCTACTCCTGCTTTTTGGCCTCCTCCG
								RPO	-	P–CCTTTTCGATCATGCCGAAGACGTAATGCTCTAGATTGGATCTTGCTGGCAC
L4.1.2	Haarlem	3013784	C	G	Rv2697c (dut)	ess/nonsyn	anti-sense	LPO_ancestral	14	GGGTTCCCTAAGGGTTGGAAATTTCTTCTCTTTCTTTCACAATAGTTGCTAGTGCAACGGGTTGAGTTGG
								LPO_mutant	30	GGGTTCCCTAAGGGTTGGACTTAACATTTAACTTCTATAACACAGTTGCTAGTGCAACGGGTTGAGTTGC
								RPO	-	P-TCGAGCTGGTCGAGGTCTCGTCGTTTCTAGATTGGATCTTGCTGGCAC
L4.2		2181026	G	C	Rv1928c	noness/syn	sense	LPO_ancestral	33	GGGTTCCCTAAGGGTTGGAACTACTTATTCTCAAACTCTAATATGCTGGCTCACATCGCAGCAGACG
								LPO_mutant	15	GGGTTCCCTAAGGGTTGGATACTTCTTTACTACAATTTACAACTGCTGGCTCACATCGCAGCAGACC
								RPO	-	P-GGCACGACCTTGCCACCTGATGTTCTAGATTGGATCTTGCTGGCAC
L4.3	LAM	1480024	G	T	Rv1318c	noness/nonsyn	anti-sense	LPO_ancestral	20	GGGTTCCCTAAGGGTTGGACTTTCTCATACTTTCAACTAATTTGCTGATCATCTCGATGGTCACATTGGTGTTC
								LPO_mutant	21	GGGTTCCCTAAGGGTTGGATCAAACTCTCAATTCTTACTTAATGCTGATCATCTCGATGGTCACATTGGTGTTA
								RPO	-	P-GGGTTCATCCTGATGTGGATCCTGGCCTTCTAGATTGGATCTTGCTGGCAC
L4.4		3966059	G	C	Rv3529c	noness/nonsyn	anti-sense	LPO_ancestral	36	GGGTTCCCTAAGGGTTGGAATTAAACAACTCTTAACTACACAAACTTGATTGCCGATCCGCTGGGTAC
								LPO_mutant	37	GGGTTCCCTAAGGGTTGGATACAACATCTCATTAACATATACAACTTGATTGCCGATCCGCTGGGTAG
								RPO	-	P-GGTGGCAGATATCTACCGGCACTTCGTCTAGATTGGATCTTGCTGGCAC
L4.5		2789341	A	C	Rv2483c (plsC)	ess/nonsyn	anti-sense	LPO_ancestral	18	GGGTTCCCTAAGGGTTGGAACACTTATCTTTCAATTCAATTACATCGCCGAAGGCCAAACCCAGCGAATCTAAGAT
								LPO_mutant	22	GGGTTCCCTAAGGGTTGGACAAACAAACATTCAAATATCAATCATCGCCGAAGGCCAAACCCAGCGAATCTAAGAG
								RPO	-	P-CGCTGGCAAGGATGGTGAGGCCTCCGCATCGCCAAGCTCTATCTAGATTGGATCTTGCTGGCAC
L4.6.1	Uganda	990626	T	A	Rv0890c	noness/nonsyn	sense	LPO_ancestral	34	GGGTTCCCTAAGGGTTGGAACTTATTTCTTCACTACTATATCAATCGCATCACCTCCTGCCAGGGCT
								LPO_mutant	35	GGGTTCCCTAAGGGTTGGACATCTTCATATCAATTCTCTTATTATCGCATCACCTCCTGCCAGGGCA
								RPO	-	P-AACTGCGCCATCAGGACCTGGTGCATTCTAGATTGGATCTTGCTGGCAC
L4.6.2	Cameroon	3191099	C	A	Rv2881c (cdsA)	ess/nonsyn	anti-sense	LPO_ancestral	19	GGGTTCCCTAAGGGTTGGAATACTTTACAAACAAATAACACACCGCAATGCTGGTCTACCCGGAAAATG
								LPO_mutant	25	GGGTTCCCTAAGGGTTGGACTTTCTTAATACATTACAACATACCGCAATGCTGGTCTACCCGGAAAATT
								RPO	-	P-GCTCGGGATGGGTGTTCTGCATGATGATTCTAGATTGGATCTTGCTGGCAC
L4.10	PGG3	1692141	C	A	Rv1501	noness/syn	sense	LPO_ancestral	13	GGGTTCCCTAAGGGTTGGACAAATACATAATCTTACATTCACTCGACTCATGATGAAGTATGACCCTCATTTC-TTTACCTTTCTTGAAATC
								LPO_mutant	26	GGGTTCCCTAAGGGTTGGATACATTCAACACTCTTAAATCAAACGACTCATGATGAAGTATGACCCTCATTTC-TTTACCTTTCTTGAAATA
								RPO	-	P-CCCGAAGTCCTAAGCATCGTTGATCGTGTGCTATCTGAAACTCTAGATTGGATCTTGCTGGCAC

Supplementary Table 3. Lineage 4-specific and sublineage-specific SNPs interrogated with Sequenom MassARRAY.

Sublineage	Genomic position of SNP in H37Rv	Ancestral base	Mutant base	Gene (nt gene)
L4.1.1 (X)	1960391	G	A	Rv1733c_0097n
L4.1.1 (X)	2603797	G	A	Rv2330c_0426s
L4.1.1 (X)	3597737	C	T	Rv3221c_0030s
L4.1.2 (Haarlem)	4352475	G	A	Rv3874_0202n
L4.1.2 (Haarlem)	891756	A	G	Rv0798c_0514s
L4.1.2 (Haarlem)	1477588	G	C	Rv1316c_0044n
L4.3 (LAM)	157292	C	T	Rv0129c_0309s
L4.3 (LAM)	2134215	T	C	Rv1884c_0047n
L4.4*	3311442	G	A	Rv2958c_0559n
L4.5	7892	G	A	Rv0006_0591s
L4.5	12555	C	T	Rv0009_0088s
L4.10 (PGG3)	7585	C	G	Rv0006_0284n
L4.10 (PGG3)	1960284	A	C	Rv1733c_0204n

* not specific for all strains of L4.4, but only for a subsublineage of L4.4

Supplementary Table 4. PCR primers for amplification and Sanger sequencing of regions flanking Lineage 4 sublineage-specific SNPs.

Sublineage	SNP	Forward Primer	Reverse Primer	Amplification Fragment Length
L4.1.1	3798451	TTGTGCACCAACTCCACAGCCG	CGTGTCTTTCTGTAGTGGATGACC	518bp
L4.1.3	4409231	AGGATTGTCAACGTTTGCGT	GGATGCGTTCGTCGATTTCAG	563bp
L4.1.2	3013784	GCGTGTCCGGCCTTGCGTTTG	CCGGCGTTGATCTCTACAGC	522bp
L4.2	2181026	CGCCTTGGAGGCGCAGTAGTGG	GGCATGTGATTCCATCAGGTATC	557bp
L4.3	1480024	GCCGGCTGGTCACCAATTGCGTC	GTTCGCCGCCCAGGCGCTCGAG	542bp
L4.4	3966059	CGGGCAAATTGCGTATCTGC	GGTACTAAAGAATCCGAGTCATC	531bp
L4.5	2789341	TAGAACGGTCCTCGCCAGATTG	GCAACTCCACCACGATCAATC	656bp
L4.6.1	990626	CGGACACCTTCGGAGTGACTG	GCCCAGGTGCTGGCGTATTGC	500bp
L4.6.2	3191099	CACGTTTGACCTGCGACTCCAC	CCTTGGTCGCTACCCATGAG	603bp
L4.10	1692141	AGGTGAATAAGCGTAGCATGATTC	GCGCGAGGTAGGTATGGTCC	540bp

Supplementary Table 5. Distribution of MTBC Lineage 4 clinical isolates per country and per sublineage.

See separate Excel file.

Supplementary Table 6. Number of fixed nonsynonymous SNPs (nsSNPs) predicted be functional among generalist and specialist sublineages.

Gene Category[1]	# nsSNPs Generalists	# nsSNPs Specialists
Cell Wall & Cell Processes	15	10
Information Pathways	7	9
Lipid Metabolism	6	14
Regulatory Proteins	1	2
Virulence	3	2

[1] Gene categories were defined as in http://tuberculist.epfl.ch/.

Supplementary Table 7. Whole genome sequence data accession codes.

See separate Excel file.

Supplementary Table 8. Characteristics of genetic diversity found in the generalist and specialist sublineages.

Sublineage	Number of strains	Number of mutations[1]	SS[2]	Categories	nsSNP	sSNP	nsSNP/sSNP	χ^2	p-value	NS mean pw dist[3]	S mean pw dist[3]
L4.6.1/Uganda	203	8054	7567.00	Epitopes	28	26	1.08	1.61	0.203	0.35	0.62
				Nonepitopes	389	243	1.60			7.8	6.25
				Essentials	1027	700	1.47	14.71	0.000	23.9	15.89
				Nonessentials	3104	1694	1.83			63.54	35.45
L4.10/PGG3	301	25678	25192.00	Epitopes	123	63	1.95	0.41	0.52	1.82	1.33
				Nonepitopes	1261	726	1.74			16.91	7.89
				Essentials	3450	2329	1.48	29.25	0.000	42.88	27.60
				Nonessentials	10234	5826	1.76			116.52	69.65
L4.3/LAM	293	19714	18930.00	Epitopes	77	43	1.79	0.09	0.755	1.22	0.43
				Nonepitopes	915	540	1.69			15.3	12.89
				Essentials	2597	1680	1.55	11.05	0.001	37.07	27.49
				Nonessentials	7544	4347	1.74			121.89	68.62
L4.1.2/Haarlem	228	15567	15108.00	Epitopes	61	45	1.36	1.12	0.29	0.97	1.12
				Nonepitopes	730	424	1.72			11.08	5.59
				Essentials	2064	1441	1.43	20.48	0.000	30.27	20.90
				Nonessentials	6041	3511	1.72			85.00	50.42

[1] Number of SNP with respect to the reconstructed ancestor sequence of MTBC.

[2] Number of polymorphic sites within each sublineage.

[3] Number of mean pairwise distances for nonsynonymous (NS) and synonymous (S) mutations.

Supplementary Table 9. Description of epitopes containing nonsingleton nonsynonymous mutations in the four sublineages analysed.

See separate Excel file.