

Supplementary Information for the manuscript

Analysis of Complete Genomes of *Mycobacterium tuberculosis* sublineage 2.1 (Proto-Beijing) Revealed the Presence of Three *pe_pgrs3-pe_pgrs4*-like Genes.

Olabisi Flora Davies-Bolorunduro^{1,2,4}, Bharkbhoom Jaemsai², Wuthiwat Ruangchai¹, Thanakron Noppanamas¹, Manon Boonbangyang¹, Thavin Bodharamik¹, Waritta Sawaengdee³, Surakameth Mahasirimongkol³, Prasit Palittapongarnpim^{1,2}

¹Pornchai Matangkasombut Center for Microbial Genomics, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok 10400 Thailand

²Department of Microbiology, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok 10400 Thailand

³Medical Life Science Institute, Department of Medical Sciences, Ministry of Public Health, Thailand.

⁴Floret Center for Advanced Genomics and Bioinformatics Research, Lagos, Nigeria.

Corresponding author: **Prasit Palittapongarnpim**

Pornchai Matangkasombut Center for Microbial Genomics, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok 10400 Thailand. E-mail: Prasit.pal@mahidol.ac.th

Supplementary Table 1. Details of isolates belonging to L2.1, L2.2.AA3 and H37Rv in this study.

Isolates	Sequencing Method	Genome Coverage	Assembly Method	Annotation Pipeline	Country	Collection Date	References
Mtb CR 170941 [CP104271.1]	-Oxford Nanopore - Illumina	38.0X	-Flye v2.9 -Minimap2 v2 -Pilon v.1.24	PGAP	Thailand	1/30/2020	This study
Mtb N0031 [NZ_CP069076.1]	-PacBio SMRT -Oxford Nanopore	50X	-Flye v.2.5	PGAP	China	6/15/1994	[12]
Mtb Sea1411716C4 [NZ_CP041797.1]	PacBio RSII	153X	Canu v1.6	PGAP	Sweden	Missing	[13]
Mtb 2279 [CP010336]	PacBio SMRT	97X	HGAP v3	Not available	China	10/25/2006	[15]
Mtb 18b [NZ_CP007299.1]	-IonTorrent -PacBio -454 -Illumina	80x (2kb PE) ^a	-HGAP v. 2.0 -Newbler v. 2.6 -MIRA v. 3.9.15	PGAP	Japan	Not available	[38]

^aGenome coverage for 454 sequencing method is 80x (2kb paired-end library).

Supplementary Table 2: Summary of structural variants in *pe-ppe* regions described in this study.

Genomic region containing <i>pe-ppe</i>	Possible <i>pe-ppe</i> within the region	Variation type	Affected genes	Affected isolates or Genotypes
<i>ppe24 - pe_pgrs31</i>	<i>ppe24, wag22, pe_pgrs31</i>	deletion	<i>wag22</i>	All L2.1 and L2.2.AA3
<i>plcB - ppe39</i>	<i>ppe38, ppe39</i>	deletion	<i>ppe38, ppe39</i>	All L2.1
<i>ppe46 - pe29</i>	<i>ppe46, pe27A, esxS, ppe47, ppe48, pe29</i>	deletion	<i>pe27A, esxS, ppe47, ppe48</i>	Two L2.1 isolates (Mtb N0031 and CR170941)
<i>Rv3134 - ppe51</i>	<i>ppe50, ppe51</i>	deletion	<i>ppe50</i>	All L2.1 and <u>Mtb2279*</u>
<i>Rv3737 - ppe67</i>	<i>ppe66, ppe67</i>	deletion	<i>ppe66</i>	All L2.1
<i>pe-pgrs3-pe-pgrs4-ppe3</i>	<i>pe-pgrs3-pe-pgrs4, †pe-pgrs3*-†pe-pgrs4*</i>	deletion	<i>pe_pgrs3</i>	All L2.1 and L2.2.AA3
		deletion [†]	<i>pe_pgrs4</i>	L2.2.AA3
<i>ppe53-moxR3</i>	<i>ppe53, †ppe53A, ††ppe53B</i>	deletion [†]	<i>ppe53</i>	All L2.1 and L2.2.AA3

**ppe50e* was identified in Mtb18b.

†*pgrs3*-pe-pgrs4** replaced *pe_pgrs3*; ††*ppe53A* and *ppe53B* replaced *ppe53* in all L2.1 and L2.2.AA3 but not found in H37Rv (NC_000962.3)

Supplementary Table 3: Pairwise SNP distances between *pe_pgrs3-pe_pgrs4*-like genes among 6 isolates belonging to L2.1, L2.2.AA3 and H37Rv.

	pe_pgrs4-H37Rv	pe_pgrs4-MtbSea_L2.1	pe_pgrs4-MtbN0031_L2.1	pe_pgrs4-MtbCR_L2.1	pe_pgrs3*-Mtb2279_L2.2	pe_pgrs3*-Mtb18b_L2.2	pe_pgrs3*-MtbN0031_L2.1	pe_pgrs3*-MtbSea-L2.1	pe_pgrs3*-MtbCR_L2.1	pe_pgrs3-H37Rv	pe_pgrs4*-Mtb2279_L2.2	pe_pgrs4*-Mtb18b_L2.2	pe_pgrs4*-MtbCR_L2.1	pe_pgrs4*-MtbSea_L2.1	pe_pgrs4*-MtbN0031_L2.1
pe_pgrs4-H37Rv	Group1														
pe_pgrs4-MtbSea_L2.1	25														
pe_pgrs4-MtbN0031_L2.1	24	1													
pe_pgrs4-MtbCR_L2.1	24	1	0												
pe_pgrs3*-Mtb2279_L2.2	414	347	346	346	Group2										
pe_pgrs3*-Mtb18b_L2.2	457	379	378	378	7										
pe_pgrs3*-MtbN0031_L2.1	458	380	379	379	10	3									
pe_pgrs3*-MtbSea-L2.1	458	380	379	379	10	3	0								
pe_pgrs3*-MtbCR_L2.1	458	380	379	379	12	5	2	2							
pe_pgrs3-H37Rv	449	397	396	396	275	273	272	272	274	Group3					
pe_pgrs4*-Mtb2279_L2.2	217	172	172	172	260	258	257	257	258	39					
pe_pgrs4*-Mtb18b_L2.2	417	364	363	363	438	466	465	465	465	247	1				
pe_pgrs4*-MtbCR_L2.1	447	398	397	397	452	480	479	479	477	216	37	37			
pe_pgrs4*-MtbSea_L2.1	446	397	396	396	451	479	478	478	478	214	36	35	4		
pe_pgrs4*-MtbN0031_L2.1	446	397	396	396	451	479	479	479	479	214	36	35	4	2	

Supplementary Table 4: Pairwise SNP distances between *pe_pgrs3-pe_pgrs4*-like genes among 9 isolates belonging to L2.1, L2.2.AA3, H37Rv, *M. bovis* and *M. bovis* BCG

	pe_pgrs4-McanCIPT	pe_pgrs4-H37Rv	pe_pgrs4-MtbSea_L2.1	pe_pgrs4-MtbN0031_L2.1	pe_pgrs4-MtbCR_L2.1	pe_pgrs4-MbovBCGTok	pe_pgrs4-MbovBCGpast	pe_pgrs4-MbovAF212	pe_pgrs3star-McanCIPT	pe_pgrs3star-MbovBCGpast	pe_pgrs3star-MbovBCGTok	pe_pgrs3star-MbovAF212	pe_pgrs3star-Mtb2279_L2.2	pe_pgrs3star-Mtb18b_L2.2	pe_pgrs3star-MtbN0031_L2.1	pe_pgrs3star-MtbSea_L2.1	pe_pgrs3star-MtbCR_L2.1	pe_pgrs3-H37Rv	pe_pgrs4-MtbCR_L2.1	pe_pgrs4-MbovBCGpast	pe_pgrs4-MbovBCGTok	pe_pgrs4-MbovAF212	pe_pgrs4-MtbSea_L2.1	pe_pgrs4-MtbN0031_L2.1	pe_pgrs4-McanCIPT	pe_pgrs4-Mtb18b_L2.2	pe_pgrs4-Mtb2279_L2.2
pe_pgrs4-McanCIPT	Group1																										
pe_pgrs4-H37Rv	96																										
pe_pgrs4-MtbSea_L2.1	75	25																									
pe_pgrs4-MtbN0031_L2.1	74	24	1																								
pe_pgrs4-MtbCR_L2.1	74	24	1	0																							
pe_pgrs4-MbovBCGTok	80	32	12	11	11																						
pe_pgrs4-MbovBCGpast	78	30	10	9	9	2																					
pe_pgrs4-MbovAF212	77	31	10	9	9	3	1																				
pe_pgrs3star-McanCIPT	372	372	318	317	317	364	364	359	Group2																		
pe_pgrs3star-MbovBCGpast	439	452	377	376	376	432	432	427	38																		
pe_pgrs3star-MbovBCGTok	440	453	378	377	377	433	433	428	38	1																	
pe_pgrs3star-MbovAF212	408	421	347	346	346	402	402	397	36	4	4																
pe_pgrs3star-Mtb2279_L2.2	437	447	371	370	370	426	426	421	34	35	36	31															
pe_pgrs3star-Mtb18b_L2.2	450	463	388	387	387	443	443	438	32	15	16	15	22														
pe_pgrs3star-MtbN0031_L2.1	451	464	389	388	388	444	444	439	31	14	15	14	25	3													
pe_pgrs3star-MtbSea-L2.1	451	464	389	388	388	444	444	439	31	14	15	14	25	3	0												
pe_pgrs3star-MtbCR_L2.1	451	464	389	388	388	444	444	439	33	16	17	16	27	5	2	2											
pe_pgrs3-H37Rv	415	454	401	400	400	446	446	441	220	280	280	251	291	274	273	273	275	Group3									
pe_pgrs4-MtbCR_L2.1	413	443	393	392	392	439	437	431	385	487	487	460	461	483	482	482	480	220									
pe_pgrs4-MbovBCGpast	414	444	394	393	393	440	438	432	384	487	487	460	461	483	482	482	482	218	5								
pe_pgrs4-MbovBCGTok	414	444	394	393	393	440	438	432	384	487	487	460	461	483	482	482	482	218	5	0							
pe_pgrs4-MbovAF212	417	447	394	393	393	443	441	439	381	486	486	459	459	481	482	482	482	232	19	14	14						
pe_pgrs4-MtbSea_L2.1	412	442	392	391	391	438	436	430	383	486	486	459	460	482	481	481	481	218	4	3	3	17					
pe_pgrs4-MtbN0031_L2.1	412	442	392	391	391	438	436	430	384	487	487	460	460	482	482	482	482	218	4	3	3	16	2				
pe_pgrs4-McanCIPT	404	434	379	378	378	428	426	420	379	468	468	439	458	480	479	479	479	249	36	35	35	49	34	34			
pe_pgrs4-Mtb18b_L2.2	407	413	359	358	358	411	409	403	368	457	457	428	447	469	468	468	468	250	36	35	35	49	34	34	28		
pe_pgrs4-Mtb2279_L2.2	227	217	171	171	171	222	222	218	193	246	246	217	261	258	257	257	258	38	36	34	34	48	35	35	27	1	

Supplementary Table 5: Pairwise SNP distances between *ppe53A* and *ppe53B* in Mtb 18b, Mtb N0031 and Mtb Sea14117p6c4 and the one in H37Rv.

	ppe53B_Mtb_18b	ppe53B_Mtb_N0031	ppe53B_Mtb_SEA	ppe53_H37Rv	ppe53A_Mtb_N0031	ppe53A_Mtb_SEA	ppe53A_Mtb_18b
ppe53B_Mtb_18b							
ppe53B_Mtb_N0031	1						
ppe53B_Mtb_SEA	1	0					
ppe53_H37Rv	366	366	366				
ppe53A_Mtb_N0031	366	366	366	2			
ppe53A_Mtb_SEA	366	366	366	2	0		
ppe53A_Mtb_18b	365	365	365	1	1	1	

ppe53 of Mtb CR170941 and Mtb 2279 were not included in the comparison. A copy of *IS6110* was inserted in the *ppe53B* of Mtb CR170941, shortening the gene. The sequence of *ppe53B* of Mtb 2279 were likely to be affected by a homopolymeric run error in the long-read sequencing, resulting in the split of the gene.

Supplementary Table 6: The positions in the H37Rv genome (NC_000962.3) of the *pe-ppe* regions with identified structural variants described in this study.

<i>pe-ppe</i> region	Position in H37Rv	Length in H37Rv (bp)
<i>ppe24 - pe_pgrs31</i>	1,981,614-2,002,470	20,857
<i>plcB - ppe39</i>	2,628,781-2,635,592	6,812
<i>ppe46 - pe29</i>	3,376,939-3,380,993	4,055
<i>Rv3134 - ppe51</i>	3,499,943-3,502,936	2,994
<i>Rv3737 - ppe67</i>	4,187,699-4,190,517	2,819
<i>pe-pgrs3-pe-pgrs4-ppe3</i>	333,437-340,974	7,538
<i>ppe53-moxR3</i>	3,527,391-3,533,905	6,515

The genes are listed according to the Figures for *pe-ppe* regions in the main text.