scientific reports

Check for updates

OPEN

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 Jaemsai[2], Wuthiwat Ruangchai[1], Thanakron Noppanamas[1], Manon Boonbangyang[1], Thavin Bodharamik[1], Waritta Sawaengdee[3], Surakameth Mahasirimongkol[3] & Prasit Palittapongarnpim[1,2]✉

Mycobacterium tuberculosis **Complex (MTBC), the etiological agent of tuberculosis (TB), demonstrates considerable genotypic diversity with distinct geographic distributions and variable virulence profiles. The *pe-ppe* gene family is especially noteworthy for its extensive variability and roles in host immune response modulation and virulence enhancement. We sequenced an Mtb genotype L2.1 isolate from Chiangrai, Northern Thailand, using second and third-generation sequencing technologies. Comparative genomic analysis with two additional L2.1 isolates and two L2.2.AA3 (Asia Ancestral 3 Beijing) isolates revealed significant *pe-ppe* gene variations. Notably, all L2.1 isolates harbored three copies of *pe_pgrs3-pe_pgrs4-like* genes (*pe_pgrs3**, *pe_pgrs4**, and *pe_pgrs4*), different from L2.2.AA3 and H37Rv strains. Additionally, *ppe53* was duplicated in all but H37Rv, and *ppe50* was deleted in L2.1 isolates, contrasting with an extended *ppe50* in an L2.2 isolate (Mtb 18b), which contains an additional SVP motif. Complete deletion of *ppe66* and loss of *wag22* were observed in L2.1 isolates. These findings highlight the high structural variability of the *pe-ppe* gene family, emphasizing its complex roles in Mtb-host immune interactions. This genetic complexity offers potentially critical insights into mycobacterial pathogenesis, with significant implications for vaccine development and diagnostics.**

Keywords *Mycobacterium tuberculosis*, *Pe-ppe*, *pe_pgrs3*, *pe_pgrs4*, Lineage 2.1, Proto-Beijing genotype

Tuberculosis (TB) is caused by *Mycobacterium tuberculosis* Complex (MTBC) which is a group of closely related bacteria. These bacteria comprise several different phylogenetic lineages, nine of which are adapted to humans whilst several others are adapted to different animal species[1–3]. Some of the human-adapted lineages of the MTBC are found in specific geographical regions such as L5, L6, and L7 while others are distributed geographically around the world such as Lineages 2 and 4[4–6].

The highly prevalent Lineage 2 (L2) (East Asian) strains have been linked to increased virulence[7,8]. This lineage is further split into two main sublineages, L2.1 (Proto-Beijing) and L2.2 (Beijing)[2]. The L2.2 spreads and contributes significantly to the global TB burden, especially in Asia where the prevalence of TB drug-resistance is highest[2,9,10]. In contrast, L2.1, the most basal sublineage of L2, has been the cause of a few reported cases in Japan[11] as well as a low percentage of cases in Southern China[10] and Southeast Asia[12–14]. In Thailand, the frequency of isolates belonging to L2.1 was 4.8% of all L2[13]. Notably, certain L2.1 isolates exhibit extensive-drug resistance[13,15], and have been clonally expanding, thus become a public health concern. The variations of geographical distributions and virulence profiles between genotypes have been attributed to both single nucleotide polymorphisms (SNPs) and structural variants. A family of highly variable genes is the *pe-ppe* gene

[1]Pornchai Matangkasombut Center for Microbial Genomics, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok 10400, Thailand. [2]Department of Microbiology, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok 10400, Thailand. [3]Department of Medical Sciences, Medical Life Science Institute, Ministry of Public Health, Nonthaburi, Thailand. [4]Floret Center for Advanced Genomics and Bioinformatics Research, Lagos, Nigeria. ✉email: prasit.pal@mahidol.ac.th

family. However, there is currently limited information available on the structural variants of *pe-ppe* genes that are specific to various sublineages, including Mtb L2.1.

The structural variants of some *pe-ppe* has been identified by genotyping with Large Sequence Polymorphism (LSP), which is based on the presence or absence of Regions of Difference (RD)[16–18]. L2 is characterized by the deletion of RD105 which is longer in L2.1 (designated as RD105ext) than L2.2, while RD207 additionally characterize L2.2 and RD181 characterize all L2.2 sublineages except L2.2.2 or L2.2.AA1 (Asia Ancestral 1)[2]. The deleted regions may contribute to the differences in virulence among various sublineages of Mtb[19].

pe-ppe is a highly polymorphic family of genes contributed to about 7% of the Mtb genome[20–22]. They are unique to the mycobacterial species and are named after their conserved N-terminal proline-glutamic acid (PE) or proline-proline-glutamic acid (PPE) domains[23].

These genes play important roles in the virulence of Mtb by interacting with host cells and modulating the host immune response[24]. They are also highly immunogenic and have been investigated as potential targets for TB vaccine development[23]. PE-PPE proteins can aid Mtb pathogenesis by negatively influencing host immunity in a TLR-dependent manner leading to apoptosis[25,26], which may be influenced by structural variations of PE/PPE. For example, deletions in the PGRS domain from PE_PGRS proteins inhibited the apoptosis[27,28].

Studies on variations of PE-PPEs present formidable challenges due to their abundance and homology, hindering precise determination of variants, functions and interactions[20,29]. Additionally, the high GC-content, reaching up to 80% in some *pe-ppe* genes, poses difficulties in sequencing, sequence mapping, and cloning[30]. The SNPs in *pe-ppe* genes, are, therefore, typically excluded from WGS studies and their structural variations remain to be completely characterized.

In this study, we sequenced an isolate of Mtb L2.1 in Chiangrai, Northern Thailand by both second and third-generation sequencing and assembled a complete genome. Comparison of the complete genome with the ones of two other L2.1 isolates (Mtb N0031 [NZ_CP069076.1] and Mtb Sea14117p6c4 [NZ_CP041797.1]) and two isolates of L2.2.AA3 (Asia Ancestral 3) (Mtb 18b [NZ_CP007299.1] and Mtb 2279 [CP010336.1]), available in NCBI identified a few missing *pe-ppe* genes common to the L2.1 isolates. Interestingly, we identified three, instead of two, contiguous genes similar to *pe_pgrs3* and *pe_pgrs4*, which provides insights into the variations of this highly polymorphic region.

Results

Sequencing and hybrid assembly of L2.1 and L2.2.AA3 isolates

An isolate, Mtb CR170941 was previously cultured from a 73-year-old male pulmonary TB patient in Chiangrai, northern Thailand in 2020. The isolate was sensitive to both isoniazid and rifampicin and was previously identified to be Mtb L2.1[2]. The complete genome of the isolate was assembled by combining Oxford Nanopore and Illumina NGS data.

We also analyzed two other complete genomes of L2.1 isolates available in National Center of Biotechnology Information (NCBI): Mtb N0031 and Mtb Sea14117p6c4. The former was collected in China in 1994 and the latter was from Sweden with unknown date of collection. Two other L2.2 isolates, belonging to the Ancestral Beijing group and classified as L2.2.AA3 (Asia Ancestral 3), were also included for comparison. Both, Mtb 2279 and Mtb 18b, were genetically closer to L2.1 than the Modern L2.2 strains[2]. Mtb 18b is a streptomycin-dependent and commonly used experimental strain[31]. The strain was originally isolated as a streptomycin-resistant mutant in Japan in 1955[32]. The accession numbers and related details of the isolates are shown in Supplementary Table 1. To facilitate recognition, the names of the L2.2 isolates will be italicized below.

pe/ppe genes in H37Rv

A total of 169 *pe-ppe* genes were initially annotated in the genome of the H37Rv strain (NC_000962.1). In subsequent annotation, NC_000962.3, only 164 genes were listed as belonging to the *pe-ppe* family. We, therefore, initially annotated the *pe-ppe* genes in NC_000962.3 using Prokka v1.14.6 (Seemann, 2014), and subsequently Panaroo v 1.2.3[33] for gene predictions.

Among the 169 pe-ppe genes, only 153 were annotated by Prokka. Eleven of the 16 missing *pe-ppe* genes were already annotated in NC_000962.3 by PGAP at NCBI. Manual analysis of NC_000962.3 revealed the presence of the coding sequences of the remaining five unannotated genes (*ppe9*, *pe10*, *ppe31*, *ppe40*, and *pe27A*). In contrast, the orthologs of 161 genes were identified in two L2.1 isolates (Mtb CR170941 and Mtb N0031). The numbers were 165 for the third L2.1 isolate (Sea14117p6c4), 167 for *Mtb 2279* and 168 for *Mtb 18b*.

ppe9 was annotated as a pseudogene with a locus_tag Rv0387c in NC_000962.3. The 1330-bp-long annotated gene in NC_000962.3 exhibits multiple internal stop codons. Notably, the first stop codon results in a coding sequence with a length of 543 bp, corresponding to the annotation of *ppe9* (Rv0388c) in NC_000962.1. The predicted 180-amino-acid-residue long PPE9 contains only the PPE domain and is categorized as a PPE-SVP protein based on phylogenetic analysis[20,34].

pe10 was not annotated in NC_000962.3. However, manual assessment revealed, the presence of an open reading frame (ORF) identical to *pe10*, annotated in NC_000962.1, which is 363 bp long and overlaps 179 bp of the 3' end of *pe9*, suggesting that they may be in the same operon. Indeed, a study revealed the production of PE10 of the size of 15 KDa, which physically interacted with PE9, becoming a TLR4 ligand[35]. *pe10* of H37Rv (NC_000962.1) did not have a PE motif and appeared different from all the other studied strains, which all have a T deletion at position 336 bp, leading to a frameshift and elongation of the gene to 444 bp (147 amino acid residues). The frameshift created a PE motif at amino acid position 120 which was not observed in H37Rv. The position is unusual for a PE protein, though.

ppe31 and *ppe40* were annotated in a NC_000962.1 as being 1,212 bp (Rv1807) and 1,848 bp (Rv2356c) long respectively.

Deleted *pe-ppe* genes in L2.1

Eight *pe-ppe* genes were deleted in the L2.1 isolates. *wag22*, *ppe38*, *ppe50* and *ppe66* were absent in all three L2.1 isolates while the other four (*pe27A*, *esxS* (*pe28*), *ppe47* and *ppe48*) were absent in two L2.1 isolates, as summarized in Supplementary Table 2.

The *wag22*-containing 20,857-bp-long region in H37Rv, extending between *ppe24* and *pe_pgrs31*, was different in each isolate. The lengths of *ppe24*, belonging to the *ppe-mptr* family, were variable among the five L2 isolates. They and H37Rv retained a highly similar 1562-bp-long sequence at the 5' terminal. Compared to H37Rv, an extra 234-bp-long *mptr* segment was inserted at nucleotide position (nt) 1562 in both L2.2.AA3 isolates as well as a L2.1 isolate, Mtb-N0031. The second 78-bp-long segment was inserted in Mtb-N0031 and *Mtb 18b*, at nt 1874 of *ppe24* of H37Rv. In contrast, nt 1563–1874 were deleted in Mtb CR170941. At nt 2856 in H37Rv, all L2 isolates have inserted extra-*mptr* segments, albeit with variable lengths, 300 bp long in two L2.1 isolates, Mtb CR170941 and Sea14117p6c4, while the ones in the L2.2.AA3 isolates were slightly shorter. At position 176 of this inserted segment in Mtb-N0031, an insertion of IS*6110* disrupted the translation of the gene. There is a different copy of IS*6110* inserted further downstream in MtbCR170941 at nt 2974 of *ppe24* in H37Rv. The 3' end of the ORF of *ppe24* in Mtb N0031 and Mtb CR170941, was thus annotated as hypothetical genes (Fig. 1).

Next to *ppe24*, *Rv1754c* encodes a 563-amino-acid-long protein, homologous to an endo-D-arabinofuranase[36]. The sequences of *Rv1754c* were similar in all L2.1 isolates and H37Rv. However, in both L2.2.AA3 isolates, a

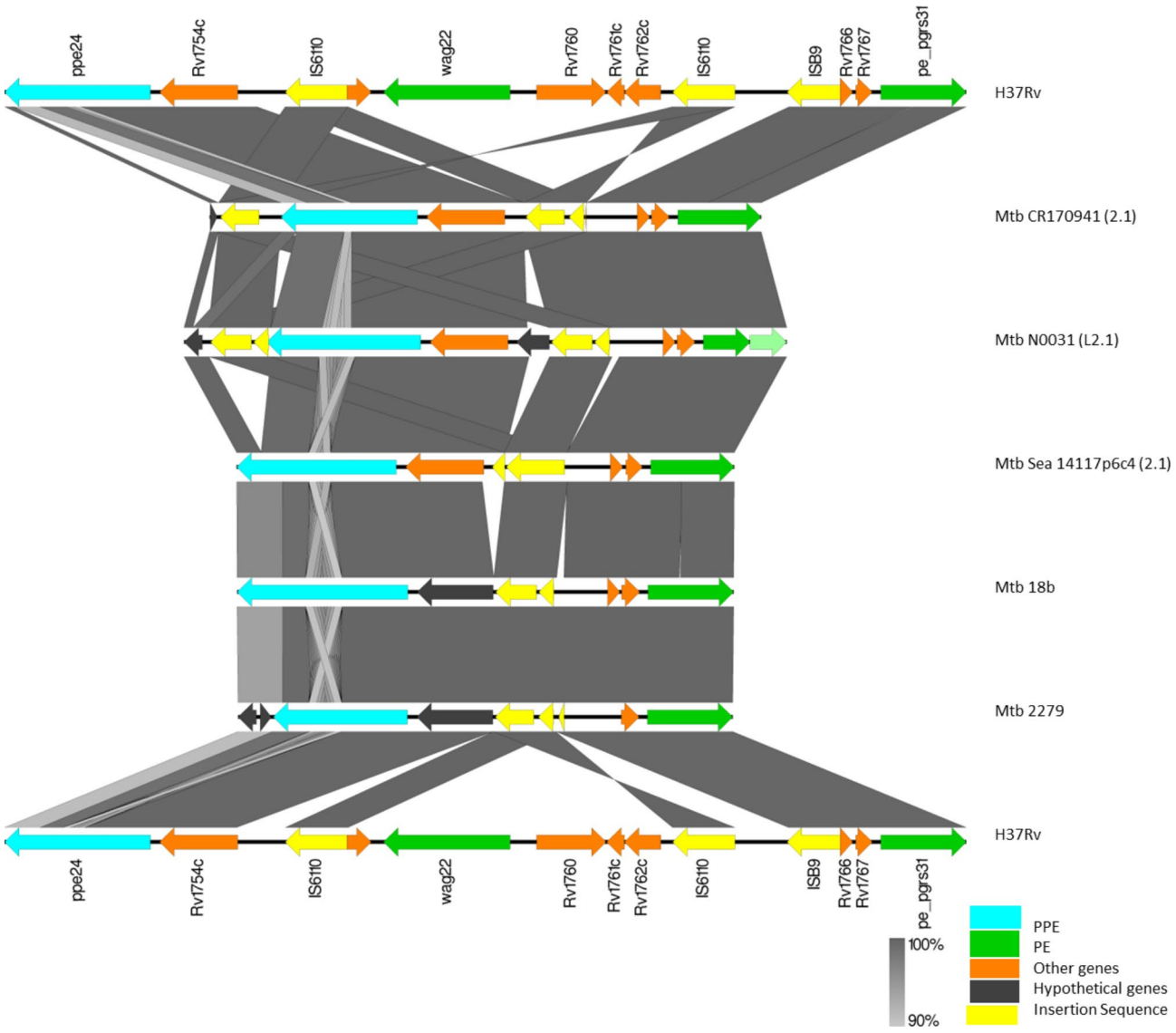

Fig. 1. Alignment of genes from *ppe24* to *pe_pgrs31* among the six isolates. The large, deleted region included 510 bp of *cut1*, *wag22* and three hypothetical genes (*Rv1760*, *Rv1761c* and *Rv1762c*) in all L2.1 and 2.2 isolates. The *pe_pgrs31* in Mtb N0031 was split into two ORFs possibly due to a homopolymeric error of long read sequencing data. However, short-read sequencing data is not available for verifying the sequence. Two coding sequences at the *pe_pgrs31* locus is, therefore, presented as predicted.

copy of IS*6110* was inserted at nt 32, disrupting the gene. Hence *Rv1754c* was likely to be non-functional in L2.2.AA3. There was a copy of IS*6110* upstream *Rv1754c* in each of the other studied isolates. The copy is located at 1032 bp upstream the start codon of *Rv1754c* in H37Rv and 356, 449 and 899 bp upstream in Mtb CR170941, Sea14117p6c4 and Mtb N0031 respectively.

The *ppe24-pe_pgrs31* region of H37Rv contains two copies of IS*6110* which flank the *wag22-Rv1760-Rv1761c-Rv1762c* segment. In all studied L2 isolates only one copy of IS*6110* remained, with variable lengths of the deleted segments, which included a part of an insertion sequence ISB9[37]. The lengths of the deleted segments were 9574 bp in Mtb N0031 and Mtb CR170941, 9737 bp in Mtb Sea14117p6c4, and 9565 bp in both L2.2.AA3 isolates, mirroring previously reported deletion in Beijing isolates in the region, LGD2 (7317 bp)[38] and RD152 (11,985 bp)[39].

Finally, next to the deleted DNA segment above, a segment containing *Rv1766-Rv1767* was largely conserved in all studied L2 isolates. However, in all L2.1 isolates, a 48-bp deletion within the *pe_pgrs31* was observed spanning from the position 673 to 720. (Fig. 1).

An alignment of the 6812-bp-long region in H37Rv covering *plcB*, *plcA*, *ppe38* and *ppe39* (*Rv2350c-Rv2353c*) revealed a large deletion that covered the first 793 bp (of 1539 bp) of *plcB*, *plcA*, and *ppe38*, in all L2.1 isolates with the insertion of a 1517-bp-long IS*6110* at the same position (Fig. 2). This suggests that the insertion and recombination of IS*6110* was the mechanism of the deletion, as previously described[13].

In contrast all four genes were identified in both L2.2.AA3 isolates, albeit with the insertion of IS*6110* at the nucleotide position 72 of *ppe38*. The ORF may be still functioning using another in-frame start codon at nt 82. However, the putative mutant protein would not have a PPE motif, which is coded by nt 22–30. Therefore, the complete *ppe38* coding sequence was essentially absent in all L2 isolates. There was the second copy of IS*6110* between *plcA* and *ppe38* in *Mtb 2279*. (Fig. 2). The *ppe38* gene region was previously shown to be hypervariable probably due to the presence of IS*6110*[40].

The Mtb H37Rv reference genome (NC_000962.3) contains a single copy of the *ppe38* (Fig. 2). In contrast, the Mtb CDC1551 genome has two copies of the gene, identified as *ppe38* and *ppe71*, which flank the *esxXY* genes (Fig. 2). Notably, *ppe38* and *ppe71* are genetically identical except for a 21-bp deletion in *ppe71*[21,40].

We analysed a 4,055-bp-long region containing *ppe46*, *pe27A*, *esxR*, *esxS* (*pe28*), *ppe47*, *ppe48*, and *pe29* in H37Rv (NC_000962.1). Interestingly, we found that in the NC_000962.3, *ppe48* was annotated to be terminated by an unusual stop codon, CAG, while a conventional stop codon, TGA, was located at nt 457–459. The downstream *ppe47* did not have any PPE signature motif, which is present only in the *ppe48*, and was annotated as a pseudogene in NC_000962.3. Alignment of the *ppe46* gene (1305-bp-long) and the 1307-bp-long *ppe47-ppe48* segment revealed that they are very similar in the first 595 bp and the last 419 bp while there were considerable variations in the middle segment. Strikingly, there was a substantial 2,439-bp-long deletion in two L2.1 isolates (Mtb CR170941 and Mtb N0031), encompassing the first 244 bp of *ppe46*, *ppe47* and most of *ppe48*, leaving only the first 244 bp of *ppe48* and resulting in the fusion of *ppe46* and *ppe48*. It should be noted, though, that the first 244 bp of *ppe48* is only one base different, G149A, from the first 244 bp of *ppe46*. The 2,439-bp-long deletion was not found in the other L2.1 isolate (Fig. 3). A metagenomic study of an 18th -century mummified body in Hungary revealed the same 2439-bp-long deletion covering *pe27A*, *esxR*, *esxS* and *ppe47*[41]. The same deletion was previously reported in an outbreak strain 7199/99 in Germany, which was a L4 strain[42]. The deletion was, therefore, homoplastic which was likely to be due to the almost identity between the 5' or 3' regions of *ppe46* and *ppe47-ppe48*. The region from *ppe46* to *pe29* was intact in the other three L2 isolates, albeit with an insertion of IS*6110*.

Alignment of a 2994-bp-long segment in H37Rv containing *Rv3134c*, *ppe50*, and *ppe51* (Fig. 4) revealed deletion of *ppe50* in all three L2.1 isolates. The deleted segment ranged from nt -12 to 331 of *ppe50* (519 bp long) and was replaced by a 242-bp-long segment. The deletion in *Mtb 2279*, a L2.2.AA3 isolate, ranged from nt -144 to 331 and was replaced by a 262-bp-long DNA segment, which was homologous to the 242-bp-long segment. Surprisingly, an insertion of a 1337-bp-long segment was observed in, the second L2.2.AA3 isolate, *Mtb 18b* at position 331 resulting in a novel 1146-bp-long *ppe* gene, tentatively named as *ppe50e* (for extended). In addition to the PPE and WWG (WxG) motifs, PPE50E also contained an SVP motif at amino acid position 312–314. It is interesting to note that PPE50 has been classified into the PPE-SVP group based only on the sequence similarity of the 519-bp-long coding sequence in H37Rv[20,34]. However, in H37Rv, *ppe50* is much shorter than the other *ppe-svp* genes and lacks the SVP coding sequence. The discovery of this longer *ppe50e* confirms its classification as *ppe-svp* (*esx5* cluster)[43].

Alignment of a 2819-bp-long segment of H37Rv, encompassing *Rv3737*, *ppe66*, and *ppe67*, revealed identical sequences among both L2.2.AA3 isolates and H37Rv. All L2.1 isolates had a duplication of a 128-bp-long segment at position 1081–1208 of *Rv3737*. This duplication caused frameshift and resulted in a stop codon at position 1275, instead of 1590 as in H37Rv. *Rv3737* is a homolog of multifunctional transporter ThrE[44]. All L2.1 isolates also lost a segment containing the last 48 nucleotides of *Rv3737* together with most of *ppe66* except for its first 42 bp as depicted in Fig. 5. Deletion of the *ppe66*-containing segment of variable lengths was reported in several sublineages of MTBC, including L2.1, which was described as RD25_tbB[45].

Novel genes in L2.1

Some novel genes similar to existing *pe-ppe* were identified in L2.1 and L2.2 isolates. These include genes like *pe_pgrs3-pe_pgrs4* and *ppe53*.

Mtb typically contains two similar contiguous genes, *pe_pgrs3* and *pe_pgrs4*. The proteins are unique among PE_PGRS as they contain two GRPLI motifs instead of one. PE_PGRS3 additionally has an arginine (R)-rich C-terminal domain, which binds to host phosphatidylinositol and cardiolipin while PE_PGRS4 does not[46].

Surprisingly, all L2.1 isolates contain three contiguous genes similar to *pe_pgrs3* or *pe_pgrs4* (Fig. 6), tentatively designated as *pe_pgrs3**, *pe_pgrs4**, and *pe_pgrs4* with a length of 2808 bp, 2700 bp and 2226 bp respectively.

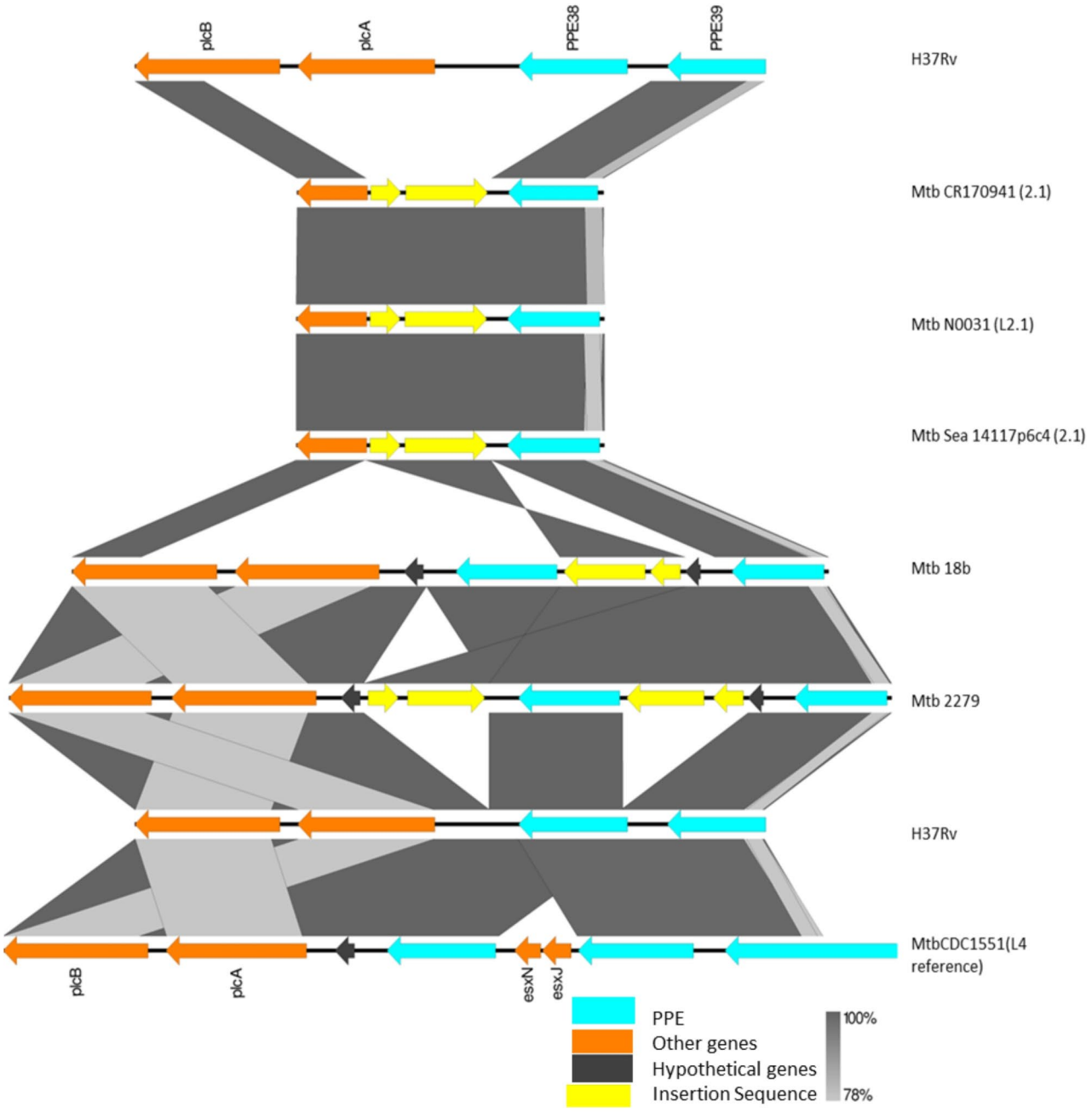

Fig. 2. Alignment of genes from *plcB* to *ppe39* among the six isolates in this study. The *ppe38* was deleted in all L2.1 isolates, replaced with IS*6110*. The homologous region of CDC1551, which contains additional genes, *esxN*, *esxJ* and *ppe71*, is also included.

They all contain the PE motif at amino acid residues positions 8 and 9, the Y***D/E motif at positions 87–91, the first GRPLI motif at positions 111-5, the second GRPLI motif at position 527–531, 522–526, and 395–399 respectively, as well as a C-terminal domain.

To properly identify the genes, we aligned their sequences with *pe_pgrs3* and *pe_pgrs4* identified in L2.2.AA3 and H37Rv. The alignment was used for constructing a SNV-based phylogenetic tree as shown in Fig. 7A and for calculation of pairwise SNV distances as shown in Supplementary Table 3. The genes at the 5'end of the region in all L2.1 isolates contained an R-rich C terminal domain and showed high similarity to the *pe_pgrs3* of the L2.2.AA3, with the average pairwise SNV distances among the L2 isolates of 5.4 bp. However, they were only distantly similar to *pe_pgrs3* in H37Rv (average pairwise SNV distances of 273.2 bp). We, therefore, designated this gene as *pe_pgrs3**.

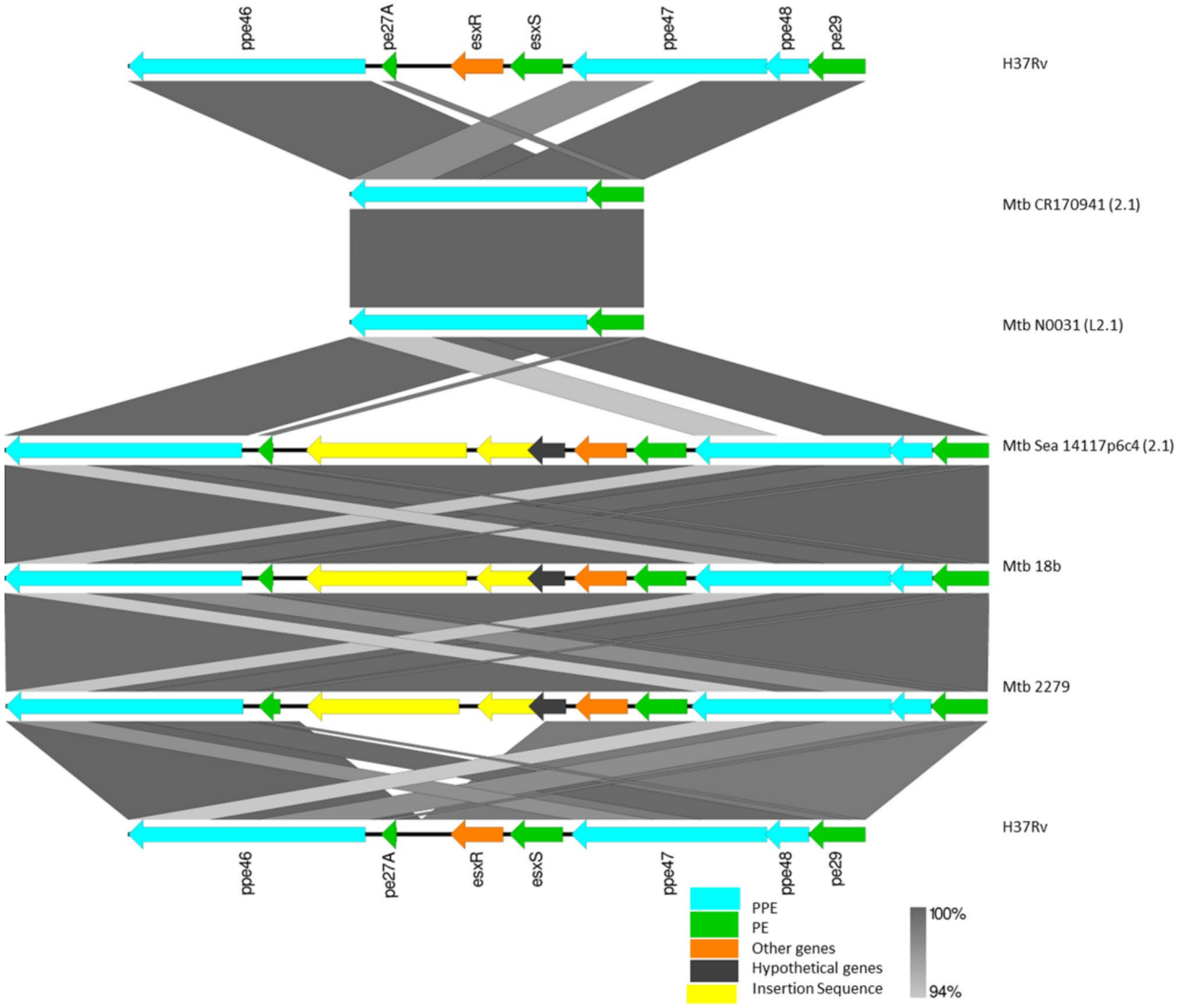

Fig. 3. Gene alignment of the region between *ppe46* and *pe29*. The deleted region of two L2.1 isolates ranged from the 5' end of *ppe46*, including *pe27A*, *esxR*, *exsS*, *ppe47* and 3' end of *ppe48*, resulting in a chimeric *ppe46/48* gene. The other L2.1 isolate (Mtb Sea14117p6c4) shared a similar gene order to the two L2.2 (*Mtb18b* and *2279*) isolates.

The middle genes in all L2.1 isolates were similar to *pe_pgrs4* of L2.2.AA3 (average pairwise SNV distances of 22.7 bp). Like H37Rv *pe_pgrs4*, they did not contain the R-rich C terminal domain but their average pairwise SNV distances from H37Rv *pe_pgrs4* were high (394.6 bp). We therefore designated this gene as *pe_pgrs4**.

The last gene at the 3' end of the region in all 2.1 isolates displayed similarity to *pe_pgrs4* in H37Rv (average pairwise SNV distances of 12.5 bp) but not similar to *pe_pgrs4* in L2.2.AA3. A diagram showing the similarity between the genes is shown in Fig. 7B.

In summary, we identify three similar genes in L2.1 isolates, which are named as *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4*. As *M. canetti* and *M. bovis* had two copies of *pe_pgrs3*[44], we then further include *pe_pgrs3-pe_pgrs4*-like genes of a sample of *M. canetti* (NC_015848), two samples of *M. bovis* BCG (NC_012207; NC_008769) and a sample of *M. bovis* (NC_002945) in the alignment. It was found that they harbor genes similar to *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4* in the same order. They do not have any gene similar to H37Rv *pe_pgrs3* as shown in Fig. 7 and Supplementary Table 4. Interestingly, PE_PGRS3* of *M. canetti*, *M. bovis* BCG and *M. bovis* do not have the R-rich C-terminal domain. This is caused by the presence of a T deletion just before the coding sequence of the R-rich region (nt 2580). The frameshift resulted in the loss of R-rich C terminal domain and shortening of C terminal domain by 22 residues.

ppe53, encoding a PPE-MPTR, was duplicated in all L2.1 and L2.2 isolates. In H37Rv, *ppe53* is 1773 bp long. There was a large insertion at nt 192 in all L2.1 and L2.2 isolates, resulting in two adjacent genes, tentatively called *ppe53A* and *ppe53B* (Fig. 8). The inserted segment provides the first 192 nucleotides of *ppe53A*, which is identical to the original *ppe53*, while the original N-terminus of *ppe53* became a part of a novel gene, *ppe53B*,

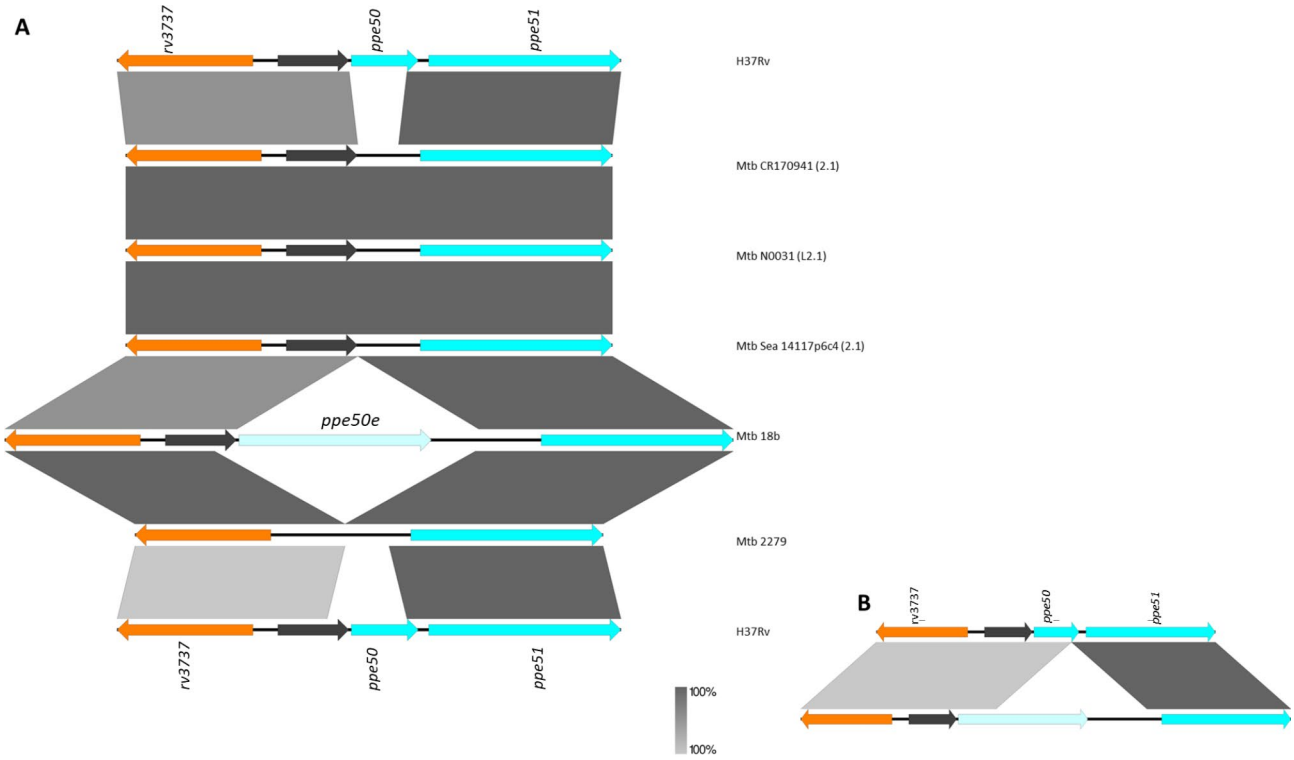

Fig. 4. Gene alignment from *Rv3134* to *ppe51*. (**A**) The alignment among the six isolates The PPE50 was deleted in all L2.1 and a L2.2 isolate (*Mtb 2279*). In contrast, *Mtb 18b* had an insertion, containing the SVP motif, that resulted in *ppe50e*, a longer version of *ppe50*. (**B**) The alignment between H37Rv (above) and *Mtb 18b* (below).

which is 1902-bp-long in most isolates. In Mtb CR170941 a copy of IS*6110* was inserted at position 820 of *ppe53B*. The putative novel PPE53B exhibited similarity in features with other known PPE-MPTRs[20,34,47]. The duplication of the *ppe53* was previously observed in all Mtb isolates, except for isolates belonging to L4.3 to L4.9 and L8, as reported by[43]. The average pairwise SNV distances among studied L2 isolates for *ppe53A* and *ppe53B* were 1.2 and 0.6 respectively. However, the average pairwise SNV distance between *ppe53B* and *ppe53A* was 365.8 (Supplementary Table 5).

Discussion

The Mtb genome holds a captivating feature in the form of *pe-ppe*, especially noteworthy for their prevalence in pathogenic mycobacteria[21]. Since their discovery in 1998, the genetic variability of Mtb *pe-ppe*, has been an interesting subject of study[48], but hampered by a large number of similar genes, high levels of structural variations and difficulties in correctly mapping NGS reads to a reference genome. The availability of long-read sequencing data and hybrid assembly pipelines allow the construction of complete genomes of Mtb clinical isolates and facilitate the study of variations of *pe-ppe*. This together with the advanced capabilities of gene annotation tools and more experimental verification, will lead to better annotation and resolution of previous discrepancy. For example, experimental deletion of the unannotated *ppe31* led to increased sensitivity to acid, reduced survival in macrophages, and decreased host cell death, suggesting that the putative *ppe31* is indeed a functioning gene[49].

Structural variations affecting some specific *pe-ppe* genes have been documented, with some of these deletions being associated with distinct RDs, genotypes and phenotypes. Lew et al.[50] reported instances of partial or total deletion of certain *pe-ppe*, some of which serve as antigenic precursors. Deletions of *pe-ppe* have the potential to affect Mtb virulence and ability by evading host immune mechanisms[38]. We have analyzed all three available complete genomes of Mtb L2.1, the genotype that is relatively rare but has been associated with MDR- or pre-XDR-TB and identified structural variants in 8 loci of *pe-ppe*, some of which were common to the studied L2.1 isolates. Although the studied isolates were limited in number and not representative of the diversity of MTBC, they illustrated the diversity of *pe-ppe* among clinical isolates.

The striking fact that Mtb L2.1 isolates contain three similar genes, *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4*, all of which are only distantly similar to H37Rv *pe-pgrs3* illustrates the difficulty to map NGS reads of clinical isolates to a correct *pe-ppe* gene. Forced mapping of NGS reads of some Mtb genotypes to the H37Rv *pe_pgrs3-pe_pgrs4* region may result in many identified SNPs, which led to the notion that *pe-pgrs3* was highly polymorphic. In the case of Mtb L2, for example, it would be the mapping of *pe_pgrs3** reads onto *pe_pgrs3*. This, therefore, supports the common precautionary practice of excluding the SNPs in the *pe-ppe* genes from phylogenetic

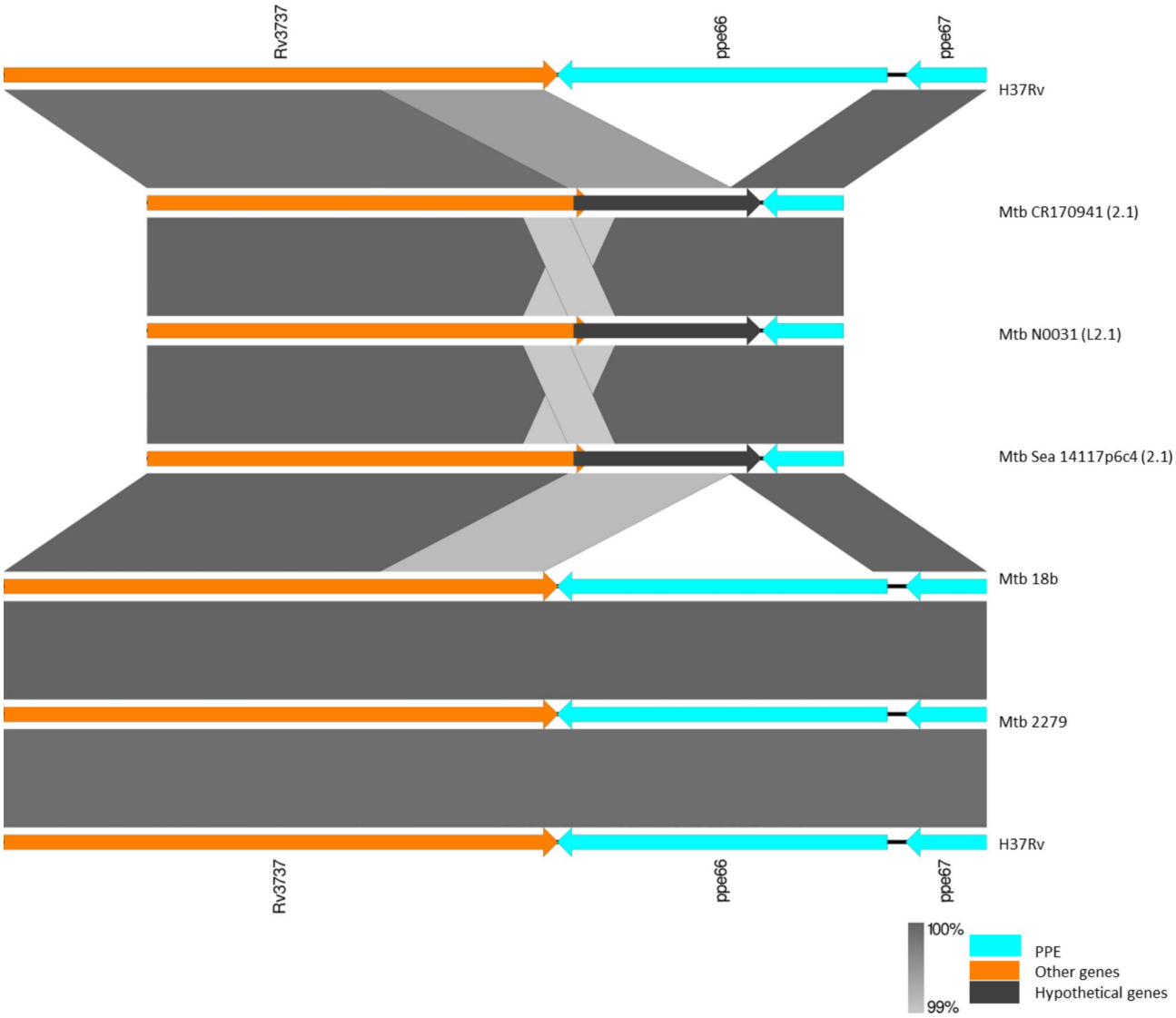

Fig. 5. Gene alignment from *Rv3737* to *ppe67* among the six isolates. The *ppe66* deletion was specific to L2.1. All L2.1 isolates had a duplication of a 128-bp-long segment at the position 1081–1208 of *Rv3737*, which resulted in frameshift and a premature stop codon. The remaining part of the coding sequence of *Rv3737* was annotated as a hypothetical gene (black).

analysis. However, this practice excludes a significant proportion of useful genomic data from investigation. A better understanding of the structural variations of *pe-ppe* would lead to the correct identification of genes and consequentially correct identification of SNPs or structural variants, which are essential for understanding the evolution and functions of the gene family.

Early branching species of MTBC, *M. canetti* and *M. bovis*[51,52] also contain the three genes of *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4*. A previous study of over 70 isolates across several lineages revealed duplication of *pe_pgrs3* in many samples[43], suggesting the presence of three genes in this set. We hypothesize that the common ancestor of MTBC was originally equipped with the three genes, one of which has been lost more recently. The lost gene may be different by sublineages, which requires further studies. The event leading to acquisition of the R-rich C-terminal domain is hard to speculate as *pe_pgrs3** of *M. canetti* and *M. bovis* also has a very similar coding sequence of the C terminal domain as the one of L2, but with only a single base deletion before the R-rich domain, causing frameshift and changing the amino acid sequence of the C-terminal domain. It is possible the ancestor of MTBC also had the gene encoding the R-rich C-terminal domain, which has been lost in some species with a reason that needs more investigation.

Although, the localization of *pe_pgrs3** in Mtb cells remains to be studied, the presence of the R-rich C-terminal domain suggests a similar function and localization to *pe-pgrs3*. The expression of the latter is induced in phosphate-depleted conditions. PE_PGRS3 is localized on Mtb cell surface with C-terminus protruding and specifically interacting with cardiolipin and phosphatidylinositol. The protein interacts with host cell membrane and mediates Mtb entry to epithelial cells[46]. It may also function as a phosphate scavenger

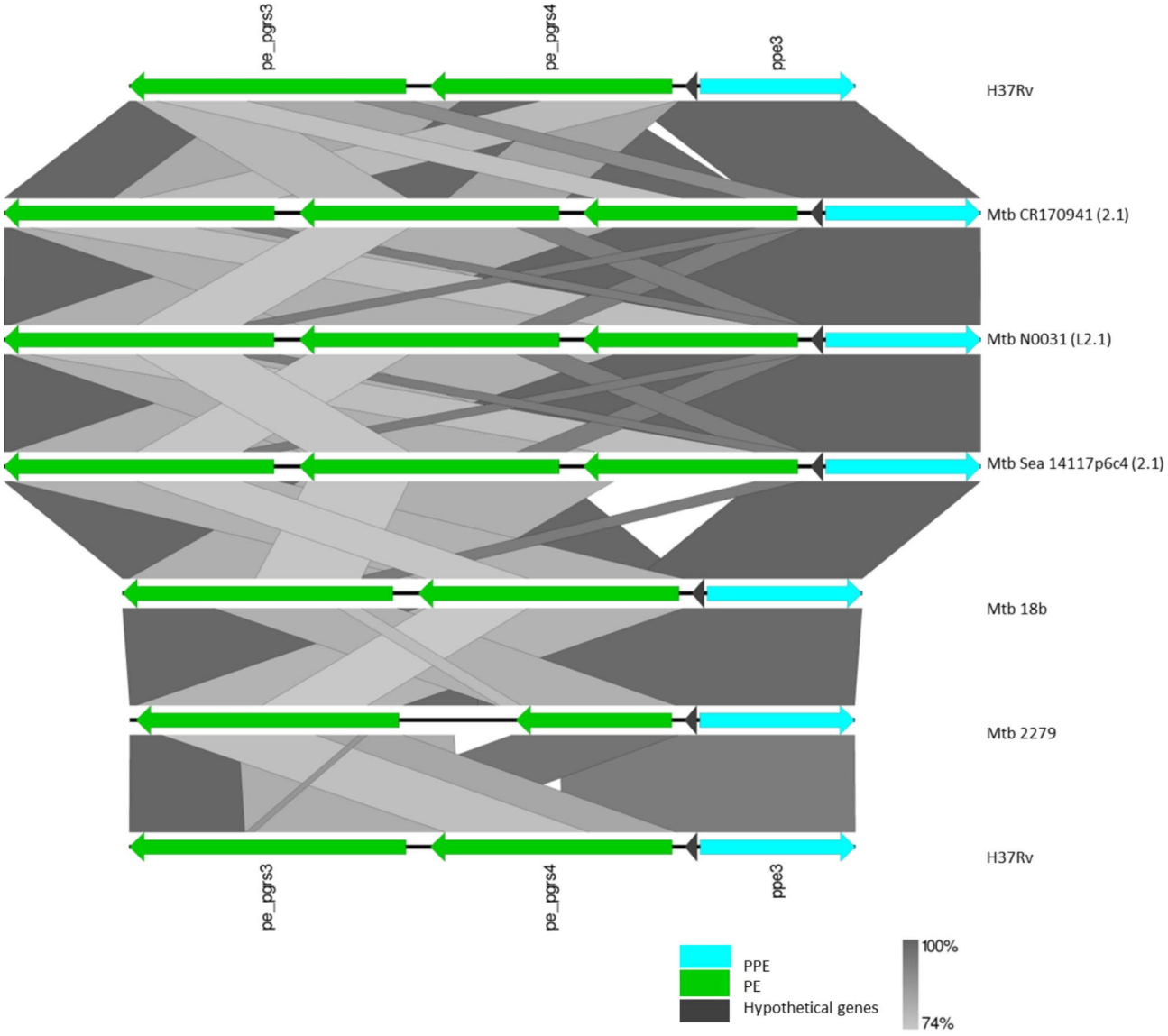

Fig. 6. Gene alignment of the *pe_pgrs3* to *ppe3* region among the six isolates. All L2.1 isolates harbor three copies of *pe_pgrs3-pe_pgrs4*-like genes, from left to right namely *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4*. The SNP distances between each pair of the three genes was approximately equal and much higher than the distances between the same genes among different isolates. None was closely similar to H37Rv *pe_pgrs3*.

in the phosphate-depleted environment of phagosome. In contrast, PE_PGRS4 and PE_PGRS4* do not have the R-rich C-terminal domain and PE_PGRS4 in H37Rv is constitutively expressed. Their functions are unknown but likely to be different from PE_PGRS3 and PE_PGRS3*. Whether PE_PGRS4 and PE_PGRS4* have the same functions remains to be confirmed. In any cases, the gene set, *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4*, is likely to play some roles in Mtb pathogenesis and require more studies.

In reverse, there were a few *pe-ppe* genes deleted from all three L2.1 genomes, notably, the deletion of *ppe50*, which was previously reported as RD516[45]. As *ppe50-ppe51* is operonic[53], the deletion of *ppe50* coding sequence and its 5' untranslated region may also disrupt the expression of *ppe51* even though the coding sequence of the latter is intact. It was shown that deletion of *ppe50-ppe51* conferred INH and RIF tolerance, which may be related to carbon metabolisms[54]. *ppe50* by itself was significantly associated with extrapulmonary TB[55]. *ppe50* deletion was previously shown to be present in L1 isolates[43,56] and more specifically in the Nonthaburi genotype (L1.2.2.2 or EAI2_NTB)[57,58]. In contrast, *Mtb 18b* contained *ppe50e*, an extended version of *ppe50*. The insertion of the SVP-motif-containing segment similar to *Mtb 18b* was previously reported in L2, L5, L6, *M. bovis* and L8[43]. The loss-and-gain variations suggest a significant role of the protein in variable environments.

The deletion of *ppe38* is especially intriguing. Deletion mutations of *ppe38* completely blocked the secretion of two large subsets of ESX-5 substrates, that is, PPE-MPTR and PE_PGRS, together comprising > 80 proteins, which may result in the decreased pro-inflammatory response[59]. IS*6110*-linked deletion of *ppe38* occurred at the branching point of the 'modern' Beijing (L2.2) sublineages and is shared by Beijing outbreak strains worldwide,

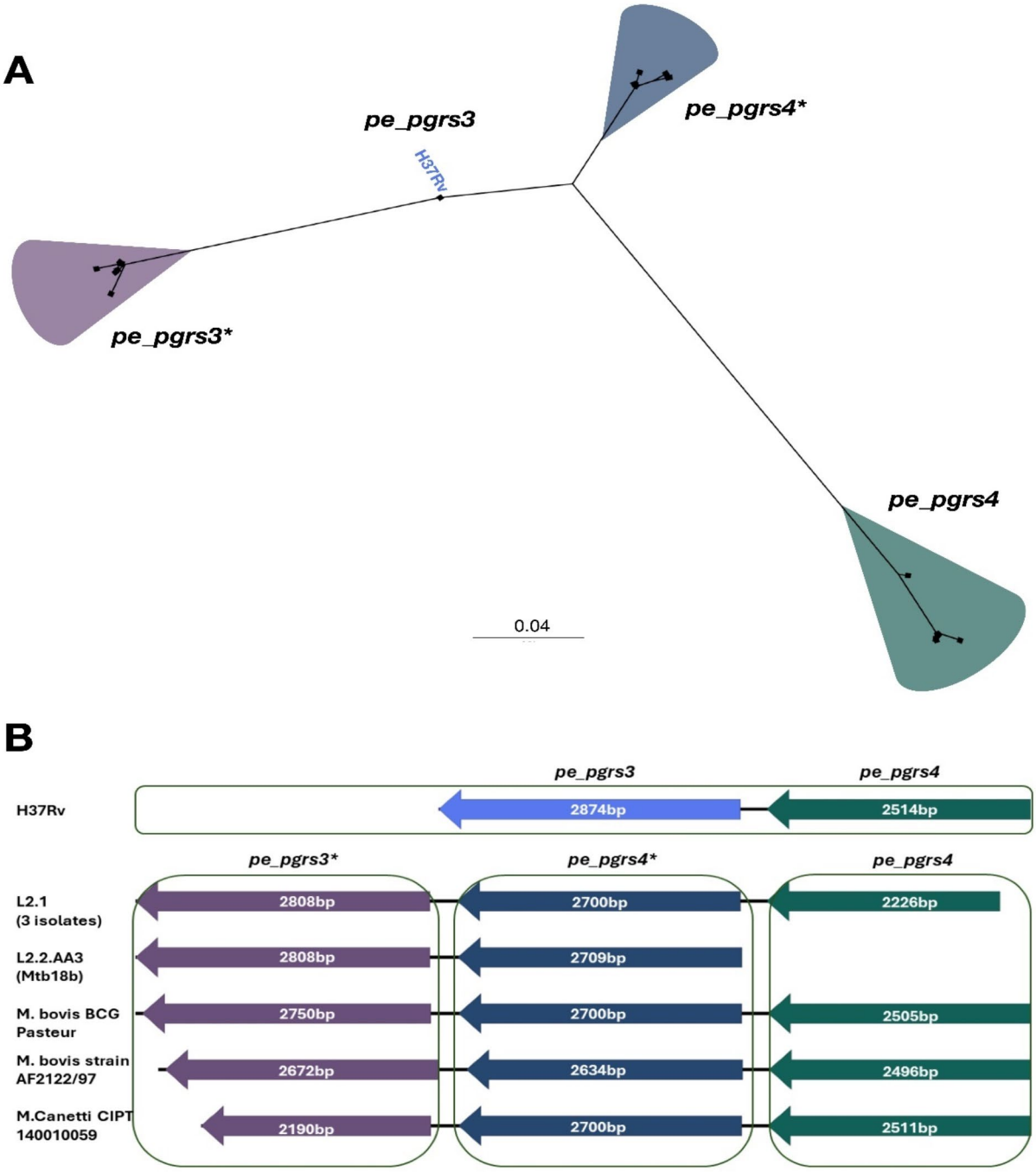

Fig. 7. The phylogenetic tree and gene mapping of *pe_pgrs3- pe_pgrs4*-like genes. (**A**) The SNP-based maximum likelihood phylogenetic tree of *pe_pgrs3*, *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4* of five L2 isolates, H37Rv, *M. canetti* CIPT, *M. bovis* AF2122/97 and *M. bovis* BCG Tokyo and Pasteur, showing that H37Rv *pe_pgrs3* is only distantly similar to the three other genes. The *pe_pgrs4* gene that is most dissimilar to the others belongs to *M. canetti* CIPT. (**B**) Gene alignment of the *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4* of representative isolates of L2.1, L2.2.AA3, *M. bovis*, *M. bovis* BCG and *M. canetti*. H37Rv *pe_pgrs3* and *pe_pgrs4* are also shown for comparison.

suggesting that this deletion may have contributed to their success and global distribution[60], including the expanding MDR/XDR W148 strains (L2.2.M4.5)[61]. Here we reveal that L2.1, which is not widespread, also lacks *ppe38*. Moreover, L2.2.AA3, an 'ancestral' sublineage, has *ppe38* with the 5' end disrupted by IS*6110*, even though it still retains most of its coding sequence. *ppe38* of L2.2.AA3 may, therefore, also be non-functioning. It should be noted, though, that L2.2.AA3 is one of the most common, or successful, 'ancestral' Beijing sublineages[2]. It

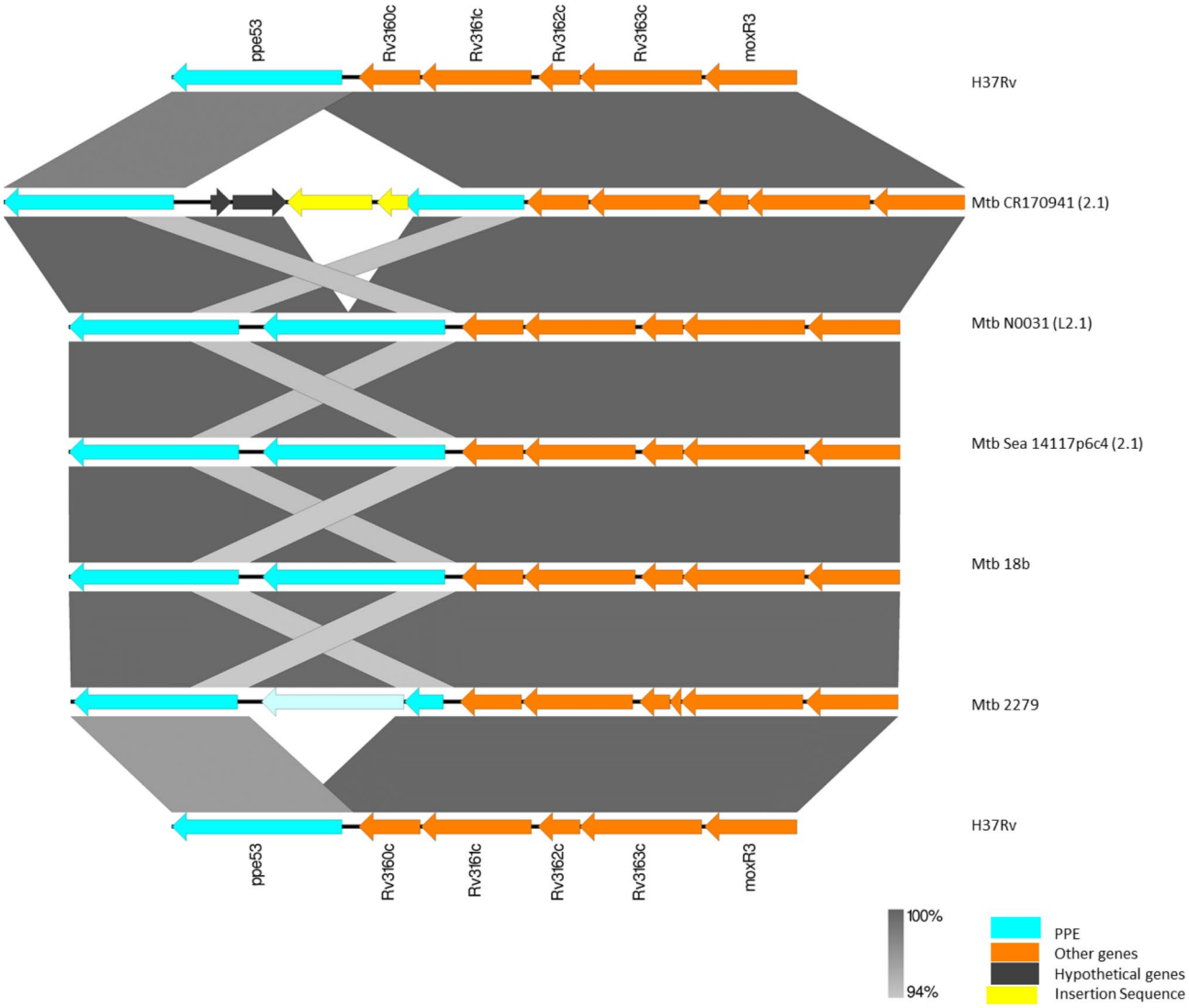

Fig. 8. Gene alignment from *ppe53* to *moxR3* showing *ppe53A* (left) and *ppe53B* (right) in L2 isolates. *ppe53B* of Mtb CR170941 was truncated by insertion of IS*6110*. The split of *ppe53B* in *Mtb 2279* into two coding sequences may be due to a homopolymeric run error in long-read sequencing. Its NGS data were not available for verification.

remains to be investigated whether the disruption of the coding sequence of *ppe38* is common among other 'ancestral' Beijing sublineages.

It is known though that the structural variants related to *ppe38* also include the presence of an extra copy of a highly similar gene, *ppe71*. *ppe71* has been reported from *M. canetti*, the reference H37Ra strain, the ATCC strain of H37Rv[40], and CDC1551 strain[62]. Moreover, a recent genomic study[63] of additional clones of H37Rv adds more complexity by identifying a complete copy of *ppe38* and a downstream truncated copy of *ppe38*, designated as *ppe38a* or *Rv2351c.2*. All these information also indicate loss-and-gain variations of the *ppe38* containing region, which is likely to be driven by homologous recombination of *ppe38* and its homolog or insertion of IS*6110* and subsequent homologous recombination.

Understanding the structural variations of *ppe38* is important as it plays important roles in Mtb pathogenesis. Apart from involvement in other PE-PPE proteins secretion, PPE38 was reported to downregulate macrophage MHC Class I expression and suppress CD8 + T cell activity[30]. Overexpression of *ppe38* may contribute to the success of Mtb MtZ strain in Aragon[64]. *ppe38* has been observed in *Mycobacterium marinum* and documented to hinder phagocytosis, while concomitantly modulating host immune responses through TLR-2 activation[65]. These factors coalesce to create a more complicated landscape of host-pathogen interactions, which may open new avenues for therapeutic interventions.

The significance of the deletions of the other two genes, *wag22* and *ppe66*, is not clear, as their functional studies are scarce. The wag22-containing *ppe24-to-pe_pgrs31* region, especially between *Rv1754c* to *Rv1765c* was previously shown to be a hot spot of IS*6110* insertions and deletion events[66]. A *pe_pgrs31* deficient mutant of H37Rv grew less in the murine Mtb infection model compared to the wild-type[67].

A recent study in China revealed the association of a SNP in *ppe66* (4,189,930; c.303G > C) with the cross-regional genetic clustering of Mtb isolates[68].

Overall, the findings in this study underscore the complex structural variations of some *pe-ppe* genes mediated by homologous recombination between highly similar genes, which may be a result of previous gene duplication, or between IS*6110*[69,70], probably through slip mispairing[71]. Reconciliation of structural variants and known phylogeny suggested multiple events during evolution and some structural variants are likely to be homoplastic. The variations may be results of adaptation of Mtb in different host populations or environments and suggest intricate interplay between *pe-ppe* gene modifications and their profound effects on mycobacterial interactions with the host immune system. Understanding the dynamics of *pe-ppe* not only furthers our grasp of mycobacterial pathogenesis but also has implications for vaccine development and diagnostic innovation. As we move forward, multidisciplinary approaches that combine genomic and molecular biology, computational analysis and immunology will be invaluable in deciphering the intricate mechanisms underlying these phenomena thus emphasizing the need for a holistic investigation into their roles and functional implications in mycobacterial pathogenesis.

Methods

DNA extraction and whole genome sequencing of Mtb CR170941

The annotated whole-genome sequence of *Mycobacterium tuberculosis* Mtb CR170941, isolated from a sputum sample of a 73-year-old male TB patient in Chiangrai province, Thailand, is presented. The Mtb CR170941 strain was cultured in Lowenstein-Jensen medium within a secure clinical microbiology laboratory in Chiangrai, following standard biosafety protocols and utilizing appropriate equipment. The NGS sequences of the sample was a part of a previous report[12], which determined the isolate to be L2.1, also known as Proto-Beijing genotype[2] The Mtb CR170941 library was prepared from 300 ng of unsheared DNA and sequenced using a MinION Flow Cell R10.4 with Kit 12 chemistry (SQK-NBD112.24). Adapters were trimmed with Porechop v0.2.4. The data were quality controlled by Filtlong v0.2.1 with a minimum mean quality more than 0.8 and minimum read length more than 1000. The complete genome was assembled with Flye v. 2.9, Minimap2 v. 2.24, and Pilon v. 1.24. The assembled genome was reorientated to match the reference genome of H37Rv, NC_000962.3, which start at *dnaA* (NP_214515.1)[56].The complete genome project has been deposited in GenBank under accession number CP104271. The NCBI Prokaryotic Genome Annotation Pipeline (PGAP) was employed for annotation to determine gene and coding sequence numbers. All laboratory works were conducted in Thailand, adhering to relevant guidelines and regulations.

Pe-ppe gene mapping, identification and annotation

The complete genome sequences of six isolates including H37Rv (NC_000962.3), 3 L2.1 isolates (Mtb CR170941 (this study), Mtb N0031 [NZ_CP069076.1] and Mtb Sea14117p6c4 [NZ_CP041797.1] and 2 L2.2.AA3 isolates (*Mtb 18b* [NZ_CP007299.1] and *Mtb 2279* [CP010336.1]) were downloaded from NCBI and annotated using Prokka v 1.14.6[72]. The predicted gene results were used as input for comparative genome analysis with Panaroo v1.2.3[33]. If a *pe-ppe* gene was not present in every isolate, the DNA segment corresponding to the *pe-ppe* region, as shown in Supplementary Table 6, was extracted from the genomic sequences manually. Multiple sequence alignment of each segment was done by Aliview v1.28[73] Genes and open reading frames were annotated manually. Structural variants were identified by comparative visualization of gene maps in each region to the one of H37Rv by Easyfig v2.2.5[74].

Phylogenetic tree reconstruction and SNP distance estimation

The phylogenetic tree of *pe_pgrs3*, *pe_pgrs3**, *pe_pgrs4** and *pe_pgrs4* was constructed by aligning each open reading frame that was homologous to *pe-pgrs3* (position 333,437 − 336,310) or *pe-pgrs4* (position 336,560 − 339,073) of H37Rv (NC_000962.3) across Mtb L2.1, L2.2.AA3, *M. bovis* BCG Tokyo 172 (NC_012207), *M bovis* BCG Pasteur 1173P2 (NC_008769), *M. bovis* (NC_002945) and *M. canetti* (NC_015848). The multiple sequence alignment was used as input to construct the maximum likelihood phylogenetic tree using IQ-TREE2[75], which determined the best-fit model (TIM2 + F + I), and also assessed tree reliability with 1,000 replicates by Ultrafast Bootstrap Approximation. Tree visualization was executed by FigTree v1.4. Pairwise nucleotide distances (number of SNP differences) were calculated using MEGA 11[76].

Data availability

All relevant data are presented as figures within the main text. The sequence data have been deposited in the NCBI database under accession number CP104271 (BioProject PRJNA877376) and can be accessed at https://www.ncbi.nlm.nih.gov. Supplementary information, including additional data, is available in Supplementary Tables S1-S6.

Received: 22 July 2024; Accepted: 8 November 2024
Published online: 28 December 2024

References

1. Gagneux, S. Ecology and evolution of Mycobacterium tuberculosis. *Nat. Rev. Microbiol.* **16**, 202–213 (2018).
2. Thawornwattana, Y. et al. Revised nomenclature and SNP barcode for Mycobacterium tuberculosis lineage 2. *Microb. Genom.* **7** (2021).
3. Napier, G. et al. Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome Med.* **12**, 114 (2020).

4. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin. Immunol.* **26**, 431–444 (2014).
5. Rutaihwa, L. K. et al. Multiple introductions of Mycobacterium tuberculosis Lineage 2–Beijing Into Africa over centuries. *Front. Ecol. Evol.* **7**, 112 (2019).
6. Stucki, D. et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48**, 1535–1543 (2016).
7. Rajwani, R. et al. Comparative whole-genomic analysis of an ancient L2 lineage Mycobacterium tuberculosis reveals a novel phylogenetic clade and common genetic determinants of hypervirulent strains. *Front. Cell. Infect. Microbiol.* **7**, 539 (2018).
8. Ribeiro, S. C. M. et al. Mycobacterium tuberculosis strains of the modern sublineage of the Beijing Family are more likely to display increased virulence than strains of the ancient sublineage. *J. Clin. Microbiol.* **52**, 2615–2624 (2014).
9. Holt, K. E. et al. Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet.* **50**, 849–856 (2018).
10. Liu, Q. et al. China's tuberculosis epidemic stems from historical expansion of four strains of Mycobacterium tuberculosis. *Nat. Ecol. Evol.* **2**, 1982–1992 (2018).
11. Yokoyama, E., Hachisu, Y., Hashimoto, R. & Kishida, K. Concordance of variable-number tandem repeat (VNTR) and large sequence polymorphism (LSP) analyses of Mycobacterium tuberculosis strains. *Infect. Genet. Evol.* **10**, 913–918 (2010).
12. Ajawatanawong, P. et al. A novel ancestral Beijing sublineage of Mycobacterium tuberculosis suggests the transition site to modern Beijing sublineages. *Sci. Rep.* **9**, 13718 (2019).
13. Srilohasin, P. et al. Genomic evidence supporting the clonal expansion of extensively drug-resistant tuberculosis bacteria belonging to a rare proto - Beijing genotype. *Emerg. Microbes Infect.* **9**, 2632–2641 (2020).
14. Phyu, A. N. et al. Genomic sequencing profiles of Mycobacterium tuberculosis in Mandalay Region, Myanmar. *Trop. Med.* **8**, 239 (2023).
15. Guyeux, C. et al. Connection between two historical tuberculosis outbreak sites in Japan, Honshu, by a new *ancestral Mycobacterium tuberculosis* L2 sublineage. *Epidemiol. Infect.* **150**, e56 (2022).
16. Alland, D. et al. Role of large sequence polymorphisms (LSPs) in Generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. *J. Clin. Microbiol.* **45**, 39–46 (2007).
17. Choudhary, R. K., Pullakhandam, R., Ehtesham, N. Z. & Hasnain, S. E. Expression and characterization of Rv2430c, a novel immunodominant antigen of Mycobacterium tuberculosis. *Protein Exp. Purif.* **36**, 249–253 (2004).
18. Delogu, G. & Brennan, M. J. Comparative Immune response to PE and PE_PGRS antigens of *Mycobacterium tuberculosis*. *Infect. Immun.* **69**, 5606–5611 (2001).
19. Bottai, D. et al. TbD1 deletion as a driver of the evolutionary success of modern epidemic Mycobacterium tuberculosis lineages. *Nat. Commun.* **11**, 684 (2020).
20. Ates, L. S. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. *Mol. Microbiol.* **113**, 4–21 (2020).
21. Fishbein, S., Van Wyk, N., Warren, R. M. & Sampson, S. L. Phylogeny to function: PE/PPE protein evolution and impact on *M ycobacterium tuberculosis* pathogenicity: evolution of PE/PPE-associated virulence. *Mol. Microbiol.* **96**, 901–916 (2015).
22. Mukhopadhyay, S. & Balaji, K. N. The PE and PPE proteins of Mycobacterium tuberculosis. *Tuberculosis* **91**, 441–447 (2011).
23. Qian, J., Chen, R., Wang, H. & Zhang, X. Role of the PE/PPE family in host–pathogen interactions and prospects for anti-tuberculosis vaccine and diagnostic tool design. *Front. Cell. Infect. Microbiol.* **10**, 594288 (2020).
24. Rahlwes, K. C., Dias, B. R. S., Campos, P. C., Alvarez-Arguedas, S. & Shiloh, M. U. Pathogenicity and virulence of *Mycobacterium tuberculosis*. *Virulence*. **14**, 2150449 (2023).
25. Bhat, K. H. et al. Proline-proline-glutamic acid (PPE) protein Rv1168c of Mycobacterium tuberculosis augments transcription from HIV-1 long terminal repeat promoter. *J. Biol. Chem.* **287**, 16930–16946 (2012).
26. Vordermeier, H. M. et al. Conserved immune recognition hierarchy of mycobacterial PE/PPE proteins during infection in natural hosts. *PLoS ONE* **7**, e40890 (2012).
27. Basu, S. et al. Prevention of nosocomial transmission of extensively drug-resistant tuberculosis in rural South African district hospitals: An epidemiological modelling study. *Lancet* **370**, 1500–1507 (2007).
28. McEvoy, C. R. E. et al. Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS ONE*. **7**, e30593 (2012).
29. Brennan, M. J. The enigmatic PE/PPE multigene family of mycobacteria and tuberculosis vaccination. *Infect. Immun.* **85**, e00969–e00916 (2017).
30. Meng, L. et al. PPE38 protein of Mycobacterium tuberculosis inhibits macrophage MHC Class I expression and dampens CD8 + T cell responses. *Front. Cell. Infect. Microbiol.* **7** (2017).
31. Benjak, A. et al. Genomic and transcriptomic analysis of the streptomycin-dependent Mycobacterium tuberculosis strain 18b. *BMC Genom.* **17**, 190 (2016).
32. Hashimoto, T. [Experimental studies on the mechanism of infection and immunity in tuberculosis from the analytical standpoint of streptomycin-dependent tubercle bacilli. 1. Isolation and biological characteristics of a streptomycin-dependent mutant, and effect of streptomycin administration on its pathogenicity in guinea-pigs]. *Kekkaku* **30**, 4–8 (1955).
33. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
34. Van Gey, N. C. et al. Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evol. Biol.* **6**, 95 (2006).
35. Tiwari, B., Ramakrishnan, U. M., Raghunand, T. R. & The *M ycobacterium tuberculosis* protein pair PE9 (Rv1088)-PE10 (Rv1089) forms heterodimers and induces macrophage apoptosis through Toll-like receptor 4: The PE9-PE10 protein pair of *M. tb* is a TLR4 ligand. *Cell Microbiol* **17**, 1653–1669 (2015).
36. Al-Jourani, O. et al. Identification of d-arabinan-degrading enzymes in mycobacteria. *Nat. Commun.* **14**, 2233 (2023).
37. Zainuddin, Z. F. & Dale, J. W. Polymorphic repetitive DNA sequences in Mycobacterium tuberculosis detected with a gene probe from a Mycobacterium fortuitum plasmid. *Microbiology* **135**, 2347–2355 (1989).
38. Zhang, Q. et al. Whole genome analysis of an MDR Beijing/W strain of Mycobacterium tuberculosis with large genomic deletions associated with resistance to isoniazid. *Gene* **582**, 128–136 (2016).
39. Cerezo-Cortés, M., Rodríguez-Castillo, J., Hernández-Pando, R. & Murcia, M. Circulation of *M. Tuberculosis* Beijing Genotype in Latin America and the Caribbean. *Pathog. Glob. Health* **113**, 336–351 (2019).
40. McEvoy, C. R., Van Helden, P. D., Warren, R. M. & Pittius, N. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. *BMC Evol. Biol.* **9**, 237 (2009). Van.
41. Chan, J. Z. M. et al. Metagenomic analysis of tuberculosis in a mummy. *N. Engl. J. Med.* **369**, 289–290 (2013).
42. Roetzer, A. et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis Outbreak: a longitudinal Molecular Epidemiological Study. *PLoS Med.* **10**, e1001387 (2013).
43. Gómez-González, P. J. et al. Functional genetic variation in pe/ppe genes contributes to diversity in Mycobacterium tuberculosis lineages and potential interactions with the human host. *Front. Microbiol.* **14**, 1244319 (2023).
44. Li, Q. et al. Rv3737 is required for Mycobacterium tuberculosis growth in vitro and in vivo and correlates with bacterial load and disease severity in human tuberculosis. *BMC Infect. Dis.* **22**, 256 (2022).
45. Liu, Z. et al. Identification of region of difference and H37Rv-related deletion in Mycobacterium tuberculosis complex by structural variant detection and genome assembly. *Front. Microbiol.* **13**, 984582 (2022).

46. De Maio, F. et al. PE_PGRS3 ensures provision of the vital phospholipids cardiolipin and phosphatidylinositols by promoting the interaction between *M. Tuberculosis* and host cells. *Virulence* **12**, 868–884 (2021).
47. D'Souza, C., Kishore, U. & Tsolaki, A. G. The PE-PPE family of Mycobacterium tuberculosis: Proteins in disguise. *Immunobiology*. **228**, 152321 (2023).
48. Delogu, G., Brennan, M. J., Manganelli, R. P. E. & Genes, P. P. E. and A tale of conservation and diversity. In *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control* (ed Gagneux, S.) Vol. 1019 191–207 (Springer International Publishing, 2017).
49. Feng, S. et al. Mycobacterium PPE31 contributes to host cell death. *Front. Cell. Infect. Microbiol.* **11**, 629836 (2021).
50. Lew, J. M., Kapopoulou, A., Jones, L. M. & Cole, S. T. TubercuList – 10 years after. *Tuberculosis* **91**, 1–7 (2011).
51. Orgeur, M., Sous, C., Madacki, J. & Brosch, R. Evolution and emergence of *Mycobacterium tuberculosis*. *FEMS Microbiol. Rev.* **48**, fuae006 (2024).
52. Orgeur, M. & Brosch, R. Evolution of virulence in the Mycobacterium tuberculosis complex. *Curr. Opin. Microbiol.* **41**, 68–75 (2018).
53. Mukku, R. P., Poornima, K., Yadav, S. & Raghunand, T. R. Delineating the functional role of the PPE50 (Rv3135) - PPE51 (Rv3136) gene cluster in the pathophysiology of Mycobacterium tuberculosis. *Microbes Infect.* **26**, 105248 (2024).
54. Martini, M. C. et al. Loss of RNase J leads to multi-drug tolerance and accumulation of highly structured mRNA fragments in Mycobacterium tuberculosis. *PLoS Pathog.* **18**, e1010705 (2022).
55. Negrete-Paz, A. M., Vázquez-Marrufo, G., Gutiérrez-Moraga, A. & Vázquez-Garcidueñas, S. Pangenome reconstruction of mycobacterium tuberculosis as a guide to reveal genomic features associated with strain clinical phenotype. *Microorganisms* **11**, 1495 (2023).
56. Thorpe, J. et al. Multi-platform whole genome sequencing for tuberculosis clinical and surveillance applications. *Sci. Rep.* **14**, 5201 (2024).
57. Coker, O. O. et al. Genetic signatures of *Mycobacterium tuberculosis* Nonthaburi genotype revealed by whole genome analysis of isolates from tuberculous meningitis patients in Thailand. *PeerJ* **4**, e1905 (2016).
58. Netikul, T. et al. Whole-genome single nucleotide variant phylogenetic analysis of Mycobacterium tuberculosis Lineage 1 in endemic regions of Asia and Africa. *Sci. Rep.* **12**, 1565 (2022).
59. Gallant, J. et al. PPE38-Secretion-dependent proteins of M. Tuberculosis Alter NF-kB signalling and inflammatory responses in macrophages. *Front. Immunol.* **12**, 702359 (2021).
60. Ates, L. S. et al. Mutations in ppe38 block PE_PGRS secretion and increase virulence of Mycobacterium tuberculosis. *Nat. Microbiol.* **3**, 181–188 (2018).
61. Merker, M. et al. Transcontinental spread and evolution of Mycobacterium tuberculosis W148 European/Russian clade toward extensively drug resistant tuberculosis. *Nat. Commun.* **13**, 5105 (2022).
62. Manca, C. et al. Mycobacterium tuberculosis CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. *J. Immunol.* **162**, 6740–6746 (1999).
63. Chitale, P. et al. A comprehensive update to the Mycobacterium tuberculosis H37Rv reference genome. *Nat. Commun.* **13**, 7068 (2022).
64. Comín, J. et al. The MtZ strain: molecular characteristics and Outbreak Investigation of the most successful Mycobacterium tuberculosis strain in Aragon using whole-genome sequencing. *Front. Cell. Infect. Microbiol.* **12**, 887134 (2022).
65. Dong, D. et al. PPE38 modulates the Innate Immune Response and is required for Mycobacterium marinum virulence. *Infect. Immun.* **80**, 43–54 (2012).
66. Sampson, S. L., Richardson, M., Van Helden, P. D. & Warren, R. M. IS *6110* -Mediated deletion polymorphism in isogenic strains of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **42**, 895–898 (2004).
67. Liu, S. et al. PE_PGRS31-S100A9 Interaction promotes mycobacterial survival in Macrophages through the regulation of NF-κB-TNF-α signaling and arachidonic acid metabolism. *Front. Microbiol.* **11**, 845 (2020).
68. Fang, W. et al. PE/PPE mutations in the transmission of Mycobacterium tuberculosis in China revealed by whole genome sequencing. *BMC Microbiol.* **24**, 206 (2024).
69. Gonzalo-Asensio, J. et al. New insights into the transposition mechanisms of IS6110 and its dynamic distribution between Mycobacterium tuberculosis Complex lineages. *PLoS Genet.* **14**, e1007282 (2018).
70. Reyes, A. et al. IS-seq: A novel high throughput survey of in vivo IS6110 transposition in multiple Mycobacterium tuberculosis genomes. *BMC Genom.* **13**, 249 (2012).
71. Warholm, P. & Light, S. Identification of a non-pentapeptide region associated with rapid mycobacterial evolution. *PLoS ONE* **11**, e0154059 (2016).
72. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
73. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
74. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: A genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011).
75. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
76. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).

Acknowledgements

Olabisi Flora Davies-Bolorunduro is supported by the International Postdoctoral Fellowship at Mahidol University. The work is funded by a Multidisciplinary Research grant and Mahidol University Strategic Research Fund (MU-SRF-WC-02B/66), and the National Science and Technology Development Agency Emerging Infectious Diseases Genomics program. We thank members of the Pornchai Matangkasombut Center for Microbial Genomics (CenMiG) team for useful discussions. We are grateful to TB/HIV Research Foundation and Chiangrai Prachanukroh Hospital for collecting sample and patient data.

Author contributions

OFD-B and PP conceived the study and wrote the manuscript, OFD-B, BJ, WR conducted sample acquisition sequence analyses. OFD-B and WR performed the gene reannotation and comparative genome analysis. OFD-B, BJ, TN and MB performed phylogenetic reconstruction, WR and TB performed genome assembly, WS, SM performed DNA extraction and sequencing. PP supervised the study. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-79351-w.

Correspondence and requests for materials should be addressed to P.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.