Heliyon 6 (2020) e05744

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Identification and *in silico* functional prediction of lineage-specific SNPs distributed in DosR-related proteins and resuscitation-promoting factor proteins of *Mycobacterium tuberculosis*

Pornpen Tantivitayakul [a], Tada Juthayothin [c], Wuthiwat Ruangchai [b], Nat Smittipat [c], Areeya Disratthakit [d], Surakameth Mahasirimongkol [d], Katsushi Tokunaga [e], Prasit Palittapongarnpim [b,c,*]

[a] *Department of Oral Microbiology, Faculty of Dentistry, Mahidol University, Bangkok, 10400, Thailand*
[b] *Pornchai Matangkasombut Center for Microbial Genomics, Department of Microbiology, Faculty of Science, Mahidol University, Rama 6 Road, Bangkok, Thailand*
[c] *National Centre for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, Phaholyothin Road, Pathumthani, Thailand*
[d] *Department of Medical Sciences, Ministry of Public Health, Tiwanon Road, Nonthaburi, Thailand*
[e] *Genome Medical Science Project, National Center for Global Health and Medicine, Tokyo, Japan*

ARTICLE INFO

Keywords:
Microbiology
Genetics
Molecular biology
Infectious disease
Lineage-specific SNP
DosR-related proteins
rpf proteins
T cell epitopes
Mycobacterium tuberculosis

ABSTRACT

One-third of the world population is infected by *Mycobacterium tuberculosis,* which may persist in the latent or dormant state. Bacteria can shift to dormancy when encountering harsh conditions such as low oxygen, nutrient starvation, high acidity and host immune defenses. Genes related to the dormancy survival regulator (DosR) regulon are responsible for the inhibition of aerobic respiration and replication, which is required to enter dormancy. Conversely, resuscitation-promoting factor (rpf) proteins participate in reactivation from dormancy and the development of active tuberculosis (TB). Many DosR regulon and rpf proteins are immunodominant T cell antigens that are highly expressed in latent TB infection. They could serve as TB vaccine candidates and be used for diagnostic development. We explored the genetic polymorphisms of 50 DosR-related genes and 5 *rpf* genes among 1,170 previously sequenced clinical *M. tuberculosis* genomes. Forty-three lineage- or sublineage-specific nonsynonymous single nucleotide polymorphisms (nsSNPs) were identified. Ten nsSNPs were specific to all Mtb isolates belonging to lineage 1 (L1). Two common sublineages, the Beijing family (L2.2) and EAI2 (L1.2.1), differed at as many as 26 lineage- or sublineage-specific SNPs. DosR regulon genes related to membrane proteins and the *rpf* family possessed mean dN/dS ratios greater than one, suggesting that they are under positive selection. Although the T cell epitope regions of DosR-related and rpf antigens were quite conserved, we found that the epitopes in L1 had higher rates of genetic polymorphisms than the other lineages. Some mutations in immunogenic epitopes of the antigens were specific to particular *M. tuberculosis* lineages. Therefore, the genetic diversity of the DosR regulon and rpf proteins might impact the adaptation of *M. tuberculosis* to the dormant state and the immunogenicity of latency antigens, which warrants further investigation.

1. Introduction

Mycobacterium tuberculosis (Mtb) is an insidious pathogen that can affect almost all human organs and all age groups of the global population. The number of deaths from tuberculosis (TB) worldwide was 1.5 million in 2018 (WHO, 2019). The organism has coevolved with the human host, allowing it to escape the host immune response and establish latent infection (Brites and Gagneux, 2015). An estimated two billion people have asymptomatic latent TB infection (LTBI). These people have a 5–10% probability of developing active TB following the initial infection (Barry et al., 2009). People with LTBI are major and important natural reservoirs of Mtb. Reactivation of Mtb in LTBI is a significant causative factor of pulmonary TB in adult patients (Lillebaek et al., 2002).

Latency, which results in delays in the pathogenesis and transmission of Mtb, is likely to be an important mechanism that allowed it to survive in human hosts when humans still lived in small groups as hunter-gatherers. The acquisition of agriculture and husbandry allowed

* Corresponding author.
E-mail address: Prasit.pal@mahidol.ac.th (P. Palittapongarnpim).

https://doi.org/10.1016/j.heliyon.2020.e05744
Received 10 September 2020; Received in revised form 21 October 2020; Accepted 11 December 2020

humans to live in settlements as large communities. This allowed Mtb to be transmitted to new hosts easily and thus decreased the significance of latency. The modern Mtb lineages (TbD1-negative lineages 2, 3 and 4) might have resulted from the adaptation of Mtb to the new human lifestyle (Gagneux, 2018). Lineage 2 (L2) comprises mostly the Beijing family, which is found all over the world, causes many outbreaks and appears to be expanding in many countries (Bifani et al., 2002). In Thailand, where L1 and L2 are equally predominant, L2 was more frequently associated with younger age (Ajawatanawong et al., 2019) and accounted for more genetic clustering than L1 (Miyahara et al., 2020). The possible explanations for this include the hypothesis that L2 patients are more infectious or L2-infected people have a shorter latency period. It is believed that pulmonary cavitation TB is more infectious than other forms of tuberculosis (Erkens et al., 2010), but the association of L2 with cavitation has never been clearly demonstrated. In contrast, it is known that W/Beijing strains constitutively express the *DosR* gene, which was suggested to be a mechanism supporting the success of the lineage (Fallow et al., 2010).

Mtb latency is primarily regulated by two sensor kinases, DosS/DosT, and one response regulator, DosR (Chauhan et al., 2011), and can be stimulated *in vitro* by hypoxia, nitric oxide (Voskuil et al., 2003), starvation (Betts et al., 2002) and acidity (Schnappinger et al., 2003). DosR affects the transcription of at least 50 genes (Park et al., 2003), including itself, which are collectively referred to as the dormancy survival regulator (DosR) regulon or DosR-related genes. DosR-related genes are involved in a wide variety of processes (Selvaraj et al., 2012) necessary for the transition of Mtb to the dormant state or latent infection. Genes belonging to the DosR regulon inhibit aerobic respiration and prevent *Mycobacterium* replication (Leistikow et al., 2010).

In contrast, shifting from latent to active disease requires resuscitation promoting factor (rpf) family proteins, which play roles in Mtb replication and reactivation. A Mtb *rpf*-deleted mutant showed impairment in reactivation during chronic infection in a mouse model (Biketov et al., 2007; Russell-Goldman et al., 2008). There are five paralogous rpf proteins (rpf A to E) in Mtb. They contain a conserved catalytic domain that is homologous to lysozyme (Cohen-Gonsaud et al., 2005). Hence, it was proposed that rpf proteins could cleave glycosidic bonds in peptidoglycan in the dormant cell wall. Subsequently, they participate in growth promotion and metabolic reactivation (Keep et al., 2006).

Many DosR-related and rpf proteins are immunogenic (Li et al., 2017; Zvi et al., 2008). They induced higher T-cell cytokine levels (especially IFN-γ response) in individuals with latent TB than in active TB patients (Black et al., 2009; Leyten et al., 2006; Schuck et al., 2009), indicating elevated gene expression during human latent infection. In addition, these antigens were experimentally proven to be potent T cell antigens that induced protective immunity (Black et al., 2009; Leyten et al., 2006; Schuck et al., 2009; Singh et al., 2014). Various DosR-related antigens (known as latency antigens) and rpf antigens are considered potential candidates for the development of multistage subunit vaccines and diagnostics for LTBI (Arroyo et al., 2018; Li et al., 2017; Niu et al., 2015). Currently, the multistage subunit vaccine is a novel approach for TB vaccine development that targets both active symptomatic and dormant asymptomatic phases (Khademi et al., 2018). The vaccines can be administered postexposure and have been developed to cope with LTBI and adult TB. Nevertheless, most studies have been based on the sequences of genes identified in the H37Rv strain, which belongs to L4. We recently demonstrated that the gene sequences of clinical isolates can be substantially different from those of the H37Rv strain and affect T cell epitopes (Tantivitayakul et al., 2020), which should be considered in vaccine development.

In the present study, we characterized genetic polymorphisms, especially lineage- and sublineage-specific single nucleotide polymorphisms (LS-SNPs), within the DosR regulon and *rpf* genes among 1,170 previously sequenced clinical isolates. We paid special attention to the differences between the ancestral (TbD1-positive) and modern (TbD1-negative) Mtb lineages as well as between L1 and L2. The LS-SNPs affecting the functions of the DosR regulons may result in differences in the latency of the bacteria and consequently the adaptability of the bacteria. Moreover, the immunogenic effects, especially on T cell epitopes, were also investigated.

2. Methods

2.1. Genome sequencing and variant calling

The genome sequences of 1,170 Mtb clinical isolates classified as 480 isolates of the Indo-Oceanic family (L1), 521 isolates of the East Asian family (L2), 11 isolates of lineage 3 (L3) and 158 of the Euro-American family (L4) were obtained from our previous study (Ajawatanawong et al., 2019). All L1 had East African Indian spoligotypes, while most L2 isolates belonged to sublineage L2.2, which is equivalent to the Beijing family. Whole genome sequencing (WGS) data from all clinical Mtb isolates were further analyzed in this study. The raw sequences were mapped to the H37Rv reference genome (GenBank accession number NC_000962_3) using the BWA program (Li and Durbin, 2009). Single nucleotide polymorphisms (SNPs) were called based on the alignment file using the GATK program (McKenna et al., 2010). The SNPs were called at a minimum depth of 20X with consensus quality scores greater than 20.

2.2. Identification of LS-SNPs

The phylogenetic trees based on 70,937 SNPs from 1,170 clinical Mtb strains that were analyzed by using maximum likelihood and Bayesian inference methods to define lineages and sublineages were previously reported (Ajawatanawong et al., 2019). LS-SNPs are the nucleotide bases present in all Mtb isolates belonging to a single lineage or sublineage that are not present in all other isolates. Nevertheless, if a sublineage contained fewer than 10 isolates, its LS-SNPs were not further studied. SNPs specific to ancestral lineages (Mtb L1, *M. africanum* L5 and L6) were determined by identifying L1-specific SNPs in the genome sequences of *M. africanum* MAL017004 (L5; accession no. KK338837) and *M. africanum* strain 25 (L6; accession no. CP010334).

2.3. Gene annotation and functional categories of DosR-related genes

Fifty DosR-related and 5 *rpf* genes (Table S1) were annotated by using the Mycobrowser database (Kapopoulou et al., 2011) (https://www.Mycobrowser.epfl.ch). The DosR-related genes were assigned to nine functional categories (Selvaraj et al., 2012; Singh et al., 2014), including *i*) redox balance metabolism and energy (n = 11), *ii*) nitrogen metabolism (n = 5), *iii*) nucleotide metabolism and repair (n = 4), *iv*) protein synthesis and cell wall synthesis (n = 2), *v*) sensor kinases and transcription regulators (n = 4), *vi*) host-pathogen interaction (n = 2), *vii*) membrane proteins (n = 4), *viii*) universal stress proteins (n = 7), and *ix*) hypothetical proteins or unknown function proteins (n = 11).

2.4. Prediction of the effects of LS-SNPs on protein function

To predict the functional consequences of SNPs, three computational tools, SNAP, PolyPhen-1 and SIFT, from the online consensus sequence prediction tool PredictSNP 1.0 (Bendl et al., 2014) were used.

2.5. Calculation of the dN/dS ratio

To assess selective pressures, the dN/dS ratio of each individual gene was calculated by using the kaks function of the *seqinr* package (Charif, 2007). The dN/dS ratio is defined as the ratio of nonsynonymous substitutions per nonsynonymous site (dN) to the number of synonymous substitutions per synonymous site (dS). A dN/dS ratio of less than one implies purifying selection, while a ratio of more than one indicates positive selection. A ratio of approximately one indicates a neutral mutation (Yang et al., 2000). The pairwise dN/dS of each Mtb isolate was calculated by using previously described protocols (Stucki et al., 2016). Briefly, the concatenated coding sequences of the 50 DosR-related and 5 *rpf* genes in each genome were compared with the homologous reference sequence of Mtb H37Rv (accession number NC_000962_3). After that, the concatenated epitope regions and concatenated nonepitope regions of the 55 target genes in each genome were constructed to evaluate the selective pressure against epitopes and nonepitopes, respectively. Statistical differences in the average dN/dS ratios between 4 Mtb lineages were evaluated using the nonparametric Kruskal-Wallis test (*p*-value of <0.05). The analyses were performed by SPSS software version 21 (SPSS, IL, USA).

2.6. T cell epitope analysis

The list of experimentally proven T cell epitopes of DosR-related antigens and rpf proteins in the H37Rv strain was obtained from the Immune Epitope Database (IEDB) (Vita et al., 2019) (https://www.iedb.org).

3. Results

3.1. Distribution of SNPs in 50 DosR-related genes and 5 rpf genes among 4 Mtb lineages

A total of 927 SNPs in coding regions and 170 SNPs in intergenic regions in the DosR regulon and *rpf* genes were identified. A total of 539 (58%) of the SNPs in coding sequences were singletons, while 89 (8%) were specific to entire Mtb lineages or sublineages. The others were found in some Mtb isolates but not an entire lineage. As L1 is the most evolutionarily distant from the H37Rv reference genome, 57 of the 89 LS-SNPs in the study were determined to belong to L1 or its sublineages. L1 isolates generally had a higher number of SNPs per genome than isolates from the other lineages. Therefore, we normalized the numbers of SNPs by dividing the number of SNPs occurring in 50 DosR-related and 5 *rpf* genes in each isolate by the total number of SNPs in the genome of the same isolate. The average ratio of isolates from the Mtb L1 lineage was significantly higher than that of isolates from the other lineages (Kruskal-Wallis test, $p < 0.001$). Furthermore, the average ratios of nonsynonymous (ns) SNPs to total SNPs of Mtb L1 and L4 were significantly higher than those of L2 and L3 (Kruskal-Wallis test, $p < 0.001$) (Figure 1).

3.2. LS-SNP distribution

There were 89 LS-SNPs with 75 in coding sequences, including a SNP that is present in all L2 and L3 isolates. Forty-three of the 75 (57%) LS-SNPs in coding regions were nonsynonymous. The numbers of LS-SNPs in DosR-related and *rpf* genes are shown in Table 1. The LS-SNPs are listed in Table 2 and supplementary Tables S2 and S3. Some intergenic LS-SNPs were found in DosR-binding sites upstream of the promoter of DosR-related genes (Table 2a) (Chauhan et al., 2011). Therefore, SNPs might affect the binding affinity of DosR and its recognition site and alter the expression level of downstream genes.

Of the 57 LS-SNPs among the L1 isolates, 16 SNPs were identified in all isolates in L1. As these SNPs were identified using H37Rv (an L4 isolate) as the reference, these SNPs signified the difference between the modern Mtb lineages L2-L4 and L1. Interestingly, 6 out of 16 SNPs were also found in *M. africanum* lineages 5 and 6 (Table 2 and Table S2). According to the fact that L1, L5 and L6 belong to a paraphyletic ancestral group (Gagneux, 2018), the 6 SNPs should actually be different between L2-L4 and the Mtb complex ancestor, with the original bases in the Mtb complex ancestor being the ones present in L1. The SNPs should be the result of mutations that occurred after the loss of the TbD1 segment during the evolution of the present modern lineages but before the branching of L2, L3 and L4, as shown in Figure 2. These mutations may have resulted in changes in latency during the adaptation of the modern Mtb lineages to agrarian society. One of the SNPs was found upstream of *hrp1* (hypoxic response protein 1), encoding a protein secreted by Mtb upon exposure to hypoxia that stimulates the pro-inflammatory response of macrophages (Sun et al., 2017). Most of the other SNPs were nonsynonymous and located in *DosS, narX, Rv0080* and *Rv2628*. Only the SNP in *ctpF* was synonymous. The Ile283Thr mutation in the major redox sensor histidine kinase, DosS, of the DosR regulon occurs in the GAF-B domain (Cho et al., 2008), which is one of the two regulatory domains of the DosS protein. DosS also modulates the autophagy pathway in a DosR regulon-independent manner (Gautam et al., 2019). *NarX* encodes a “fused nitrate reductase”, a protein with homology to parts of the 3 other nitrate reductases, NarG, NarJ, and NarI (Wang et al., 2011). Rv2628 is an immunogenic protein inducing strong IFN-γ production in individuals with latent TB infection (Goletti et al., 2010).

The other 10 SNPs were truly L1-specific. There were two L1-specific SNPs in intergenic regions that affected the predicted DosR binding sites of *Rv1996* (universal stress protein) and *Rv1997* (*ctpF*) (Table 2a and Figure 3). Remarkably, both genes also contained L1-specific nsSNPs in the coding regions. CtpF is a P-type ATPase that mediates calcium transport across the mycobacterial plasma membrane (Maya-Hoyos et al., 2019). Other nonsynonymous mutations were found in *Rv2003c* (conserved hypothetical protein) and *otsB1* (trehalose-6-phosphate phosphatase). Trehalose serves as a carbon and energy source for Mtb during dormancy (Shleeva et al., 2017). The true L1-specific SNPs resulted from mutations that occurred after the separation of *M. tuberculosis* sensu stricto into the L1 and modern Mtb lineages. Hence, they may be the result of the adaptation of L1 to agrarian society as well.

There were three nsSNPs specific to L1.2.1, which is also known as EAI2 according to spoligotyping. It is a common sublineage of L1 but localized to Southeast Asia, where L1.2.1 was found in approximately 80% of all TB patients in the Philippines (Phelan et al., 2019); L1.2.1.2 is common in Thailand and Myanmar (Palittapongarnpim et al., 2018). The L1.2.1-specific SNPs were located in the coding sequences of *narK2* (a proton/nitrate transporter) (Giffin et al., 2012), *rpfB* and *Rv1996* (universal stress protein) (Table 2b). rpfB is the sole member of rpf indispensable for resuscitation *in vivo* and has been investigated as a possible drug target (Ruggiero et al., 2013). *RpfB* deletion mutant strains showed

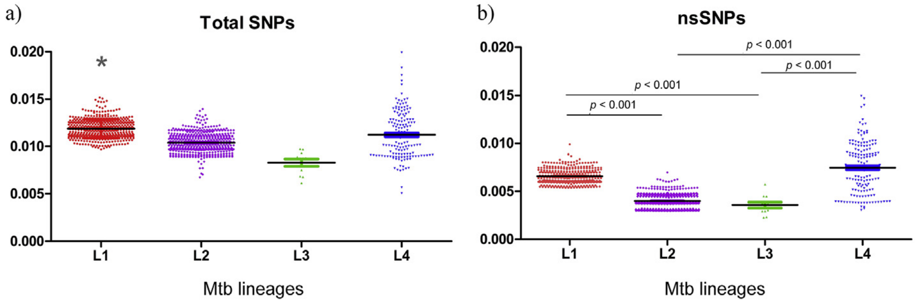

Figure 1. a) Ratios of the numbers of SNPs (nsSNPs and sSNPs) occurring in the 50 DosR-related and 5 *rpf* genes of each Mtb isolate per total number of SNPs present in the same genome. A total of 1,170 Mtb isolates were categorized as L1-L4. The asterisk indicates that the average ratio of Mtb L1 was significantly higher than that of the other lineages (Kruskal-Wallis test, $p < 0.001$), while b) the ratios of the nsSNPs of each isolate to the total SNPs in the same isolate of L1 and L4 were significantly higher than those of L2 and L3 (Kruskal-Wallis test, $p < 0.001$).

delayed reactivation in chronic TB infection in a mouse model (Tufariello et al., 2006).

Apart from these SNPs, there were also some SNPs specific to other sublineages of L1. Among these, three more nsSNPs were found in *ctpF* that were specific to L1.1.1.6, L1.1.3.3 and L1.2.2 (Figure 3). There were three more nsSNPs specific to L1.2.2 in *nrdZ* (a class II ribonucleotide reductase), *Rv2627* and *Rv2630*. The other 11 nsSNPs were specific to other minor sublineages, including three more that were specific to L1.1.1.6, as shown in Table 2b.

There was a synonymous SNP in the DosR regulon specific to all L2 and L3 isolates but none that were specific to all L2 isolates. There were 3 LS-SNPs specific to L2.1 or the proto-Beijing strains. However, due to the limited number of L2.1 isolates, the specific SNPs need to be confirmed. Two SNPs in *rpfE* and *Rv2629* were specific to L2.2, which is equivalent to the Beijing family. There was an additional nsSNP in *rpfB* specific to L2.2.1.1 (Pacific RD150) (Table 2b).

There was a synonymous LS-SNP in L4 and only one nonsynonymous LS-SNP in L3. As there were limited numbers of strains in this study, the L3-specific SNPs need further confirmation. There were three nsSNPs specific to L4.5, a common sublineage of L4 primarily found in East Asia (Brynildsrud et al., 2018), in *Rv0570*, *Rv0571c* and *otsB1*.

3.3. Selective pressures on DosR-related and rpf genes

To better understand the selective pressure on the DosR-related and *rpf* genes, we evaluated the dN/dS of the genes categorized into nine functional groups (Table S1). The average dN/dS values of all functional categories of DosR-related genes were less than one except for the membrane protein and conserved uncharacterized protein categories, indicating that most DosR-related genes were highly conserved (Figure 4a). In fact, only 8 DosR-related genes had dN/dS ratios greater than one (Table S1). In contrast to DosR-related genes, most *rpf* genes had high dN/dS ratios, with rpfB and rpfE values of 5.52 and 1.85, respectively. Moreover, only nonsynonymous mutations were identified in *rpfC* and *rpfD,* so their dN/dS values could not be calculated. Only rpfA had a dN/dS ratio of less than 1.

The selective pressure was different for each lineage (Figure 4b). Mtb L4 and L1 had significantly higher average dN/dS ratios than L2 and L3 (Kruskal-Wallis test, $p < 0.001$). The L4 isolates had a wide range of dN/dS values, with 55 isolates (35%) having dN/dS ratios higher than one.

3.4. LS-SNPs affecting human T cell epitopes in the latency and rpf antigens

In addition to physiological functions, DosR-related and rpf proteins may be affected by host immune responses (Arroyo et al., 2016). The dN/dS ratios of the T cell and non-T cell epitopes were analyzed. Based on the Immune Epitope Database (IEDB), we identified 27 dormancy-related and 3 rpf proteins containing 243 experimentally proven T cell epitopes. These antigens induce IFN-γ secretion by T lymphocytes (Black et al., 2009; Leyten et al., 2006; Schuck et al., 2009; Singh et al., 2014). The average dN/dS value of proteins harboring T cell epitopes is 0.87, which is slightly less than the value of 0.99 of proteins without known T cell epitopes. Seventy-five of the T cell epitopes (31%) were found to carry amino acid substitutions in this study. Six nonsynonymous LS-SNPs in experimentally proven T cell epitopes were identified and are shown in Table 3.

Genetic variations in T cell epitopes and non-T cell epitopes among the 4 different Mtb lineages were analyzed. Both epitope and nonepitope regions were under purifying selection. This conformed to the studies of Coscolla et al. (2015) and Comas et al. (2010). Nevertheless, the T cell epitopes in L1 had a significantly higher mean dN/dS ratio than those in other lineages (Kruskal-Wallis test, $p < 0.001$) (Figure 5a). Nonepitope regions in L1 and L4 had significantly higher mean ratios than those in L2 and L3 isolates (Kruskal-Wallis test, $p < 0.001$) (Figure 5b). Furthermore, the dN/dS of T cell epitopes of DosR-related and rpf antigens in L1 had a higher mean than the non-T cell epitopes, suggesting the lower purifying pressure on the T cell epitopes in this lineage.

4. Discussion

In general, single nucleotide mutations occur randomly throughout the genome. Some may have devastating effects on the encoded proteins, which decrease the fitness of the mutated bacteria. They would therefore disappear from the population quickly. Some mutations are evolutionarily neutral and may randomly disappear or may persist in the population essentially by chance. Some rare mutations that increase the fitness of the bacteria enhance the chance of the bacteria surviving, expanding and becoming recognizable clades. Sublineages are usually defined as clades with a considerable number of members, which may reflect their adequate fitness to thrive in human populations. Some LS-SNPs may contribute to fitness and hence become interesting targets for investigations. In contrast, some LS-SNPs may be merely associated mutations. As DosR-related and *rpf* genes are involved in an important part of the life cycle of Mtb, mutations that confer advantageous phenotypes are expected to be fairly frequent and, if they persist, may be recognized as LS-SNPs, conforming to the findings in this study. The fitness of a sublineage may also be attributed to other types of mutations, such as insertions/deletions.

The Mtb complex has evolved for millennia and currently comprises 7 major lineages that cause TB in humans, lineages 1–7, based on SNP phylogeny. Lineages 1, 5 and 6 are called “Ancestral”, as they harbor the

Table 1. The distribution of total SNPs and LS-SNPs in 50 DosR-related genes and 5 *rpf* genes among 4 major Mtb lineages. The percentages of LS-SNPs were calculated by dividing the number of LS-SNPs by the total number of SNPs identified in each lineage.

SNP types	Lineage (no. of Mtb isolates)			
	L1 (480)	L2 (521)	L3 (11)	L4 (158)
- Total number of SNPs in 50 DosR-related and 5 *rpf* genes in each lineage[#]	531	227	20	159
- Number of LS-SNP in 50 DosR-related and 5 *rpf* genes in each lineage	57	15[*]	4[*]	14
- Number of all LS-SNPs in intergenic regions (%)	12 (2.3%)	0	1 (5%)	1 (0.6%)
Number of SNPs specific to major lineages or common sublineages in intergenic region[$] (%)	3	0	1	1
- Number of all LS-SNPs in coding region (%)	45 (8.5%)	15 (6.6%)	3 (15%)	13 (8.2%)
Number of LS-nonsynonymous SNPs/LS-synonymous SNPs	29/16	7/8	1/2	6/7
- Number of SNPs specific to major lineages or common sublineages in coding region[$] (%)	21 (3.9%)	7[*] (3.1%)	3[*] (12%)	5 (3.1%)
Number of LS-Nonsynonymous SNPs/LS-Synonymous SNPs	12/9	2/5	1/2	3/2

SNPs residing within 500 bp upstream of the start codons from annotated genes were identified as intergenic SNPs.

[#] 10 SNPs were found in more than one Mtb lineages.

[*] A synonymous SNP in a coding region was present in all isolates of both L2 and L3.

[$] represents major lineages (L1, L2, L3, L4) and common Mtb sublineages which comprise more than 90 Mtb isolates including L1.1.1 (269), L1.2.1 (108) and L4.5 (93).

Table 2. Lists of lineage-specific SNPs in intergenic regions and LS-nonsynonymous SNPs in coding regions.

SNP position	REF	ALT	No. of isolates	Lineage	Changing nucleotide sequence at
a) Intergenic SNPs specific to major Mtb lineages. The other SNPs in intergenic regions were listed in Table S3.					
2238930	A	C	480	L1	Primary DosR binding site upstream *Rv1996* TTAGGGACCATCGCCTCCTG to TTCGGGACCATCGCCTCCTG
2240062	C	G	480	L1	Primary DosR binding site upstream *Rv1997 (ctpF)* CTGGACCGTAGGTCCCTG to CTGGACCGTAGGTCGCTG
2056184	G	A	11	L3	72 bp upstream to start codon of *Rv1813c*
2953307	C	A	480	L 1*	314 bp upstream to start codon of *Rv2626c (hrp1)*

Position	Gene (Rv)	Gene Annotation	Functional Category	NT change	AA change	Lineage	Amino acid change	Predicted SNP effect by		
								PP-1	SIFT	SNAP
b) List of LS-nonsynonymous SNPs in coding regions with prediction of the mutational effect on the protein function. The full list of LS-SNP was in Table S2.										
SNP specific to major Mtb lineages or common sublineages$^{\$}$										
89200	*Rv0080*	Uncharacterized protein	HP	179G/T	Gly60Val	L 1*	Hydrophobic to hydrophobic	N	N	N
661524	*Rv0570*	Ribonucleoside-diphosphate reductase, nrdZ	NMR	230T/C	Leu77Pro	L 4.5	Hydrophobic to hydrophobic	**A**	**A**	N
663803	*Rv0571c*	Phosphoribosyl transferase	NMR	1016C/T	Ala339Val	L 4.5	Hydrophobic to hydrophobic	**A**	**A**	**A**
1128814	*Rv1009*	Resuscitation promoting factor, rpfB	RPF	724G/A	Glu242Lys	L 1.2.1	Acidic to basic	N	**N**	N
1963957	*Rv1736c*	Nitrate reductase-like protein, narX	NM	230A/G	Asp77Gly	L 1*	Acidic to hydrophobic	**A**	N	N
1964913	*Rv1737c*	Nitrate/nitrite transporter, narK2	NM	458G/A	Arg153Gln	L 1.2.1	Basic to polar	N	N	N
2239160	*Rv1996*	Universal stress protein	SP	157G/C	Gly53Arg	L 1.2.1	Hydrophobic to basic	N	N	N
2239055	*Rv1996*	Universal stress protein	SP	52C/T	Pro18Ser	L 1	Hydrophobic to polar	**A**	**A**	**A**
2242808	*Rv1997*	Cation-transporting ATPase F, ctpF	MP	2650G/A	Ala884Thr	L 1	Hydrophobic to polar	N	N	**A**
2249035	*Rv2003c*	Uncharacterized protein	RD	386T/C	Met129Thr	L 1	Hydrophobic to polar	N	N	N
2253701	*Rv2006*	Glycosyl hydrolase, otsB1	RD	1700C/G	Ala567Gly	L 4.5	Hydrophobic to hydrophobic	N	N	N
2255942	*Rv2006*	otsB1	RD	3941G/T	Arg1314Leu	L 1	Basic to hydrophobic	N	N	**A**
2275764	*Rv2029c*	Phosphofructokinase B, pfkB	RD	661C/T	Leu221Phe	L 3	Hydrophobic to hydrophobic	**A**	N	N
2752122	*Rv2450c*	Resuscitation promoting factor, rpfE	RPF	59C/G	Thr20Arg	L 2.2	Polar to basic	N	N	**A**
2955233	*Rv2628*	Uncharacterized protein	HP	176C/T	Ser59Leu	L 1*	Polar to hydrophobic	N	**A**	N
2955957	*Rv2629*	Uncharacterized protein	HP	191A/C	Asp64Ala	L 2.2	Acidic to hydrophobic	N	N	**A**
3496264	*Rv3130c*	Diacyglycerol O-acyl transferase, tgs1	RD	103G/A	Ala35Thr	L 1.1.1	Hydrophobic to polar	N	N	N
3498418	*Rv3132c*	Redox sensor histidine kinase response regulator, devS	TR	848T/C	Ile283Thr	L 1*	Hydrophobic to polar	N	N	N
SNP specific to sublineages										
89179	*Rv0080*	Uncharacterized protein	HP	158T/C	Val53Ala	L 1.1.3.1	Hydrophobic to hydrophobic	N	**A**	N
89272	*Rv0080*	Uncharacterized protein	HP	251A/G	Tyr84Cys	L 1.1.2.1	polar to basic	A	**A**	N
89454	*Rv0080*	Uncharacterized protein	HP	433A/G	Ile145Val	L 1.1.3.2	Hydrophobic to hydrophobic	N	N	N
89623	*Rv0081*	HTH-type transcriptional regulator	TR	49C/T	Leu17Phe	L 1.1.1.2	Hydrophobic to hydrophobic	**A**	**A**	N
90111	*Rv0082*	Probable oxidoreductase	RD	188A/G	Glu63Gly	L 2.1	Acidic to hydrophobic	N	A	N
91649	*Rv0083*	Probable oxidoreductase	RD	1250C/T	Ala417Val	L 2.1	Hydrophobic to hydrophobic	N	A	N
661929	*Rv0570*	nrdZ	NMR	635C/A	Ala212Asp	L 1.2.2	Hydrophobic to acidic	A	A	A
662624	*Rv0570*	nrdZ	NMR	1330A/G	Ile444Val	L 2.1	Hydrophobic to hydrophobic	N	**N**	N
662900	*Rv0570*	nrdZ	NMR	1606C/T	Pro536Ser	L 1.1.3.3	Hydrophobic to polar	A	**A**	**A**
664065	*Rv0571c*	Phosphoribosyl transferase	NMR	754C/T	Pro252Ser	L 1.1.1.2	Hydrophobic to polar	N	N	N
665293	*Rv0572c*	Uncharacterized protein	HP	91T/C	Phe31Leu	L 1.2.1.1	Hydrophobic to hydrophobic	A	N	N
1128883	*Rv1009*	rpfB	RPF	793G/A	Val265Met	L 2.2.1.1	Hydrophobic to hydrophobic	A	**A**	A
2241296	*Rv1997*	ctpF	MP	1138G/A	Ala380Thr	L 1.1.1.6	Hydrophobic to polar	N	**A**	**A**
2241494	*Rv1997*	ctpF	MP	1336G/A	Glu446Lys	L 1.1.3.3	Acidic to Basic	**N**	**A**	**N**

(*continued on next page*)

Table 2 (*continued*)

Position	Gene (Rv)	Gene Annotation	Functional Category	NT change	AA change	Lineage	Amino acid change	Predicted SNP effect by		
								PP-1	SIFT	SNAP
2242238	*Rv1997*	ctpF	MP	2080G/A	Val694Ile	L 1.2.2	Hydrophobic to hydrophobic	**N**	N	N
2242860	*Rv1997*	ctpF	MP	2702G/A	Arg901Gln	L 4.5.3	Basic to polar	**A**	**A**	**A**
2248981	*Rv2003c*	Uncharacterized protein	RD	440C/T	Pro147Leu	L 4.5.3	Hydrophobic to hydrophobic	**A**	**A**	**A**
2254519	*Rv2006*	otsB1	RD	2518A/G	Thr840Ala	L 2.2 Asia Ancestral 4	Polar to hydrophobic	**N**	**N**	**N**
2254852	*Rv2006*	otsB1	RD	2851G/C	Gly951Arg	L 1.1.2.1	Hydrophobic to basic	**N**	**A**	A
2278333	*Rv2030c*	Uncharacterized protein	HP	154C/A	Pro52Thr	L 1.1.1.6	Hydrophobic to polar	N	N	**A**
2950836	*Rv2624c*	Universal stress protein	SP	472A/T	Thr158Ser	L 1.1.1.6	Polar to polar	N	N	N
2951648	*Rv2625c*	Putative zinc metalloprotease	MP	856G/A	Asp286Asn	L 1.1.1.6	Acidic to polar	N	N	N
2953871	*Rv2627c*	Uncharacterized protein	HP	878G/T	Gly293Val	L 1.2.2	Hydrophobic to hydrophobic	**N**	**N**	**N**
2954571	*Rv2627c*	Uncharacterized protein	HP	178T/G	Leu60Val	L 4.2	Hydrophobic to hydrophobic	A	**A**	A
2957418	*Rv2630*	Probable proteinarchease	NMR	526A/G	Thr176Ala	L 1.2.2	Polar to hydrophobic	N	N	N

REF = nucleotide sequence of Mtb reference H37Rv strain.
ALT = nucleotide sequence changing from Mtb reference H37Rv strain.
NT change = nucleotide sequence changing from Mtb reference H37Rv strain.
AA change = amino acid sequence changing from Mtb reference H37Rv strain.
Mutational effect on protein function was predicted by three programs, polyphen-1 (PP-1), SIFT, SNAP.
*corresponds to SNPs specific to Ancestral Mtb strains including Mtb L1, *M. africanum* L5 and L6.
*SNPs specific to Ancestral Mtb lineages including Mtb L1, *M. africanum* L5 and L6.
$ represents major lineages (L1, L2, L3, L4) and common Mtb sublineages which comprise more than 90 Mtb isolates including L1.1.1 (269), L1.2.1 (108) and L4.5 (93).

TbD1 DNA segment, which is similar to that in the Mtb ancestor. L2, 3 and 4 lost the TbD1 segment and are designated "Modern", as they may be more highly adapted to modern human lifestyles that emerged in larger and denser settlements. Latency is thought to be essential for Mtb survival in small human populations to allow the bacteria to wait for new generations of unimmunized hosts. In large human populations, in which unimmunized members are born continuously, the need for the long latency of Mtb may not be significant, and rapid transmission may be a more attractive survival strategy. This phenomenon was supported by the success of the Beijing family (L2.2) with shorter latency periods as well as possibly higher infectivity. Mutations specific to the Beijing family in the DosR regulon and *rpf* genes are therefore of particular interest.

Generally, LS-SNPs are regarded as informative and are used as genetic markers for discriminating bacterial populations. Many researchers have increasingly focused on exploring the functional roles of LS-SNPs. An interesting study by Rose et al. (2013) revealed a SNP specific to the Beijing family (C3500149T) upstream of the start codon of *DosR*. The Beijing-specific SNP creates a new transcriptional start site (TSS) in DosR and results in constitutive expression of DosR and DosR-related genes. This evidence was supported by the transcriptome analysis, which revealed that the Mtb Beijing family carrying the C3500149T SNP had a higher gene expression level of the DosR regulon than other Mtb lineages (Domenech et al., 2017; Homolka et al., 2010). We confirmed the ubiquitous presence of the C3500149T SNP in the Beijing family as a synonymous LS-SNP in *Rv3134c,* as shown in Table S2.

L2 is classified into two major sublineages, the rare L2.1 and very common L2.2 lineages (the Beijing family). The considerable differences between the incidences of both sublineages suggested their differences in fitness and conformed to the finding that L2.1 and L2.2 did not share any SNPs in the DosR regulon and *rpf* genes, which suggested that they had developed different latency processes.

We identified two more Beijing-specific nsSNPs in the coding regions of *Rv2629* and *rpfE*. The 191A/C mutation (Asp64Ala) in *Rv2629* is a well-known Beijing-specific SNP (Homolka et al., 2009; Zhang et al., 2014). Rv2629 overexpression delayed entry of Mtb into the exponential growth phase (Liu et al., 2017). The Rv2629 antigen contains several epitopes that induce strong cytotoxic T cell responses, which are predictably restricted by HLA-A2, and have been suggested to be vaccine candidates (Bai et al., 2018). Furthermore, the 59C/G mutation in *rpfE* alters the amino acid residue from threonine (polar) to arginine (positively charged) and disrupts the hydrophobic region of the signal peptide (Mukamolova et al., 2002), which normally interacts with the secA

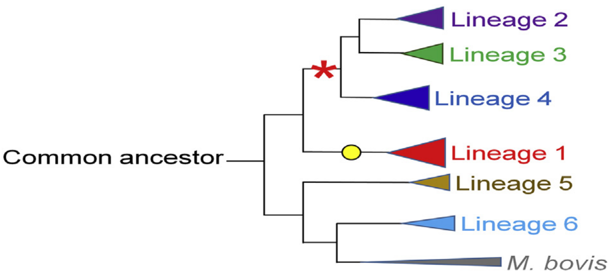

Figure 2. A diagram showing the proposed evolutionary path of the *M. tuberculosis* complex (Comas et al., 2013). The ancestral Mtb lineages comprise L1 (Indo-Oceanic) and *M. africanum* L5 and L6. The modern lineages are characterized by the loss of the TbD1 segment, and they branched into L4 (Euro-American), L2 (Beijing family) and L3 (CAS). The asterisk represents the loss of the TbD1 segment and the presence of the 6 SNPs that were different between L2-L4 and L1, L5 and L6, while the circle represents the 10 truly L1-specific SNPs.

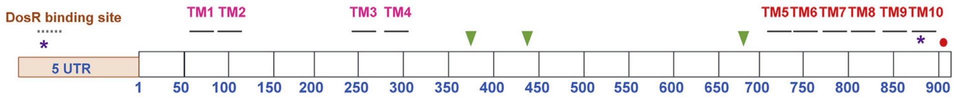

Figure 3. A diagram of ctpF, which is composed of 905 amino acids with 10 transmembrane segments (TM), as shown in the black bar. The DosR binding site upstream of the start codon was mapped as a dotted line. The locations of LS-SNPs affecting the DosR binding site and the amino acid sequence of ctpF are shown. The asterisks indicate SNPs specific to all L1 isolates in the DosR binding site and the TM10 region. The triangles represent the sublineage-specific nsSNPs of L1, and the circles represent the sublineage-specific SNPs of L4.5.3.

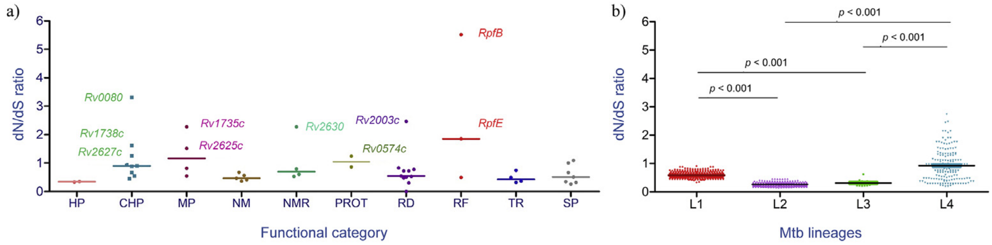

Figure 4. a) dN/dS ratios of 5 *rpf* genes and 50 DosR-related genes belonging to 10 functional categories (Singh et al., 2014) including *i)* host-pathogen interactions (HP, n = 2), *ii)* uncharacterized proteins (CHP, n = 11), *iii)* membrane proteins (MP, n = 4), *iv)* nitrogen metabolism (NM, n = 5), *v)* nucleotide metabolism and repair (NMR, n = 4), *vi)* protein synthesis and cell wall synthesis (PROT, n = 2), *vii)* redox balance metabolism and energy (RD, n = 11), *viii)* resuscitation-promoting factor (RF, n = 5), *ix)* sensor kinases and transcription regulators (TR, n = 4), and *x)* universal stress proteins (SP, n = 7). The bars show the average dN/dS ratios. b) Pairwise dN/dS ratios of 50 DosR-related genes and 5 *rpf* genes among 4 major Mtb lineages. *represents a significant difference with $p <$ 0.001, as calculated by the nonparametric Kruskal-Wallis test.

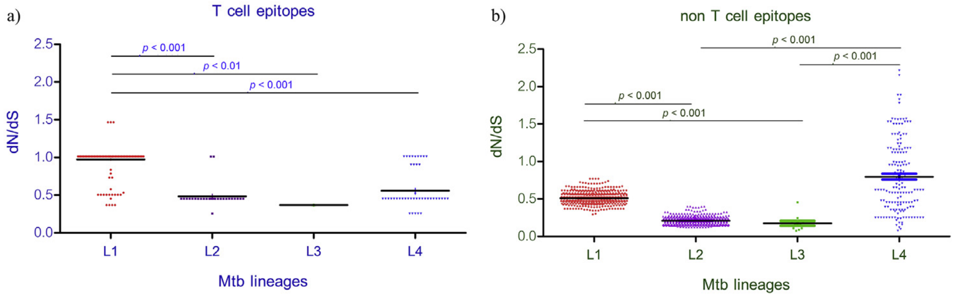

Figure 5. Comparison of the dN/dS ratios of a) concatenated T cell epitope and b) concatenated non-T cell epitope regions in 27 dormancy-related and 3 rpf antigens among isolates belonging to 4 Mtb lineages. Statistical analysis of the dN/dS values of T cell epitopes and non-T cell epitopes was performed with the nonparametric Kruskal-Wallis test. Only in L1 were the dN/dS ratios of T cell epitopes significantly higher than those of non-T cell epitopes (Kruskal-Wallis test, $p < 0.001$).

component (Mori et al., 1997) to transport the rpfE protein across the cell membrane. Hence, the mutation in the signal peptide region might interfere with the translocation of rpfE. How the mutation of *rpfE* affects the resuscitation of Mtb from dormancy is still unknown. However, rpfE interacts with the TLR-4 receptor of dendritic cells and induces Th1 and Th17 cell activation (Choi et al., 2015). This results in the secretion of IFN-γ, IL-2 and IL-17A, which play a major role in the host response to Mtb. Hence, the mutation may partially explain the relatively low induction of the immune response and the resulting benefit to the Beijing strains. It should be noted that the Leu330Arg substitution in rpfB was

Table 3. Nonsynonymous LS-SNPs affecting experimentally-proven T cell epitopes.

Gene (Rv)	Epitope	IEDB_ID	Epitope sequences	LS-SNP (No. isolates)	NT change	AA change
Rv1733c	T	177591	AGTAV**Q**DSRSHVYAH	All Lineages except L4.8	G204T	Gln68His
Rv2029c (pfkB)	T	154515	EPEQLAAAHE**L**IDRGRAEVV	L3 (11)	C661T	Leu221Phe
	T	155102	**L**IDRGRAEVVVVSLGSQGAL			
Rv2627	T	38529	LPIARPTIA**L**AAQAFRDEIV	L4.2 (10)	T178G	Leu60Val
	T	4520	ASLEEGLACAIL**G**VPVADLI	L1.2.2 (18)	G878T	Gly293Val
Rv2628	T	106585	KVQSATIYQVTDR**S**H	L1 (480)	C176T	Ser59Leu
Rv1009 (rpfB)	T	229352	LPVANVVVTPAHEA**V**	L2.2.1.1 (22)	G793A	Val265Met

identified in 111 of 146 (76%) Mtb isolates belonging to L2.2.1, which is also known as Asia Ancestral 3 and is the most common Ancestral Beijing sublineage in Thailand. The amino acid residue at position 330 is located within the transglycosylase catalytic domain of rpfB, which forms a hydrophobic pocket that binds the N-acetylglucosamine moiety (NAG) of peptidoglycan (Squeglia et al., 2013). Therefore, the change from leucine (hydrophobic) to arginine (positively charged) should affect the functions of the protein.

We identified additional L1-specific SNPs in the DosR-related genes. The differences in 16 LS-SNPs between L1 and the modern Mtb lineages suggest that there are some differences in the control of latency between L1 and L2-4. Some mutations occurred during the evolution of modern Mtb strains before the separation of L2, L3 and L4, which warrants further study. Several mutations also occurred during the evolution of L1 before its separation into sublineages. Among the DosR-related genes, *ctpF* appeared to be frequently affected. The protein contributes to calcium efflux, and its defect impairs tolerance to oxidative and nitrosative stress (Maya-Hoyos et al., 2019). The multiple mutations found in L1 suggested some changes in the latency process of L1 compared to that of the Mtb common ancestor, and L1 may not be as phenotypically ancient as its name implies.

The finding that the dN/dS ratio of most rpf proteins is more than one suggests that they are under positive selective pressure. This might be because the rpf proteins are located on the cell surface of *M. tuberculosis*. These molecules are exposed to the external environment and putatively target the host immune system. Therefore, genetic variations identified in rpf proteins are likely to result in improved survival for the bacteria.

Our study demonstrated that the T cell epitopes in L1 had higher genetic diversity than those in other lineages. Variation in the epitope regions in L1 may represent ongoing evolution, which may benefit the bacterial population by promoting, for example, evasion of T cell recognition or interaction with various HLA alleles among different human populations. Remarkably, the variation in epitopes in L1 may have an impact on the efficacy of candidate vaccine antigens comprising DosR-related and rpf antigens that use Mtb H37Rv as a reference sequence.

5. Conclusion

This study demonstrated that there were considerable variations in SNPs in the DosR regulon and rpf family specific to various Mtb lineages and sublineages. Two common sublineages, the Beijing family or L2.2 and EAI2 or L1.2.1, differed at as many as 26 SNPs. These factors should affect the process of dormancy and reactivation. Thus, the information revealed in this study would be useful for further analysis of the effects of SNPs on TB phenotypes, including adaptation to dormant states as well as differential induction of host immune responses against latency antigens.

Declarations

Author contribution statement

Pornpen Tantivitayakul, Prasit Palittapongarnpim: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Tada Juthayothin, Wuthiwat Ruangchai, Nat Smittipat, Areeya Disratthakit, Surakameth Mahasirimongkol, Katsushi Tokunaga: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by the Science and Technology Research Partnership for Sustainable Development (SATREPS) project (Grant no. JP17jm0110010) by AMED and the Japan International Cooperation Agency (JICA) and Department of Medical Sciences, Ministry of Public Health, Thailand. PT and PP were supported by Mahidol University Research Center grants.

Data availability statement

Data associated with this study has been deposited at European Nucleotide Archive (ENA) under the accession number ERP006738.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e05744.

References

Ajawatanawong, P., Yanai, H., Smittipat, N., Disratthakit, A., Yamada, N., Miyahara, R., Nedsuwan, S., Imasanguan, W., Kantipong, P., Chaiyasirinroje, B., Wongyai, J., Plitphonganphim, S., Tantivitayakul, P., Phelan, J., Parkhill, J., Clark, T.G., Hibberd, M.L., Ruangchai, W., Palittapongarnpim, P., Juthayothin, T., Thawornwattana, Y., Viratyosin, W., Tongsima, S., Mahasirimongkol, S., Tokunaga, K., Palittapongarnpim, P., 2019. A novel Ancestral Beijing sublineage of *Mycobacterium tuberculosis* suggests the transition site to Modern Beijing sublineages. Sci. Rep. 9, 13718.

Arroyo, L., Marin, D., Franken, K., Ottenhoff, T.H.M., Barrera, L.F., 2018. Potential of DosR and Rpf antigens from *Mycobacterium tuberculosis* to discriminate between latent and active tuberculosis in a tuberculosis endemic population of Medellin Colombia. BMC Infect. Dis. 18, 26.

Arroyo, L., Rojas, M., Franken, K.L., Ottenhoff, T.H., Barrera, L.F., 2016. Multifunctional T cell response to DosR and rpf antigens is associated with protection in long-term *Mycobacterium tuberculosis*-infected individuals in Colombia. Clin. Vaccine Immunol. 23, 813–824.

Bai, X., Wang, D., Liu, Y., Xiao, L., Liang, Y., Yang, Y., Zhang, J., Lin, M., Wu, X., 2018. Novel epitopes identified from *Mycobacterium tuberculosis* antigen Rv2629 induces cytotoxic T lymphocyte response. Immunol. Lett. 203, 21–28.

Barry 3rd, C.E., Boshoff, H.I., Dartois, V., Dick, T., Ehrt, S., Flynn, J., Schnappinger, D., Wilkinson, R.J., Young, D., 2009. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. Nat. Rev. Microbiol. 7, 845–855.

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J., Damborsky, J., 2014. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput. Biol. 10, e1003440.

Betts, J.C., Lukey, P.T., Robb, L.C., McAdam, R.A., Duncan, K., 2002. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. Mol. Microbiol. 43, 717–731.

Bifani, P.J., Mathema, B., Kurepina, N.E., Kreiswirth, B.N., 2002. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol. 10, 45–52.

Biketov, S., Potapov, V., Ganina, E., Downing, K., Kana, B.D., Kaprelyants, A., 2007. The role of resuscitation promoting factors in pathogenesis and reactivation of *Mycobacterium tuberculosis* during intra-peritoneal infection in mice. BMC Infect. Dis. 7, 146.

Black, G.F., Thiel, B.A., Ota, M.O., Parida, S.K., Adegbola, R., Boom, W.H., Dockrell, H.M., Franken, K.L., Friggen, A.H., Hill, P.C., Klein, M.R., Lalor, M.K., Mayanja, H., Schoolnik, G., Stanley, K., Weldingh, K., Kaufmann, S.H., Walzl, G., Ottenhoff, T.H., Consortium, G.B.f.T., 2009. Immunogenicity of novel DosR regulon-encoded candidate antigens of *Mycobacterium tuberculosis* in three high-burden populations in Africa. Clin. Vaccine Immunol. 16, 1203–1212.

Brites, D., Gagneux, S., 2015. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. Immunol. Rev. 264, 6–24.

Brynildsrud, O.B., Pepperell, C.S., Suffys, P., Grandjean, L., Monteserin, J., Debech, N., Bohlin, J., Alfsnes, K., Pettersson, J.O., Kirkeleite, I., Fandinho, F., da Silva, M.A., Perdigao, J., Portugal, I., Viveiros, M., Clark, T., Caws, M., Dunstan, S., Thai, P.V.K., Lopez, B., Ritacco, V., Kitchen, A., Brown, T.S., van Soolingen, D., O'Neill, M.B., Holt, K.E., Feil, E.J., Mathema, B., Balloux, F., Eldholm, V., 2018. Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation. Sci. Adv. 4, eaat5869.

Charif, D.L., 2007. SeqinR 1.0-2: a contributed package to the R-project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla, U., Porto, M.E.R.H., Vendruscolo, M. (Eds.), Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Springer Verlag.

Chauhan, S., Sharma, D., Singh, A., Surolia, A., Tyagi, J.S., 2011. Comprehensive insights into *Mycobacterium tuberculosis* DevR (DosR) regulon activation switch. Nucleic Acids Res. 39, 7400–7414.

Cho, H.Y., Cho, H.J., Kim, Y.M., Oh, J.I., Kang, B.S., 2008. Crystallization and preliminary crystallographic analysis of the second GAF domain of DevS from *Mycobacterium smegmatis*. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 64, 274–276.

Choi, H.G., Kim, W.S., Back, Y.W., Kim, H., Kwon, K.W., Kim, J.S., Shin, S.J., Kim, H.J., 2015. *Mycobacterium tuberculosis* RpfE promotes simultaneous Th1- and Th17-type T-cell immunity via TLR4-dependent maturation of dendritic cells. Eur. J. Immunol. 45, 1957–1971.

Cohen-Gonsaud, M., Barthe, P., Bagneris, C., Henderson, B., Ward, J., Roumestand, C., Keep, N.H., 2005. The structure of a resuscitation-promoting factor domain from *Mycobacterium tuberculosis* shows homology to lysozymes. Nat. Struct. Mol. Biol. 12, 270–273.

Comas, I., Chakravartti, J., Small, P.M., Galagan, J., Niemann, S., Kremer, K., Ernst, J.D., Gagneux, S., 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat. Genet. 42, 498–503.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S.R., Niemann, S., Diel, R., Aseffa, A., Gao, Q., Young, D., Gagneux, S., 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat. Genet. 45, 1176–1182.

Coscolla, M., Copin, R., Sutherland, J., Gehre, F., de Jong, B., Owolabi, O., Mbayo, G., Giardina, F., Ernst, J.D., Gagneux, S., 2015. *M. tuberculosis* T cell epitope analysis reveals paucity of antigenic variation and identifies rare variable TB antigens. Cell Host Microbe 18, 538–548.

Domenech, P., Zou, J., Averback, A., Syed, N., Curtis, D., Donato, S., Reed, M.B., 2017. Unique regulation of the DosR regulon in the beijing lineage of *Mycobacterium tuberculosis*. J. Bacteriol. 199.

Erkens, C.G., Kamphorst, M., Abubakar, I., Bothamley, G.H., Chemtob, D., Haas, W., Migliori, G.B., Rieder, H.L., Zellweger, J.P., Lange, C., 2010. Tuberculosis contact investigation in low prevalence countries: a European consensus. Eur. Respir. J. 36, 925–949.

Fallow, A., Domenech, P., Reed, M.B., 2010. Strains of the East Asian (W/Beijing) lineage of *Mycobacterium tuberculosis* are DosS/DosT-DosR two-component regulatory system natural mutants. J. Bacteriol. 192, 2228–2238.

Gagneux, S., 2018. Ecology and evolution of *Mycobacterium tuberculosis*. Nat. Rev. Microbiol. 16, 202–213.

Gautam, U.S., Mehra, S., Kumari, P., Alvarez, X., Niu, T., Tyagi, J.S., Kaushal, D., 2019. *Mycobacterium tuberculosis* sensor kinase DosS modulates the autophagosome in a DosR-independent manner. Commun. Biol. 2, 349.

Giffin, M.M., Raab, R.W., Morganstern, M., Sohaskey, C.D., 2012. Mutational analysis of the respiratory nitrate transporter NarK2 of *Mycobacterium tuberculosis*. PloS One 7, e45459.

Goletti, D., Butera, O., Vanini, V., Lauria, F.N., Lange, C., Franken, K.L., Angeletti, C., Ottenhoff, T.H., Girardi, E., 2010. Response to Rv2628 latency antigen associates with cured tuberculosis and remote infection. Eur. Respir. J. 36, 135–142.

Homolka, S., Koser, C., Archer, J., Rusch-Gerdes, S., Niemann, S., 2009. Single-nucleotide polymorphisms in *Rv2629* are specific for *Mycobacterium tuberculosis* genotypes Beijing and Ghana but not associated with rifampin resistance. J. Clin. Microbiol. 47, 223–226.

Homolka, S., Niemann, S., Russell, D.G., Rohde, K.H., 2010. Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: delineation of conserved core and lineage-specific transcriptomes during intracellular survival. PLoS Pathog. 6, e1000988.

Kapopoulou, A., Lew, J.M., Cole, S.T., 2011. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. Tuberculosis (Edinb). 91, 8–13.

Keep, N.H., Ward, J.M., Robertson, G., Cohen-Gonsaud, M., Henderson, B., 2006. Bacterial resuscitation factors: revival of viable but non-culturable bacteria. Cell. Mol. Life Sci. 63, 2555–2559.

Khademi, F., Derakhshan, M., Yousefi-Avarvand, A., Tafaghodi, M., Soleimanpour, S., 2018. Multi-stage subunit vaccines against *Mycobacterium tuberculosis*: an alternative to the BCG vaccine or a BCG-prime boost? Expert Rev. Vaccines 17, 31–44.

Leistikow, R.L., Morton, R.A., Bartek, I.L., Frimpong, I., Wagner, K., Voskuil, M.I., 2010. The *Mycobacterium tuberculosis* DosR regulon assists in metabolic homeostasis and enables rapid recovery from nonrespiring dormancy. J. Bacteriol. 192, 1662–1670.

Leyten, E.M., Lin, M.Y., Franken, K.L., Friggen, A.H., Prins, C., van Meijgaarden, K.E., Voskuil, M.I., Weldingh, K., Andersen, P., Schoolnik, G.K., Arend, S.M., Ottenhoff, T.H., Klein, M.R., 2006. Human T-cell responses to 25 novel antigens encoded by genes of the dormancy regulon of *Mycobacterium tuberculosis*. Microb. Infect. 8, 2052–2060.

Li, F., Kang, H., Li, J., Zhang, D., Zhang, Y., Dannenberg Jr., A.M., Liu, X., Niu, H., Ma, L., Tang, R., Han, X., Gan, C., Ma, X., Tan, J., Zhu, B., 2017. Subunit vaccines consisting of antigens from dormant and replicating bacteria show promising therapeutic effect against *Mycobacterium bovis* BCG latent infection. Scand. J. Immunol. 85, 425–432.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Lillebaek, T., Dirksen, A., Baess, I., Strunge, B., Thomsen, V.O., Andersen, A.B., 2002. Molecular evidence of endogenous reactivation of *Mycobacterium tuberculosis* after 33 years of latent infection. J. Infect. Dis. 185, 401–404.

Liu, D., Hao, K., Wang, W., Peng, C., Dai, Y., Jin, R., Xu, W., He, L., Wang, H., Wang, H., Zhang, L., Wang, Q., 2017. Rv2629 overexpression delays *Mycobacterium smegmatis* and *Mycobacteria tuberculosis* entry into log-phase and increases pathogenicity of *Mycobacterium smegmatis* in mice. Front. Microbiol. 8, 2231.

Maya-Hoyos, M., Rosales, C., Novoa-Aponte, L., Castillo, E., Soto, C.Y., 2019. The P-type ATPase CtpF is a plasma membrane transporter mediating calcium efflux in *Mycobacterium tuberculosis* cells. Heliyon 5, e02852.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Miyahara, R., Smittipat, N., Juthayothin, T., Yanai, H., Disratthakit, A., Imsanguan, W., Intralawan, D., Nedsuwan, S., Chaiyasirinroje, B., Bupachat, S., Tokunaga, K., Mahasirimongkol, S., Palittapongarnpim, P., 2020. Risk factors associated with large clusters of tuberculosis patients determined by whole-genome sequencing in a high-tuberculosis-burden country. Tuberculosis (Edinb) 125, 101991.

Mori, H., Araki, M., Hikita, C., Tagaya, M., Mizushima, S., 1997. The hydrophobic region of signal peptides is involved in the interaction with membrane-bound SecA. Biochim. Biophys. Acta 1326, 23–36.

Mukamolova, G.V., Turapov, O.A., Young, D.I., Kaprelyants, A.S., Kell, D.B., Young, M., 2002. A family of autocrine growth factors in *Mycobacterium tuberculosis*. Mol. Microbiol. 46, 623–635.

Niu, H., Peng, J., Bai, C., Liu, X., Hu, L., Luo, Y., Wang, B., Zhang, Y., Chen, J., Yu, H., Xian, Q., Zhu, B., 2015. Multi-Stage tuberculosis subunit vaccine candidate LT69 provides high protection against *Mycobacterium tuberculosis* infection in mice. PloS One 10, e0130641.

Palittapongarnpim, P., Ajawatanawong, P., Viratyosin, W., Smittipat, N., Disratthakit, A., Mahasirimongkol, S., Yanai, H., Yamada, N., Nedsuwan, S., Imasanguan, W., Kantipong, P., Chaiyasirinroje, B., Wongyai, J., Toyo-Oka, L., Phelan, J., Parkhill, J., Clark, T.G., Hibberd, M.L., Ruengchai, W., Palittapongarnpim, P., Juthayothin, T., Tongsima, S., Tokunaga, K., 2018. Evidence for host-bacterial Co-evolution via genome sequence analysis of 480 Thai *Mycobacterium tuberculosis* lineage 1 isolates. Sci. Rep. 8, 11597.

Park, H.D., Guinn, K.M., Harrell, M.I., Liao, R., Voskuil, M.I., Tompa, M., Schoolnik, G.K., Sherman, D.R., 2003. Rv3133c/dosR is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis*. Mol. Microbiol. 48, 833–843.

Phelan, J.E., Lim, D.R., Mitarai, S., de Sessions, P.F., Tujan, M.A.A., Reyes, L.T., Medado, I.A.P., Palparan, A.G., Naim, A.N.M., Jie, S., Segubre-Mercado, E., Simoes, B., Campino, S., Hafalla, J.C., Murase, Y., Morishige, Y., Hibberd, M.L., Kato, S., Ama, M.C.G., Clark, T.G., 2019. *Mycobacterium tuberculosis* whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. Sci. Rep. 9, 9305.

Rose, G., Cortes, T., Comas, I., Coscolla, M., Gagneux, S., Young, D.B., 2013. Mapping of genotype-phenotype diversity among clinical isolates of *Mycobacterium tuberculosis* by sequence-based transcriptional profiling. Genome Biol. Evol. 5, 1849–1862.

Ruggiero, A., Marchant, J., Squeglia, F., Makarov, V., De Simone, A., Berisio, R., 2013. Molecular determinants of inactivation of the resuscitation promoting factor B from *Mycobacterium tuberculosis*. J. Biomol. Struct. Dyn. 31, 195–205.

Russell-Goldman, E., Xu, J., Wang, X., Chan, J., Tufariello, J.M., 2008. A *Mycobacterium tuberculosis Rpf* double-knockout strain exhibits profound defects in reactivation from chronic tuberculosis and innate immunity phenotypes. Infect. Immun. 76, 4269–4281.

Schnappinger, D., Ehrt, S., Voskuil, M.I., Liu, Y., Mangan, J.A., Monahan, I.M., Dolganov, G., Efron, B., Butcher, P.D., Nathan, C., Schoolnik, G.K., 2003. Transcriptional adaptation of *Mycobacterium tuberculosis* within macrophages: insights into the phagosomal environment. J. Exp. Med. 198, 693–704.

Schuck, S.D., Mueller, H., Kunitz, F., Neher, A., Hoffmann, H., Franken, K.L., Repsilber, D., Ottenhoff, T.H., Kaufmann, S.H., Jacobsen, M., 2009. Identification of T-cell antigens specific for latent *Mycobacterium tuberculosis* infection. PloS One 4, e5590.

Selvaraj, S., Sambandam, V., Sardar, D., Anishetty, S., 2012. In silico analysis of DosR regulon proteins of *Mycobacterium tuberculosis*. Gene 506, 233–241.

Shleeva, M.O., Trutneva, K.A., Demina, G.R., Zinin, A.I., Sorokoumova, G.M., Laptinskaya, P.K., Shumkova, E.S., Kaprelyants, A.S., 2017. Free trehalose accumulation in dormant *Mycobacterium smegmatis* cells and its breakdown in early resuscitation phase. Front. Microbiol. 8, 524.

Singh, S., Saraav, I., Sharma, S., 2014. Immunogenic potential of latency associated antigens against Mycobacterium tuberculosis. Vaccine. 32, 712–716.

Squeglia, F., Romano, M., Ruggiero, A., Vitagliano, L., De Simone, A., Berisio, R., 2013. Carbohydrate recognition by RpfB from *Mycobacterium tuberculosis* unveiled by crystallographic and molecular dynamics analyses. Biophys. J. 104, 2530–2539.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., Fenner, L., Rutaihwa, L., Borrell, S., Luo, T., Gao, Q., Kato-Maeda, M., Ballif, M., Egger, M., Macedo, R., Mardassi, H., Moreno, M., Tudo Vilanova, G., Fyfe, J., Globan, M., Thomas, J., Jamieson, F., Guthrie, J.L., Asante-Poku, A., Yeboah-Manu, D., Wampande, E., Ssengooba, W., Joloba, M., Henry Boom, W., Basu, I., Bower, J., Saraiva, M., Vaconcellos, S.E.G., Suffys, P., Koch, A., Wilkinson, R., Gail-Bekker, L., Malla, B., Ley, S.D., Beck, H.P., de Jong, B.C., Toit, K., Sanchez-Padilla, E., Bonnet, M., Gil-Brusola, A., Frank, M., Penlap Beng, V.N., Eisenach, K., Alani, I., Wangui Ndung'u, P., Revathi, G., Gehre, F., Akter, S., Ntoumi, F., Stewart-Isherwood, L., Ntinginya, N.E., Rachow, A., Hoelscher, M., Cirillo, D.M., Skenders, G., Hoffner, S., Bakonyte, D., Stakenas, P., Diel, R., Crudu, V., Moldovan, O., Al-Hajoj, S., Otero, L., Barletta, F., Jane Carter, E., Diero, L., Supply, P., Comas, I., Niemann, S., Gagneux, S., 2016. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nat. Genet. 48, 1535–1543.

Sun, C., Yang, G., Yuan, J., Peng, X., Zhang, C., Zhai, X., Luo, T., Bao, L., 2017. *Mycobacterium tuberculosis* hypoxic response protein 1 (Hrp1) augments the pro-inflammatory response and enhances the survival of *Mycobacterium smegmatis* in murine macrophages. J. Med. Microbiol. 66, 1033–1044.

Tantivitayakul, P., Ruangchai, W., Juthayothin, T., Smittipat, N., Disratthakit, A., Mahasirimongkol, S., Viratyosin, W., Tokunaga, K., Palittapongarnpim, P., 2020. Homoplastic single nucleotide polymorphisms contributed to phenotypic diversity in Mycobacterium tuberculosis. Sci Rep. 10, 8024.

Tufariello, J.M., Mi, K., Xu, J., Manabe, Y.C., Kesavan, A.K., Drumm, J., Tanaka, K., Jacobs Jr., W.R., Chan, J., 2006. Deletion of the *Mycobacterium tuberculosis* resuscitation-promoting factor Rv1009 gene results in delayed reactivation from chronic tuberculosis. Infect. Immun. 74, 2985–2995.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., Peters, B., 2019. The immune epitope database (IEDB): 2018 update. Nucleic Acids Res. 47, D339–D343.

Voskuil, M.I., Schnappinger, D., Visconti, K.C., Harrell, M.I., Dolganov, G.M., Sherman, D.R., Schoolnik, G.K., 2003. Inhibition of respiration by nitric oxide induces a *Mycobacterium tuberculosis* dormancy program. J. Exp. Med. 198, 705–713.

Wang, X., Wang, H., Xie, J., 2011. Genes and regulatory networks involved in persistence of *Mycobacterium tuberculosis*. Sci. China Life Sci. 54, 300–310.

Who, 2019. Global Tuberculosis Report 2019.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.

Zhang, L., Xu, W., Cui, Z., Liu, Y., Wang, W., Wang, J., Hu, D., Liu, D., Wang, H., 2014. A novel method of identifying *Mycobacterium tuberculosis* Beijing strains by detecting SNPs in *Rv0444c* and *Rv2629*. Curr. Microbiol. 68, 381–386.

Zvi, A., Ariel, N., Fulkerson, J., Sadoff, J.C., Shafferman, A., 2008. Whole genome identification of *Mycobacterium tuberculosis* vaccine candidates by comprehensive data mining and bioinformatic analyses. BMC Med. Genom. 1, 18.