Journal of Tuberculosis Research, 2014, 2, 181-198
Published Online December 2014 in SciRes. http://www.scirp.org/journal/jtr
http://dx.doi.org/10.4236/jtr.2014.24023

Scientific Research

Evolution of PE35 and PPE68 Gene Families in *Mycobacterium*: Roles of Horizontal Gene Transfer and Evolutionary Constraints

Ashay Bavishi, Lin Lin, Madhusudan Choudhary, Todd P. Primm*

Department of Biological Sciences, Sam Houston State University, Huntsville, USA
Email: *tprimm@shsu.edu

Received 20 October 2014; revised 24 November 2014; accepted 4 December 2014

Abstract

***Mycobacterium* is a genus of bacteria with over a hundred non-pathogenic and pathogenic species, best recognized for certain members known to cause diseases such as tuberculosis and leprosy. Two novel protein families important in the pathogenesis of *Mycobacterium* species are the PE and PPE families. These two protein families affect the antigenic profiles, disturbing host immunity. To better understand the origin and evolution of these gene families and the differences in their composition between pathogenic and non-pathogenic strains, several bioinformatic analyses were conducted both among *Mycobacterium* and closely related species that contain PE35 and PPE68 gene homologs. The methods included protein homology searches (BLASTP), horizontal gene transfer analysis (IslandViewer), phylogenetic analysis, gene cluster analysis and structural and functional constraints. Results revealed that PE and PPE gene homologs were not only limited to *Mycobacterium*, but also existed in three other non-mycobacterial genera, *Rhodococcus*, *Tsukamurella* and *Segniliparus*, and were possibly initially acquired from non-mycobacterial microorganisms by multiple horizontal gene transfers. Results also demonstrated that PE and PPE genes were more diverse and more rapidly evolving in pathogenic *Mycobacterium* as compared with non-pathogenic *Mycobacterium* and other non-mycobacterial species. These findings possibly shed light on the diverse functions and origins of the PE/PPE proteins among these organisms.**

Keywords

PE35, PPE68, Horizontal Gene Transfer, *Mycobacterium*

1. Introduction

Mycobacteria are a genus of acid-fast bacteria with over a hundred non-pathogenic and pathogenic species, best

*Corresponding author.

How to cite this paper: Bavishi, A., Lin, L., Choudhary, M. and Primm, T.P. (2014) Evolution of PE35 and PPE68 Gene Families in *Mycobacterium*: Roles of Horizontal Gene Transfer and Evolutionary Constraints. *Journal of Tuberculosis Research*, **2**, 181-198. http://dx.doi.org/10.4236/jtr.2014.24023

recognized for certain members such as *Mycobacterium tuberculosis* and *Mycobacterium leprae*, which cause the diseases tuberculosis [1] and leprosy [2], respectively. Growth rate and pathogenicity are two defining characteristics of the *Mycobacterium* species. Pathogenic strains generally grow slowly, forming colonies on solid media in several weeks to months, while environmental species are non-pathogenic and grow rapidly within a week [3] [4].

Pathogenicity is defined as the ability of a microorganism to cause disease that is harmful to the host. Furthermore, pathogenic characteristics result from specific interactions between the pathogen and its host, and the components of these characteristics are coded by their genomes. PE and PPE gene families are a surprising discovery in the genome of *M. tuberculosis* and they represent ~10% of its genome [5] [6]. The genome of *M. tuberculosis* strain H37Rv is annotated with 99 and 69 PE and PPE proteins, respectively. Members of the PE protein family have conserved ~110 amino acid N-terminal domains with the proline-glutamic acid motif at positions 8 - 9 [5] [7] [8]. Members of the PPE family have longer ~180 amino acid conserved N-terminal domains with the proline-proline-glutamic acid motif at positions 7 - 9 [5] [8]. Compared with the conserved N-terminal domains, the C-terminal domains of both PE and PPE vary in sequence and length, often containing repetitive regions.

PE and PPE genes are further classified into subfamilies based on consensus motifs in the C-terminus regions [8]. For example, the polymorphic GC-rich-repetitive sequence (PGRS), which contains repeats of glycine-glycine-alanine or glycine-glycine-asparagine, is the largest PE subfamily of *M. tuberculosis* [8] [9]. On the other hand, the PPE family has two major subfamilies: SVP (GxxSVPxxW) and major polymorphic tandem repeat (MPTR) [5] [10]. Several genes in PE-PGRS subfamilies serve as cell surface constituents necessary for cell-cell interactions, cellular structures and infectivity of host cells [1] [11]-[15]. These genes encode surface-exposed proteins associated the cell wall, which provide antigenic diversity. A majority of PE and PPE coding genes are differentially expressed under different experimental growth conditions [6] [16]-[19], which suggests that the differential expression of these genes may contribute to the varied antigenic potential to the changing microenvironments within the host. Further, several PE and PPE family members are secreted by the ESX-5 protein secretion system in *M. marinum* [20]. This is consistent with ESX-5 being related to ESAT-6, and PE/PPE gene members often being located within ESAT-6 clusters in mycobacteria.

Fast growing species possess fewer PE/PPE genes than slow growing species suggesting that these gene families originated in rapidly growing mycobacteria and then laterally transferred into and expanded further in slow growing mycobacterial species. However, some slow growing species, such as *M. ulcerans* and *M. avium* subsp. *Paratuberculosis*, have fewer number of PE/PPE genes [21]. PE and PPE genes from the ESAT-6 cluster region 1, *M. tuberculosis* Rv3872 (PE35) and Rv3873 (PPE68) respectively, were considered as the ancestors of PE/PPE families [8]. PE35 and PPE68 were more conserved among MTB strains as compared with other PE and PPE family members, and this idea was compatible with the gene duplication model for ESAT-6 clusters at that time. The authors also concluded, consistent with others, that no PE/PPE gene homologs were present in species outside of the genus *Mycobacterium*.

The two following hypotheses were investigated in the present study. First, PE and PPE genes were initially acquired by horizontal gene transfers (HGT) from related species and then those genes further spread into multiple copies using gene duplication and/or transposition in the genome. Second, PE/PPE genes in pathogenic mycobacterial species evolved faster than their homologs in the non-pathogenic species, and therefore they would be less evolutionary constrained and more duplicated than the homologs present in non-pathogenic species. In this study, the identified ancestral PE (PE35) and PPE (PPE68) genes were analyzed among mycobacterial and related species to ascertain their possible evolutionary relationships. A number of bioinformatics approaches (protein homologies, evolutionary constraints, horizontal gene transfer analyses, gene cluster analyses and phylogenetic analyses) were used to ascertain the evolutionary relationship of PE/PPE proteins and the level of selection they experienced.

2. Methods

2.1. Identification of PE35/PPE68 Gene Pair Homologs among Species

Since the PE35 and PPE68 gene pair are considered the most ancestral genes of the two respective gene families [8], protein similarity searches were conducted to determine the highest matches for PE and PPE protein-pairs.

To help elucidate the origin and evolution of this pair across different organisms, the PE35 and PPE68 protein sequences for each *Mycobacterium* were compared against the National Center for Biotechnology Information (NCBI) microbial database using BLASTP [22]. Proteins were selected using the lowest e-value homolog found for respective organisms.

The reference PPE68 and PE35 proteins were selected from *Mycobacterium tuberculosis* CDC1551 (100% identical in H37Rv) and blasted against fully sequenced *Mycobacterium* genomes found in the NCBI database. For each *Mycobacterium*, the protein with lowest e-value was selected for further analysis. Subsequently, each of these selected PPE68/PE35 protein homologs was blasted against fully sequenced non-*Mycobacterium* organisms in the NCBI microbial database. Given the potential rapid divergence of such gene families and the likelihood of ancient relationships between gene homologs, a simple filtering of genes based on a single measure such as e-value or bit-score would likely miss relevant relationships. Rather, all genes which showed homology to *Mycobacterium* reference sequences were selected for further scrutiny. For these genes, inspection of sequences and similarity regions was performed to ascertain which genes were to be included in subsequent analysis. If multiple genes within a single organism were found with significant homology to the reference strains, then usually the gene with lowest e-value was selected.

For the organisms for which a significant match was obtained with a PE/PPE gene, their corresponding 16S rRNA gene sequences were additionally obtained to serve as a comparison group. Upon identification of homologs, full length DNA and protein sequences were obtained through the NCBI database using their respective accession numbers. The 16S rRNA accession numbers can be found listed in **Table 1** and the PE35/PPE68 numbers in **Table S3**.

2.2. Functional Constraints Analysis

Protein sequence alignments were carried out using MUSCLE (Multiple Sequence Comparison by Log-Expectation) [23], a program known for its accuracy and speed. For the functional constraints analysis, comparisons were conducted across all *Mycobacterium* strains whose genomes were sequenced and annotated. More specifically, the functional constraints analyses were performed independently for the PE and PPE genes. The synonymous rates (K_s) and nonsynonymous substitution rates (K_a) along with the nonsynonymous-synonymous substitution rate ratio (ω) were calculated using the modified Yang-Nielsen method [24] [25] using K_a/K_s calculator [26].

2.3. Phylogenetic Analysis

Geneious 4.6 was used to organize and perform the protein similarity searches, generate alignments, and construct phylogenetic trees [27]. Only organisms with completely sequenced genomes were chosen to avoid poor or incomplete sequence data from shotgun or partial genome sequencing projects.

The 16S rRNA nucleotide sequences as well as PE35 and PPE68 homolog nucleotide sequences for all species were obtained from the NCBI gene database. Phylogenetic analysis was performed using PhyML [28] with the Tamura-Nei (TN) model [29] to generate unrooted, maximum likelihood trees. For all trees, bootstrap values were calculated using 100 replications**.**

2.4. Gene Cluster Analysis

For all relevant organisms, information concerning gene location, direction, and content was obtained from the NCBI database. More specifically, the genes adjacent to the PE35 and PPE68 homologs were analyzed in content, direction, and length to look for similarities and differences in organization and structure. Relative gene maps were then constructed showing the distribution of genes around PE35/PPE68. Some gene maps were subsequently grouped together to provide for a visual comparison of related gene clusters.

2.5. PE/PPE Copy Number

The N-terminal domains of PE35 (conserved N-terminal domain, residues 5-162) and PPE68 (conserved N-terminal domain, residues 5-163) as identified by Pfam [30] were compared to their respective genomes using BLASTP. All proteins with homology above an e-value of 0.01 were then further classified as PPE or PE.

Table 1. Genomic characteristics of*Mycobacterium*and their related species.

Species	Accession Number	Genome Size (Kb)	GC (%)	Number of Protein Coding Genes	Protein Coding (%)	Growth Type	Potential Host Preference
Mycobacterial Species							
M. abscessus ATCC 19977	NC_010397.1	5,067,172	64.1	4,920	92	fast	Pathogenic
M. avium 104	NC_008595.1	5,475,491	69.0	5,120	88	slow	Pathogenic
M. avium subsp. *Paratuberculosis* K-10	NC_002944.2	4,829,781	69.3	4,350	91	slow	Pathogenic
M. bovis AF2122/97	NC_002945.3	4,345,492	65.6	3,918	90	slow	Pathogenic
M. bovis BCG str. Pasteur 1173P2	NC_008769.1	4,374,522	65.6	3,949	90	slow	Pathogenic
M. bovis BCG str. Tokyo 172	NC_012207.1	4,371,711	65.6	3,944	90	slow	Pathogenic
M. gilvum PYR-GCK	NC_009338.1	5,619,607	67.7	5,241	92	fast	Non-Pathogenic
M. leprae Br4923	NC_011896.1	3,268,071	57.8	1,604	49	slow	Pathogenic
M. leprae TN	NC_002677.1	3,268,203	57.8	1,605	49	slow	Pathogenic
M. marinum M	NC_010612.1	6,636,827	65.7	5,423	89	slow	Pathogenic
M. smegmatis str. MC^2 155	NC_008596.1	6,988,209	67.4	6,716	90	fast	Non-Pathogenic
M. sp. JLS	NC_009077.1	6,048,425	68.4	5,739	92	fast	Non-Pathogenic
M. sp. KMS	NC_008705.1	5,737,227	68.2	5,460	92	fast	Non-Pathogenic
M. sp. MCS	NC_008146.1	5,705,448	68.4	5,391	92	fast	Non-Pathogenic
M. sp. Spyr1	NC_014814.1	5,547,747	67.0	5,130	91	fast	Non-Pathogenic
M. tuberculosis CDC1551	NC_002755.2	4,403,837	65.6	4,189	90	slow	Pathogenic
M. tuberculosis F11	NC_009565.1	4,424,435	65.6	3,941	90	slow	Pathogenic
M. tuberculosis H37Ra	NC_009525.1	4,419,977	65.6	4,034	90	slow	Pathogenic
M. tuberculosis H37Rv	NC_000962.2	4,411,532	65.6	3,988	90	slow	Pathogenic
M. tuberculosis KZN 1435	NC_012943.1	4,398,250	65.6	4,059	91	slow	Pathogenic
M. ulcerans Agy99	NC_008611.1	5,631,606	65.4	4,160	72	slow	Pathogenic
M. vanbaalenii PYR-1	NC_008726.1	6,491,865	67.8	5,979	91	fast	Non-Pathogenic
Mycobacterial-Related Species							
Rhodococcus equi 103S	NC_014659	5,043,170	68.0	4,512	90	N/A	Pathogenic
Rhodococcus equi ATCC 33707	NZ_ADNW00000000	5,255,557	68.0	5,030	91	N/A	Pathogenic
Rhodococcus erythropolis SK121	NZ_ACNO00000000	6,785,398	62.0	6,713	91	N/A	Pathogenic
Rhodococcus jostii RHA1	NC_008268	7,804,765	67.0	7,211	91	N/A	Pathogenic
Rhodococcus opacus B4	NC_012522	7,913,450	67.0	7,246	91	N/A	Non-Pathogenic
Segniliparus rotundus DSM 44985	NC_014168	3,157,527	66.0	3,006	90	N/A	Pathogenic
Segniliparus rugosus ATCC BAA974	NZ_ACZI00000000	3,567,567	68.0	3,516	88	N/A	Pathogenic
Tsukamurella paurometabola DSM 20162	NC_014158	4,379,918	68.0	4,157	91	N/A	Pathogenic

N/A: not Applicable; Accession number is for the respective complete genome sequence.

3. Results

3.1. Genome Characteristics and Life Styles of Mycobacterial Species

The detailed genome and lifestyle characteristics of 22 *Mycobacterium* and 8 other related species are shown in **Table 1**. *Mycobacterium* species exhibit varying levels of genome sizes, ranging from ~3.3 Mbp of *M. leprae* to ~7.0 Mbp of *M. smegmatis*. The percentage GC content varies from 57.8% of *M. leprae* to 69.3% of *M. bovis*. Genomes of most mycobacterial species have ~90% coding capabilities with the exception of *M. leprae* and *M. ulcerans* which have 49% and 72% coding capabilities, respectively.

3.2. PE35 and PPE68 Protein Families in *Mycobacterium*

Pairwise amino acid identities between PPE68 and PE35 homologs across the genomes of mycobacteria and the related species are listed in **Table S1** and **Table S2**, respectively.

The protein homology searches indicate that PE and PPE proteins are not limited to the genus *Mycobacterium*. More specifically, organisms within the genera *Rhodococcus*, *Segniliparus*, and *Tsukamurella* were found to contain genes with significant homology (>50%) to the ancestralPE35 and PPE68 genes found in *Mycobacterium*. Consistent with this, the genes from the non-mycobacterial species are annotated as PE/PPE family members (except the PE68 homolog entry in *R. opacus*, listed as a hypothetical protein, but does include the PE domain) in their respective database entries (**Table S3**).

The blast of the conserved N-terminal domains, for both PE35 and PPE68, to their own respective genomes shows that pathogenic *Mycobacterium* have high copy numbers of PPE68 homologs (≥30 gene copies) as compared to non-pathogenic (<10 PPE gene copies) with an exception of *M. leprae*, whose genome contains only 4 gene copies of PPE (**Table 2**). The numbers of PE homologs in mycobacteria were lower than the numbers of PPE copies in their respective genomes, except in three non-pathogenic species, including *M. smegmatis*. In addition, pathogenic mycobacterial species contain more copies of PE as compared to the PE gene copies in non-pathogenic mycobacterial species. Pathogenic *Mycobacterium* contain 8.3 ± 5.4 PE genes and 51.2 ± 30.5 PPE genes, while non-pathogenic contain 3.1 ± 1.9 PE and 2.6 ± 0.8 PPE genes. Thus pathogenic species have experienced a strong expansion of the PPE family. Other related species copy numbers for PE/PPE were not greater than seven combined. Three of the eight species did not have any noted PE genes: *Rhodococcus equi* ATCC 33707, *Rhodococcus jostii* RHA1, and *Tsukamurella paurometabola* DSM 20162.

PPE68 and PE35 GC composition were relatively similar to their genome GC composition in mycobacteria. However, within the related species three significant differences of PPE GC content to genome GC content were noticed. *Rhodococcus equi*103S PPE GC content is 76.4% while its genome GC content is 68%. Similarly, *Rhodococcus jostii* RHA1 and *Rhodococcus opacus* B4 have PPE GC contents of 71.7% and 73.7%, respectively, while their genome GC composition are both 67%, suggesting recent HGT.

None of the PE/PPE genes were found in predicted HGT regions except for the PE35 gene homologs of *M. smegmatis* and *M. avium* subsp. *Paratuberculosis*. Data was obtained using the program IslandViewer and its datasets [31]. However, PE35 and PPE68 genes in other species were not found to be in HGT regions identified by employing the IslandViewer program.

3.3. Phylogenetic Analysis of *Mycobacterium* and Other Related Species

Phylogenetic relationships based on 16S rRNA gene sequences are a standard tool to reflect evolutionary histories of species. As such, the phylogenetic tree shown in **Figure 1** revealed that slow growing pathogenic species and rapid growing environmental species of *Mycobacterium* form two distinct evolutionary groups, as found in numerous other studies. Also, organisms within the genera *Rhodococcus*, *Segniliparus*, and *Tsukamurella*, which contain genes with significant homology to PE/PPE genes found in *Mycobacterium*, are found together as an "out-group" at the base of the tree and are distantly related to mycobacterial species. The closest relative to-mycobacteria was *Tsukamurella paurometabola* and the most distant was *Segniliparus rotundus*. The *Rhodococcus* genus is grouped between the *Tsukamurella* and *Segniliparus* genera.

The phylogenetic trees based on PPE68 and PE35 homolog gene sequences are shown in **Figure 2** and **Figure 3**, respectively, and these two gene trees do not completely parallel the 16S ribosomal tree shown in **Figure 1**, or each other. The PPE68 tree, as shown in **Figure 2**, reveals that genes in the MTB complex, *Mycobacterium leprae*, *and Mycobacterium marinum* genes are closely related to *Tsukamurella paurometabola*, while non-patho-

Table 2. Sequence characteristics of PE35 and PPE68 homologs of *Mycobacterium* and their related species.

Species	GC Content of PPE68 (%)	Number of PPE Family	GC Content of PE35 (%)	Number of PE Family	Genome GC Content	Potential Host Preference
M. abscessus ATCC 19977	66.9	7	72.2	3	64.1	Pathogenic
M. avium 104	70.4	35	77.7	5	69.0	Pathogenic
***M. avium* subsp. *Paratuberculosis* K-10**	68.1	35	77.0	7	69.3	Pathogenic
M. bovis AF2122/97	67.5	61	65.3	12	65.6	Pathogenic
M. bovis BCG str. Pasteur 1173P2	64.4	58	66.7	6	65.6	Pathogenic
M. bovis BCG str. Tokyo 172	64.4	58	66.7	5	65.6	Pathogenic
M. gilvum PYR-GCK	69.1	3	70.7	4	67.7	Non-Pathogenic
M. leprae Br4923	59.0	4	-	0	57.8	Pathogenic
M. leprae TN	59.0	4	-	0	57.8	Pathogenic
M. marinum M	68.6	106	66.1	20	65.7	Pathogenic
***M. smegmatis* str. MC^2 155**	70.0	2	68.7	7	67.4	Non-Pathogenic
M. sp. Spyr1	69.1	2	70.7	3	67.0	Non-Pathogenic
M. sp. JLS	71.2	2	72.4	2	68.4	Non-Pathogenic
M. sp. KMS	71.2	4	72.4	2	68.2	Non-Pathogenic
M. sp. MCS	71.2	3	72.4	2	68.4	Non-Pathogenic
M. tuberculosis CDC1551	67.4	59	62.8	11	65.6	Pathogenic
M. tuberculosis F11	67.5	65	65.3	11	65.6	Pathogenic
M. tuberculosis H37Ra	67.5	95	65.3	15	65.6	Pathogenic
M. tuberculosis H37Rv	67.5	73	65.3	10	65.6	Pathogenic
M. tuberculosis KZN 1435	67.4	64	65.3	10	65.6	Pathogenic
M. ulcerans Agy99	64.8	44	65.8	10	65.4	Pathogenic
M. vanbaalenii PYR-1	68.2	2	69.0	2	67.8	Non-Pathogenic
Mycobacterial-Related Species						
Rhodococcus equi 103S	76.4	1	71.2	1	68.0	Pathogenic
Rhodococcus equi ATCC 33707	-	2	-	0	68.0	Pathogenic
Rhodococcus erythropolis SK121	-	1	-	1	62.0	Pathogenic
Rhodococcus jostii RHA1	71.7	1	-	0	67.0	Pathogenic
Rhodococcus opacus B4	73.7	3	70.7	3	67.0	Non-Pathogenic
Segniliparus rotundus DSM 44985	69.5	2	67.6	2	66.0	Pathogenic
Segniliparus rugosus ATCC BAA-974	-	4	-	3	68.0	Pathogenic
Tsukamurella paurometabola DSM 20162	69.7	1	-	0	68.0	Pathogenic

BOLD: Indicative of recent horizontal gene transfer of PE35 homolog as ascertained by IslandViewer. Genes were found by using the conserved domains (residues listed in text) of the PE35 and PPE68 genes from MTB CDC1551, which were compared to the respective genome sequences, and included with a minimum e-value of 0.01. These numbers are not the same as the current genome annotations. GC content is from the closest homolog to PPE68 or PE35 in that respective genome.

genic and several other pathogenic species, such as *Mycobacterium avium*, *Mycobacterium bovis*, *Mycobacterium ulcerans*, and *Mycobacterium abscessus* are more closely related to *Segniliparus rotundus* and *Segniliparus rugosus* as compared to their counterpart mycobacterial species, as reflected in **Figure 1**. On the other hand, the PE35 tree, as shown in **Figure 3**, reveals that the PE35 gene homologs of several pathogenic mycobacterial species are closely related to *Rhodococcus erythropolis*, non-pathogenic, environmental mycobacterial species are clustered together, and evolutionary groups, both pathogenic and non-pathogenicare related to *Rhodococcus opacus*. It is interesting to note, that the pathogenic species *Mycobacterium avium* and *Mycobacterium abscessus* are clustered with two non-mycobacterial species, *Segniliparus rotundus* and *Segniliparus rugosus*. These results strongly suggest that both PE35 and PPE68 genes have been acquired among these species by HGT.

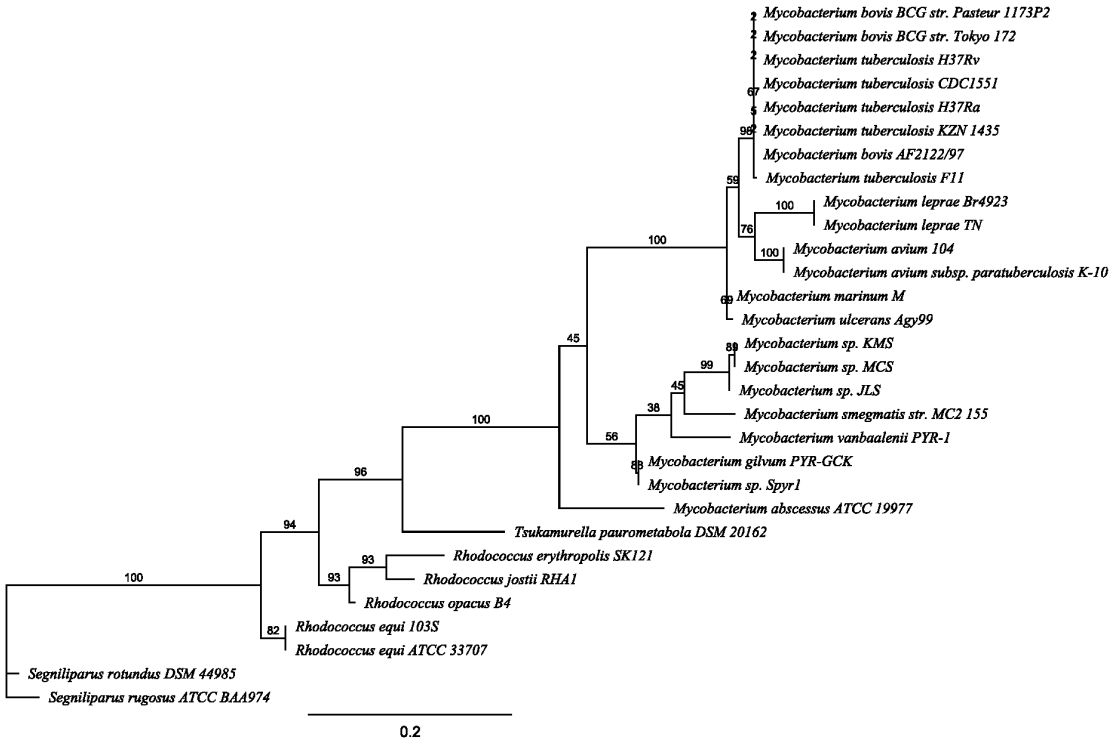


Figure 1. The phylogenetic relationship of 16S rRNA genes among 22 *Mycobacterium* and 8 *Mycobacterium*-related species. Maximum likelihood trees were developed using the Tamura-Nei (TN) model [29]. Scale bar at bottom of tree allows for gauge of numbers of substitutions per site and numbers on the tree branches reflect bootstrap values.

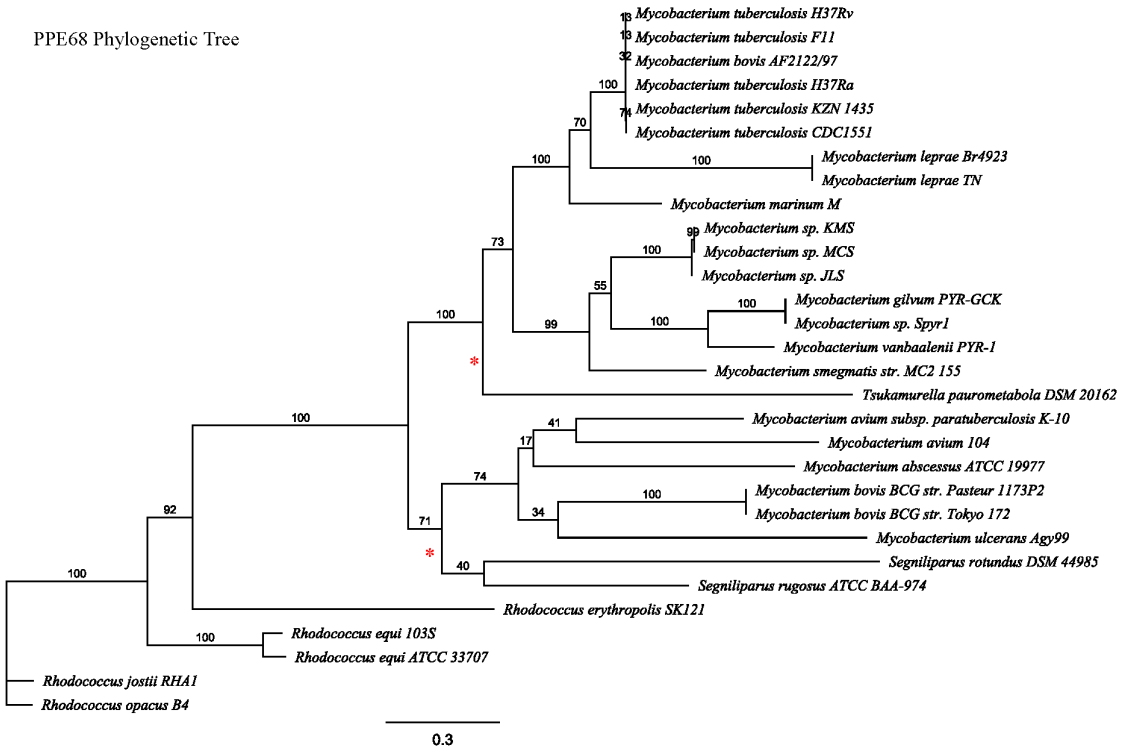


Figure 2. The phylogenetic relationship of PPE68 gene homologs among 22 *Mycobacterium* and 8 *Mycobacterium*-related species. Maximum likelihood trees were developed using the Tamura-Nei (TN) model [29]. Scale bar at bottom of tree allows for gauge of numbers of substitutions per site and numbers on the tree branches reflect bootstrap values.

3.4. Functional Constraints Analysis

For the functional constraints analysis, pairwise comparisons were conducted between each PE35 and PPE68 best homolog of the 22 mycobacterial strains. The relationship between K_a and K_s of PPE68 and PE35 homologs are shown in **Figure 4**. The results revealed that the PE35 and PPE68 genes from non-pathogenic *Mycobacte-*

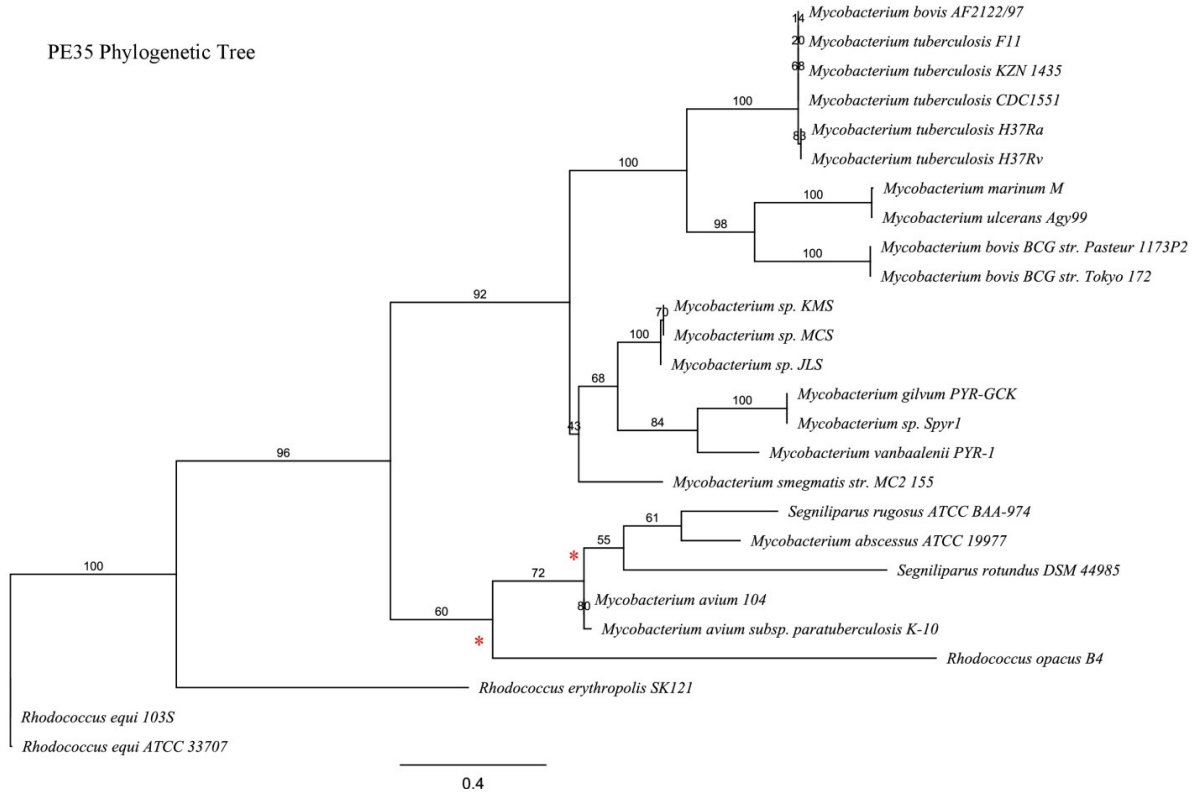


Figure 3. The phylogenetic relationship of PE35 gene homologs among 20 *Mycobacterium* and 6 *Mycobacterium*-related species. Maximum likelihood trees were developed using the Tamura-Nei (TN) model [29]. Scale bar at bottom of tree allows for gauge of numbers of substitutions per site and numbers on the tree branches reflect bootstrap values.

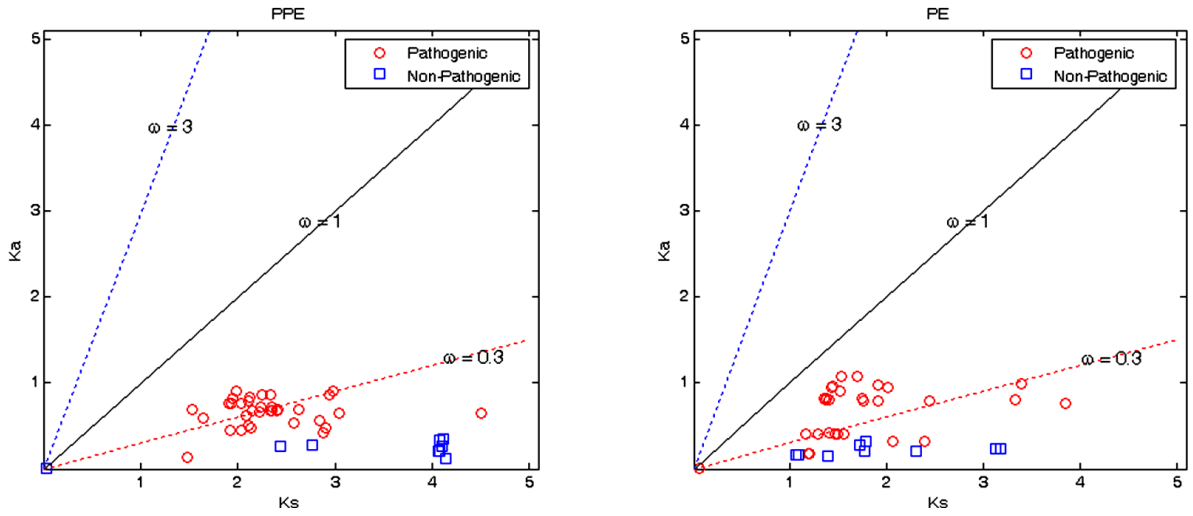


Figure 4. K_a-K_s correlations of PPE68 (22 *Mycobacterium* and 8 *Mycobacterium*-related species) and PE35 gene homologs (20 *Mycobacterium* and 6 *Mycobacterium*-related species). K_a and K_s values were estimated using MYN (Modified Yang-Nielsen algorithm). $\omega = 0.3$, 1 and 3 were used for negative, neutral, and positive selection, respectively.

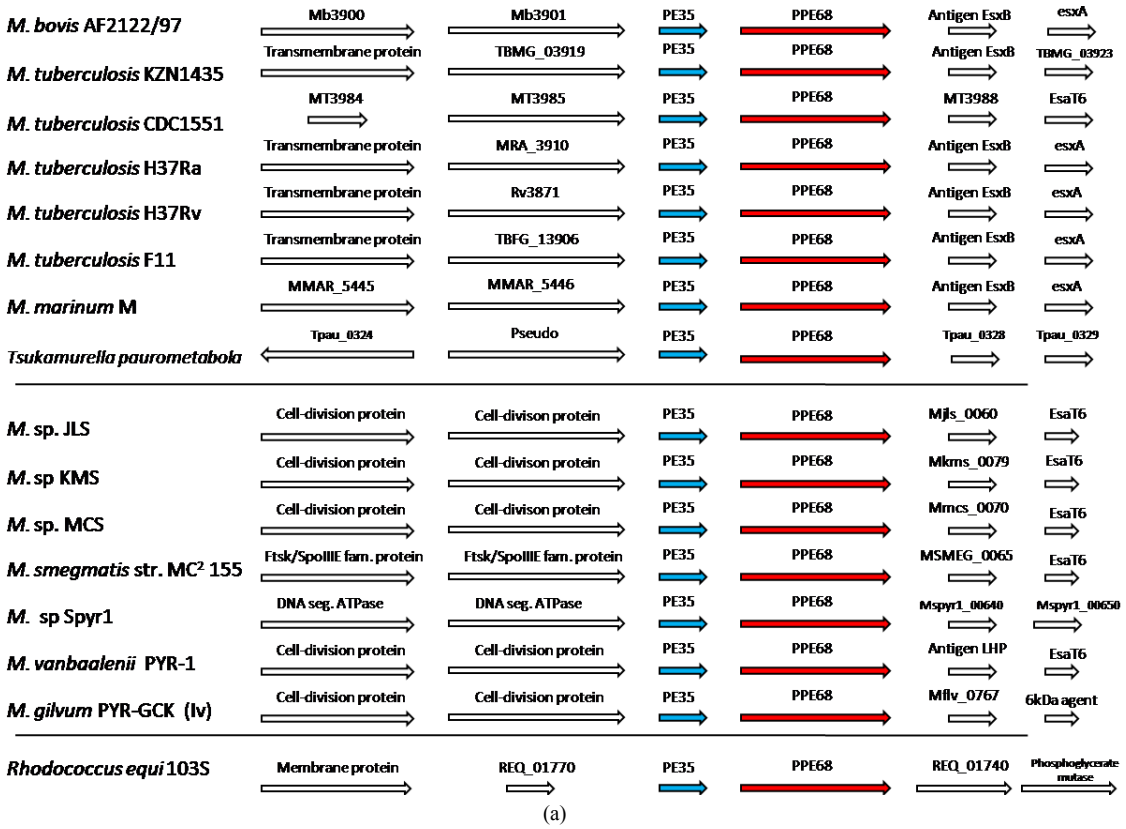
rium strains are under strong negative selection ($\omega < 0.3$), while in many of the pathogenic mycobacterial species these two genes have evolved under relaxed or neutral constraints($0.3 < \omega < 1$). The Ka/Ks (mean ± standard deviation) of PPE68 homologs in pathogenic *Mycobacterium* is 0.291 ± 0.092 while in non-pathogenic is 0.078 ± 0.036. The K_a/K_s of PE35 homologs in pathogenic species is 0.359 ± 0.165 while in non-pathogenic is 0.115 ± 0.041. The constraint analyses also revealed that the PPE family overall is experiencing more intense negative selection as compared to the PE family (Welch's *t*-test two-tailed *p*-value =0.0162). Furthermore, there is a greater degree of variation and lesser constraint among pathogenic strains as compared to non-pathogenic strains.

3.5. Gene Mapping Analysis of PE35 and PPE68 Genes among *Mycobacterium* and Related Species

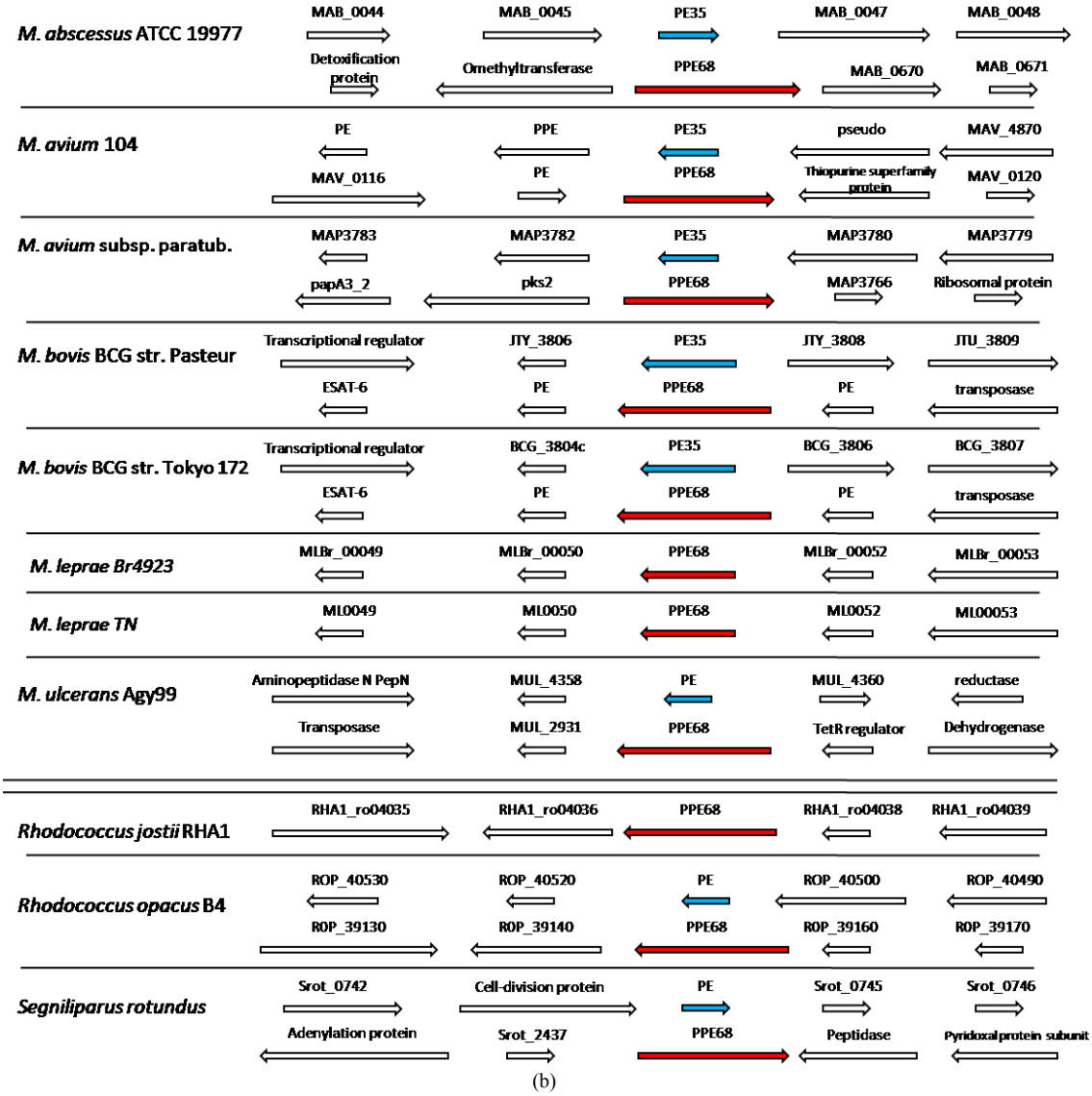
The organization of genes around PE35 and PPE68 homologs for the *Mycobacterium* and their related species are shown in **Figure 5(a)** and **Figure 5(b)**. As revealed in the upper section of **Figure 5(a)**, the PE35/PPE68 genecluster homologs in the pathogenic MTB complex members (*M. tuberculosis*, *M. bovis*, and *M. marinum*) and *Tsukamurella paurometabola* are all flanked by transmembrane protein genes located upstream from PE35 and EsxB and EsaT-6 (esxA) protein genes located directly downstream from PPE68. The only exceptions are the BCG stains, which have extensive genomic rearrangements due to growth in laboratory culture.

Furthermore, among these pathogenic mycobacterial genomes only one of the three *M. bovis* strains, *M. bovis* AF2122/97 is present while, from **Figure 5(b)**, the other two, *M. bovis* BCG str. Tokyo and *M. bovis* BCG str. Pasteur, show significantly different organization, containing genes that encode transposase enzymes involved in the movement of a DNA fragment from one site in the genome to another, and PE35 and PPE68 homologs not located adjacent to each other. It is unclear at this time if this genomic rearrangement is related to the attenuation of these strains.

In the second section of **Figure 5(a)**, a majority of the nonpathogenic *Mycobacterium* also have PPE68/PE35 gene homologs flanked by identical sets of genes. More specifically, downstream from PE35 there are two cell-division proteins while upstream from PPE68 they tend to have antigen and EsaT6 proteins.



(a)



(b)

Figure 5. (a) Gene maps of regions surrounding PPE68/PE35 homologs. In this figure PPE68 and PE35 are located adjacent to each other. Organisms were grouped together based on similarities seen based on flanking regions. Two genes downstream from PE35 and two genes upstream from PPE68 are shown. (b) Gene maps of regions surrounding PPE68/PE35 homologs. PPE68 and PE35 are not located adjacent to each other and have been illustrated in their respective regions. Two genes upstream and downstream from both PE35 and PPE68 are shown.

In the third section of **Figure 5(a)**, *Rhodococcusequi* 103S is shown to have significant similarity to the pathogenic mycobacterium in the first section with the notable difference of a *Phosphoglycerate mutase* located upstream from PPE68 instead of EsaT6.

Figure 5(b) contains all the *Mycobacterium* and their related species from which gene clusterpatterns could not be distinguished. Also all of the species in **Figure 5(b)** with the exception of *M. leprae* Br4923 and *M. leprae* TN—for which no PE35 homologs were found—have PE35 and PPE68 homologs that are not located directly adjacent to each other.

4. Discussion

4.1 Genomic Characteristics

The data revealed that the pathogenic strains tend to have smaller genome sizes, with the exception of *M. mari-*

num and *M. ulcerans* Agy99. Smaller genome size has been attributed to a narrow host range of the pathogen [32]. However, the above two strains are outliers which can be most likely ascribed to more recent acquisition of virulence and thus insufficient time to delete those genes which are necessary only to the free-living lifestyle [32] [33]. It should be noted that two pathogenic strains in particular, *Mycobacterium leprae* Br4923 and *Mycobacterium leprae* TN, have the smallest genome sizes among all pathogenic species. Moreover, *M. leprae* utilizes the smallest gene set among these organisms, as its genome codes for only ~1600 proteins (49% of genome), which reveals an extreme case of reductive evolution and massive degeneration resulting from the obligate intracellular pathogenic lifestyle. Despite causing a chronic infection in humans, *M. leprae* has lost almost all of the PE/PPE genes, thus these have little role in pathogenesis of that organism.

4.2. PE35 and PPE68 Gene Homologs in Mycobacteria and Their Related Species

Organisms within the genera *Rhodococcus*, *Segniliparus*, and *Tsukamurella* were found to contain genes with significant homology to several PE/PPE genes found in mycobacteria. It is interesting to note that PPE68 gene homologs are more diverse and found in all these species while *M. leprae* lacks any significant PE35 gene homolog, consistent with previous findings [8]. It should be noted that five PE genes are annotated in the *M. leprae* genome, based on homology of partial blocks of sequence.

The results showed that pathogenic mycobacteria had higher copy numbers of PPE genes as compared to non-pathogenic, again, with the exclusion of *M. leprae*. This may be indicative of the *leprae* species splitting off before amplification of these genes occurred in the sister lineage or gene decay in the existing genome. Only five PPE genes are annotated in the *M. leprae* genome. Furthermore, the low PPE copy number (<11) for all the non-pathogenic mycobacteria is indicative that these genes are specifically needed for virulence and host infectivity and may not be needed in the environment, consistent with previous literature [8]. Aquatic environmental mycobacteria may have pathogenic phases in their life cycles, interacting with protozoa [34].

The low copy numbers of PE, as compared to PPE, indicate that PPE is more diverged than PE. The related species also have PE/PPE genes but those that are pathogenic have low copy numbers for both PE and PPE gene families. This may be indicative of three possible scenarios: 1) PE/PPE genes have less important roles in virulence and host-infectivity, 2) it is possible that this HGT is more recent and thus these related species have not been given sufficient time to expand, or 3) these related pathogenic species are acute infections that do not need to evade the host and thus have little need for the PPE and PE antigen variation. Yet *Rhodococcus* can cause persistent pneumonia and other infections in horses. Upon examination of the respective trees it is evident that since these species have not diverged more recently, the most likely suggestion is that PPE and PE gene families are not prominent in virulence for these species. This is a fertile area for experimental investigation.

4.3. Horizontal Transfer of PE35 and PPE68 Genes among *Mycobacterium* and Related Species

Only PE35 gene homologs in *M. smegmatis* and *M. avium* subsp. *Paratuberculosis* were found in predicted horizontal transfer regions. However, this is likely because the IslandViewer program only takes recent HGTs into account and cannot account for ancestral events. This is important to note as older HGTs would likely be homogenized with the surrounding genome and would be substantially more difficult if not impossible to find using conventional bioinformatics techniques. Consistently, GC content in the PE/PPE genes and surrounding genomes is similar in most cases.

As compared to the 16S rRNA tree, both the PPE and PE phylogenetic trees show significant differences. This is possible due to a variety of scenarios. It can be indicative of HGT taking place in the form of three different scenarios: from *Mycobacterium* to *Mycobacterium*, *Mycobacterium* to its related-species and from related-species to *Mycobacterium*. On the other hand or in conjunction, different rates of divergence of the genes could result in alternative trees. With inductive reasoning it is possible to posit transfer events based on the ordering of phylogenies and the placement of tree nodes. More specifically, the rearrangement of PE/PPE nodes on a tree can be suggestive of gene transfer events in comparison to tree structures expected from 16S comparisons. In the PPE68 tree, the clades containing *M. gilvum*, *M.* sp. Spyr1, *M. vanbaalenii* and *M. smegmatis* show slight reordering as compared to the 16S rRNA reference tree. However, although ordering may be slightly different, they are still clustered together very well, the cluster is within the expected tree segment, and they have not diverged very significantly so it is unlikely that any ascertainable transfer events occurred, as this is more

easily explainable by simple genetic drift. However, there are three notable HGTs that may have taken place using the aforementioned methodology: two between a *Mycobacterium* and a *Mycobacterium*-related species (depicted by asterisks in **Figure 2**). *Tsukamurella paurometabola* DSM 20162 moved up the PPE tree from its original 16S rRNA placement, indicative of a transfer by a common ancestor of several pathogenic mycobacterial species to this organism. Similarly, the location of *Segniliparus* is incongruent with its expected tree placement. As such, it is suggestible that it experienced an HGT event from a *Mycobacterium* ancestor in the nearby clade and obtained its PPE68 homolog as such. Furthermore, although the differences within these trees signify that significant HGT of PPE68 and PE35 has taken place, since the mechanisms and host/environment interactions of the eight *Mycobacterium*-related species are not fully understood it is not possible to identify what role exactly the PPE and PE genes hold within these organisms. For non-pathogens, these cell envelope-located proteins may play roles in cell attachment to other cells and surfaces.

For the PE35 trees, *M. leprae* was not found to have any significant PE35 homolog and hence is missing from the tree. Otherwise, there are three notable differences within the PE35 tree compared to the 16S rRNA tree. First, similar to the PPE68 tree, the clades containing *M. gilvum*, *M.* sp. Spyr1, *M. vanbaalenii* and *M. smegmatis* show slight reordering as compared to the 16S rRNA tree. Once again, however, although ordering may be slightly different, they are still clustered together very well and are not diverged very significantly so it is unlikely that these represent transfer events either. The other two differences are more notable and may represent HGTs that have taken place, specifically between *Mycobacterium* and *Mycobacterium*-related species (depicted by asterisks in **Figure 3**). The first involves *Segniliparus* obtaining a PE35 homolog. It is unlikely that it acquired its PE35 homolog from convergent evolution because of the close relationship established in the PPE tree. As such an HGT event, likely involving the PE/PPE region, occurred whereby either the PE35 homolog from *M. avium* or a precursor of *M. abscessus* was transferred to it. The second event may be similar in which *Rhodococcusopacus*B4 or its precursor potentially obtained its PE35 gene homolog from a *Mycobacterium* ancestor in the nearby clade. However, its gene location is not significantly incongruent and may be the result of convergent evolution of retained homology from a common ancestor to the *Mycobacterium* in the associated clade.

To note, within both the PE35 and PPE68 trees, *M. avium* is far removed from its location within the 16S rRNA tree and is in a place that it should not be if divergence patterns and rates were fairly consistent across the *Mycobacterium* organisms. As such, this may be indicative of a HGT event of PPE from another *Mycobacterium* into the *M. avium* group. Alternatively, since the *M. avium* complex members are opportunistic pathogens with a dominantly environmental lifestyle, mutations may have accumulated in these genes in the absence of host selection.

On the other hand, the PE35 and PPE68 trees fairly clearly suggest that the genes originated from a common ancestor of the *Mycobacterium* and *Mycobacterium*-related species or that an ancient gene transfer event occurred among the ancestors of these two groups of organisms. Subsequently, the genes diverged and specialized their functions in the *Mycobacterium* species. For instance, the PE and PPE homolog genes are not very diverged in the *Rhodococcus* organisms in comparison to the *Mycobacterium* and of the five *Rhodococcus* that contain a PPE homolog, one does not contain a PE homolog suggesting that it may have been deleted. To reiterate then, these genes may not play as integral a role in pathogenicity in these organisms. Conversely, the rapid divergence of the PE and PPE genes in *Mycobacterium* may lend credence to its success as a chronic pathogen.

4.4. Gene Mapping Analysis

As revealed in **Figure 5(a)**, PPE68 gene homologs are flanked by identical sets of genes in the majority of pathogenic mycobacterial genomes and *Tsukamurella paurometabola*. However, *M. bovis* BCG str. Tokyo and *M. bovis* BCG str. Pasteur show significantly different organization within that region, which also contains genes that encode transposases. This is not surprising as the BCG strains have been cultured *in vitro* throughout the years and have large deleted and rearranged regions in their genomes [35].

Another point to note is that *Tsukamurella paurometabola* DSM 21062 has a similar upstream and downstream pattern to the other pathogenic *Mycobacterium* species in that section. This gives credence to the HGT hypothesis established earlier, in which it was stated that *Tsukamurella* obtained its PPE68 gene homolog from several pathogenic *Mycobacterium* species. However, given the hypothetical nature of the *Tsukamurella* proteins in **Figure 5(a)**, further analysis was conducted. More specifically, the two proteins downstream to PE35 and the two proteins upstream to PPE68 in *Tsukamurella* were compared via MUSCLE pairwise alignment [23] to their corresponding proteins in *M. tuberculosis F*11. The two proteins downstream to PE35 in *Tsukamurella*,

Tpau_0324 and Tpau_0325 (labeled "Pseudo") have sequence lengths of 1,551 bp and 3,546bp respectively. Their corresponding proteins in *M. tuberculosis F*11, TBFG_13905 (labeled "Transmembrane protein") and TBFG_13906, have sequence lengths of 2244 and 1776. Upon comparison it was found that the comparisons of Tpau_0324 to TBFG_13906 and Tpau_0325 to TBFG_13905 showed greater pairwise identities (45%, 37.5%) versus their comparisons to their corresponding genes, specifically Tpau_0324 to TBFG_13905 and Tpau_0325 to TBFG_13906 (39.5%, 31.3%). Also, taking into account that the sequence lengths are more similar across these "crossed" pairs (1551 and 1776 & 2244 and 3546) it may be inferred that these two genes in *Tsukamurella* were rearranged. For the two genes upstream to PPE68 in *Tsukamurella* (Tpau_0328 and Tpau_0329), the same information was gathered. The pairwise identities are as follows: Tpau_0328 to TBFG_13909 48%, Tpau_ 0328 to TBFG_13910 44.3%, Tpau_0329 to TBFG_13909 47.9% and Tpau_0329 to TBFG_13910 43%. In this case, there were no significant differences in pairwise identity between the sets and no significant differences in sequence lengths (all four have lengths of ~300 bp) and hence it may be suggested that these genes (Tpau_0328 and Tpau_0329 in *Tsukamurella paurometabola* DSM 20162) did not rearrange but rather stayed in a similar gene pattern. Furthermore, these may represent duplicate gene pairs.

Lastly, in the upper section of **Figure 5(a)**, it is notable that MTB complex members show very similar patterns. This is consistent with low genetic variation seen in *Mycobacterium tuberculosis* [36]. More specifically, given that PE35 and PPE68 are involved in pathogenicity, it is not surprising that the surrounding regions in the obligate pathogen *Mycobacterium tuberculosis* are highly conserved, as they are likely integral for continued organism success. Furthermore, it has been proposed that this low level of genetic variation suggests that the entire population resulted from clonal expansion following an evolutionary bottleneck around 35,000 years ago.

4.5. Diversification of PE and PPE Protein Families in Pathogenic Mycobacteria Is Due to Less Evolutionary Constraint

PE35 and PPE68 genes of non-pathogenic and non-mycobacterial species are under strong negative selection ($\omega < 0.3$) while pathogenic mycobacterial species are under relaxed or neutral constraint ($0.3 < \omega < 1$). Also, there is a greater degree of variation and lesser constraint among pathogenic strains as compared to non-pathogenic strain. These results suggest, as expected, that the antigenic variation function of PE/PPE genes and pressure from the host immune system has resulted in amplification and divergence of these genes in MTB and other pathogens. A significant question that remains is the function of PE/PPE genes outside the host, if any.

5. Conclusions

In this study, we have showed that significant homologs to the ancestral PE/PPE genes exist outside the mycobacterial lineage. *Mycobacterium* and their related species have acquired PE and PPE genes through horizontal gene transfers from each other.

The study reveal that PE and PPE gene homologs are not only limited to *Mycobacterium,* but also exist in at least three other non-mycobacterial genera, *Rhodococcus*, *Tsukamurella* and *Segniliparus*. All these genera are in the suborder Corynebacterineae, a group of the Actinomycetes. Results also demonstrate that PE and PPE genes are more diverse and are maintained at low evolutionary constraint in pathogenic *Mycobacterium* as compared with non-pathogenic *Mycobacterium* and other non-mycobacterial species. These findings possibly shed light on the diverse functions and pathogenicity of the PE/PPE proteins among these organisms.

References

[1] Banu, S., Honore, N., Saint-Joanis, B., Philpott, D., Prevost, M.C. and Cole, S.T. (2002) Are the PE-PGRS Proteins of Mycobacterium Tuberculosis Variable Surface Antigens? *Molecular Microbiology*, **44**, 9-19. http://dx.doi.org/10.1046/j.1365-2958.2002.02813.x

[2] Brennan, P.J. and Vissa, V.D. (2001) Genomic Evidence for the Retention of the Essential Mycobacterial Cell Wall in the Otherwise Defective Mycobacterium Leprae. *Leprosy Review*, **72**, 415-428.

[3] Lambrecht, R.S., Carriere, J.F. and Collins, M.T. (1988) A Model for Analyzing Growth Kinetics of a Slowly Growing Mycobacterium sp. *Applied and Environmental Microbiology*, **54**, 910-916.

[4] Tsukamura, M. (1966) Adansonian Classification of Mycobacteria. *Journal of General Microbiology*, **45**, 253-273. http://dx.doi.org/10.1099/00221287-45-2-253

[5] Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry

3rd, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S. and Barrell, B.G. (1998) Deciphering the Biology of Mycobacterium Tuberculosis from the Complete Genome Sequence. *Nature*, **393**, 537-544. http://dx.doi.org/10.1038/31159

[6] Camus, J.C., Pryor, M.J., Medigue, C. and Cole, S.T. (2002) Re-Annotation of the Genome Sequence of Mycobacterium Tuberculosis H37Rv. *Microbiology*, **148**, 2967-2973.

[7] Chaitra, M.G., Hariharaputran, S., Chandra, N.R., Shaila, M.S. and Nayak, R. (2005) Defining Putative T Cell Epitopes from PE and PPE Families of Proteins of Mycobacterium Tuberculosis with Vaccine Potential. *Vaccine*, **23**, 1265-1272. http://dx.doi.org/10.1016/j.vaccine.2004.08.046

[8] Gey van Pittius, N.C., Sampson, S.L., Lee, H., Kim, Y., van Helden, P.D. and Warren, R.M. (2006) 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 Evolutionary Biology*, **6**, 95. http://dx.doi.org/10.1186/1471-2148-6-95

[9] Gordon, S.V., Eiglmeier, K., Brosch, R., Garnier, T., Honoré, N., Barrell, B.G. and Cole, S.T. (2009) Chapter 5. Genomics of Mycobacterium Tuberculosis and Mycobacterium Leprae. In: Ratledge, C. and Dale, J., Eds., *Mycobacteria*: *Molecular Biology and Virulence*, Wiley-Blackwell, 93-109. http://onlinelibrary.wiley.com/book/10.1002/9781444311433

[10] Adindla, S. and Guruprasad, L. (2003) Sequence Analysis Corresponding to the PPE and PE Proteins in Mycobacterium Tuberculosis and Other Genomes. *Journal of Biosciences*, **28**, 169-179. http://dx.doi.org/10.1007/BF02706216

[11] Brennan, M.J., Delogu, G., Chen, Y., Bardarov, S., Kriakov, J., Alavi, M. and Jacobs Jr., W.R. (2001) Evidence That Mycobacterial PE_PGRS Proteins Are Cell Surface Constituents That Influence Interactions with Other Cells. *Infection and Immunity*, **69**, 7326-7333. http://dx.doi.org/10.1128/IAI.69.12.7326-7333.2001

[12] Espitia, C., Laclette, J.P., Mondragon-Palomino, M., Amador, A., Campuzano, J., Martens, A., Singh, M., Cicero, R., Zhang, Y. and Moreno, C. (1999) The PE-PGRS Glycine-Rich Proteins of *Mycobacterium tuberculosis*: A New Family of Fibronectin-Binding Proteins? *Microbiology*, **145**, 3487-3495.

[13] Delogu, G., Pusceddu, C., Bua, A., Fadda, G., Brennan, M.J. and Zanetti, S. (2004) Rv1818c-Encoded PE_PGRS Protein of *Mycobacterium tuberculosis* Is Surface Exposed and Influences Bacterial Cell Structure. *Molecular Microbiology*, **52**, 725-733. http://dx.doi.org/10.1111/j.1365-2958.2004.04007.x

[14] Delogu, G. and Brennan, M.J. (2001) Comparative Immune Response to PE and PE_PGRS Antigens of *Mycobacterium tuberculosis*. *Infection and Immunity*, **69**, 5606-5611. http://dx.doi.org/10.1128/IAI.69.9.5606-5611.2001

[15] Singh, P.P., Parra, M., Cadieux, N. and Brennan, M.J. (2008) A Comparative Study of Host Response to Three *Mycobacterium tuberculosis* PE_PGRS Proteins. *Microbiology*, **154**, 3469-3479. http://dx.doi.org/10.1099/mic.0.2008/019968-0

[16] Delogu, G., Sanguinetti, M., Pusceddu, C., Bua, A., Brennan, M.J., Zanetti, S. and Fadda, G. (2006) PE_PGRS Proteins Are Differentially Expressed by *Mycobacterium tuberculosis* in Host Tissues. *Microbes and Infection*, **8**, 2061-2067. http://dx.doi.org/10.1016/j.micinf.2006.03.015

[17] Dheenadhayalan, V., Delogu, G., Sanguinetti, M., Fadda, G. and Brennan, M.J. (2006) Variable Expression Patterns of *Mycobacterium tuberculosis* PE_PGRS Genes: Evidence That PE_PGRS16 and PE_PGRS26 Are Inversely Regulated *in Vivo*. *Journal of Bacteriology*, **188**, 3721-3725. http://dx.doi.org/10.1128/JB.188.10.3721-3725.2006

[18] Li, Y., Miltner, E., Wu, M., Petrofsky, M. and Bermudez, L.E. (2005) A *Mycobacterium avium* PPE Gene Is Associated with the Ability of the Bacterium to Grow in Macrophages and Virulence in Mice. *Cellular Microbiology*, **7**, 539-548. http://dx.doi.org/10.1111/j.1462-5822.2004.00484.x

[19] Voskuil, M.I., Schnappinger, D., Rutherford, R., Liu, Y. and Schoolnik, G.K. (2004) Regulation of the *Mycobacterium tuberculosis* PE/PPE Genes. *Tuberculosis*, **84**, 256-262. http://dx.doi.org/10.1016/j.tube.2003.12.014

[20] Brodin, P., Rosenkrands, I., Andersen, P., Cole, S.T. and Brosch, R. (2004) ESAT-6 Proteins: Protective Antigens and Virulence Factors? *Trends in Microbiology*, **12**, 500-508. http://dx.doi.org/10.1016/j.tim.2004.09.007

[21] Marri, P.R., Bannantine, J.P. and Golding, G.B. (2006) Comparative Genomics of Metabolic Pathways in Mycobacterium Species: Gene Duplication, Gene Decay and Lateral Gene Transfer. *FEMS Microbiology Reviews*, **30**, 906-925. http://dx.doi.org/10.1111/j.1574-6976.2006.00041.x

[22] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, **25**, 3389-3402. http://dx.doi.org/10.1093/nar/25.17.3389

[23] Edgar, R.C. (2004) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research*, **32**, 1792-1797. http://dx.doi.org/10.1093/nar/gkh340

[24] Yang, Z. and Nielsen, R. (2000) Estimating Synonymous and Nonsynonymous Substitution Rates under Realistic Evolutionary Models. *Molecular Biology and Evolution*, **17**, 32-43. http://dx.doi.org/10.1093/oxfordjournals.molbev.a026236

[25] Zhang, Z., Li, J. and Yu, J. (2006) Computing K_a and K_s with a Consideration of Unequal Transitional Substitutions. *BMC Evolutionary Biology*, **6**, 44. http://dx.doi.org/10.1186/1471-2148-6-44

[26] Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K. and Yu, J. (2006) K_aK_s_Calculator: Calculating K_a and K_s through Model Selection and Model Averaging. *Genomics*, *Proteomics & Bioinformatics*, **4**, 259-263. http://dx.doi.org/10.1016/S1672-0229(07)60007-2

[27] Drummond, A.J., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Thierer, T. and Wilson, A. (2009) Geneious v4.6.

[28] Guindon, S. and Gascuel, O. (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, **52**, 696-704. http://dx.doi.org/10.1080/10635150390235520

[29] Tamura, K. and Nei, M. (1993) Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees. *Molecular Biology and Evolution*, **10**, 512-526.

[30] Bateman, A., Coil, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S. and Sonnhammer, E.L.L. (2004) The Pfam Protein Families Database. *Nucleic Acids Research*, **32**, D138-D141. http://dx.doi.org/10.1093/nar/gkh121

[31] Langille, M.G. and Brinkman, F.S. (2009) Island Viewer: An Integrated Interface for Computational Identification and Visualization of Genomic Islands. *Bioinformatics*, **25**, 664-665. http://dx.doi.org/10.1093/bioinformatics/btp030

[32] Stinear, T.P., Seemann, T., Harrison, P.F., Jenkin, G.A., Davies, J.K., Johnson, P.D., Abdellah, Z., Arrowsmith, C., Chillingworth, T., Churcher, C., Clarke, K., Cronin, A., Davis, P., Goodhead, I., Holroyd, N., Jagels, K., Lord, A., Moule, S., Mungall, K., Norbertczak, H., Quail, M.A., Rabbinowitsch, E., Walker, D., White, B., Whitehead, S., Small, P.L., Brosch, R., Ramakrishnan, L., Fischbach, M.A., Parkhill, J. and Cole, S.T. (2008) Insights from the Complete Genome Sequence of *Mycobacterium marinum* on the Evolution of *Mycobacterium tuberculosis*. *Genome Research*, **18**, 729-741. http://dx.doi.org/10.1101/gr.075069.107

[33] Stinear, T.P., Seemann, T., Pidot, S., Frigui, W., Reysset, G., Garnier, T., Meurice, G., Simon, D., Bouchier, C., Ma, L., Tichit, M., Porter, J.L., Ryan, J., Johnson, P.D., Davies, J.K., Jenkin, G.A., Small, P.L., Jones, L.M., Tekaia, F., Laval, F., Daffe, M., Parkhill, J. and Cole, S.T. (2007) Reductive Evolution and Niche Adaptation Inferred from the Genome of *Mycobacterium ulcerans*, the Causative Agent of Buruli Ulcer. *Genome Research*, **17**, 192-200. http://dx.doi.org/10.1101/gr.5942807

[34] Primm, T.P., Lucero, C.A. and Falkinham 3rd, J.O. (2004) Health Impacts of Environmental Mycobacteria. *Clinical Microbiology Reviews*, **17**, 98-106. http://dx.doi.org/10.1128/CMR.17.1.98-106.2004

[35] Mahairas, G.G., Sabo, P.J., Hickey, M.J., Singh, D.C. and Stover, C.K. (1996) Molecular Analysis of Genetic Differences between *Mycobacterium bovis* BCG and Virulent *M. bovis*. *Journal of Bacteriology*, **178**, 1274-1282.

[36] Gutierrez, M.C., Brisse, S., Brosch, R., Fabre, M., Omais, B., Marmiesse, M., Supply, P. and Vincent, V. (2005) Ancient Origin and Gene Mosaicism of the Progenitor of *Mycobacterium tuberculosis*. *PLoS Pathogens*, **1**, e5. http://dx.doi.org/10.1371/journal.ppat.0010005

Supplementary

Table S1. Pairwise protein identities between PPE68 protein homologs of *Mycobacterium* and their related species.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	100																													
2	65	100																												
3	38	71	100																											
4	46	50	48	100																										
5	57	65	70	42	100																									
6	57	65	70	42	100	100																								
7	38	43	41	41	43	43	100																							
8	36	35	36	45	37	37	40	100																						
9	36	35	36	45	37	37	40	100	100																					
10	46	46	45	72	50	50	41	45	45	100																				
11	46	49	44	62	44	44	53	29	29	48	100																			
12	44	49	45	45	44	44	57	31	31	65	63	100																		
13	44	49	45	44	44	44	57	30	30	47	62	99	100																	
14	44	49	45	44	44	44	57	30	30	47	62	99	100	100																
15	38	43	41	41	43	43	100	40	40	43	50	56	55	55	100															
16	45	50	47	100	43	43	40	44	44	70	63	63	63	63	40	100														
17	46	50	48	100	43	43	40	44	44	70	63	63	63	63	40	100	100													
18	46	50	48	100	43	43	40	44	44	70	63	63	63	63	40	100	100	100												
19	46	50	48	100	43	43	40	44	44	70	63	63	63	63	40	100	100	100	100											
20	45	50	47	100	43	43	40	44	44	70	63	63	63	63	40	100	100	100	100	100										
21	48	55	55	42	52	52	38	32	32	43	38	42	42	42	38	42	42	42	42	42	100									
22	44	46	44	41	41	41	78	30	30	43	50	59	59	59	78	41	41	41	41	41	40	100								
23	32	38	35	32	34	34	32	25	25	28	30	30	30	30	32	32	32	32	32	32	27	29	100							
24	32	38	35	32	34	34	28	24	24	30	30	30	30	30	28	32	32	32	32	32	27	30	99	100						
25	29	31	32	33	29	29	28	36	36	26	31	33	33	33	28	33	33	33	33	33	28	30	43	47	100					
26	29	34	32	29	30	30	32	38	38	29	29	33	33	33	32	29	29	29	29	29	27	30	50	57	53	100				
27	30	35	32	29	30	30	32	38	38	29	30	34	34	34	32	29	29	29	29	29	27	32	51	58	54	93	100			
28	34	58	58	46	43	43	40	42	42	44	45	45	45	45	40	46	46	46	46	46	42	43	35	35	32	33	33	100		
29	61	62	66	46	58	58	36	29	29	47	48	47	47	47	36	46	46	46	46	46	49	42	38	30	34	34	35	40	100	
30	41	43	41	40	35	35	38	35	35	40	44	33	33	33	38	40	40	40	40	40	33	42	27	27	23	27	25	43	42	100

Species ID: *Mycobacterium abscessus* ATCC 19977 (1), *Mycobacterium avium* 104 (2), *Mycobacterium avium* subsp. *Paratuberculosis* K-10 (3), *Mycobacterium bovis* AF2122/97 (4), *Mycobacterium bovis* BCG str. Pasteur 1173P2 (5), *Mycobacterium bovis* BCG str. Tokyo 172 (6), *Mycobacterium gilvum* PYR-GCK (7), *Mycobacterium leprae* Br4923 (8), *Mycobacterium leprae* TN (9), *Mycobacterium marinum* M (10), *Mycobacterium smegmatis* str. MC^2 155 (11), *Mycobacterium* sp. JLS (12), *Mycobacterium* sp. KMS (13), *Mycobacterium* sp. MCS (14), *Mycobacterium* sp. Spyr1 (15), *Mycobacterium tuberculosis* CDC1551 (16), *Mycobacterium tuberculosis* F11 (17), *Mycobacterium tuberculosis* H37Ra (18), *Mycobacterium tuberculosis* H37Rv (19), *Mycobacterium tuberculosis* KZN 1435 (20), *Mycobacterium ulcerans* Agy99 (21), *Mycobacterium vanbaalenii* PYR-1 (22), *Rhodococcusequi* 103S (23), *Rhodococcusequi* ATCC 33707 (24), *Rhodococcus erythropolis* SK121 (25), *Rhodococcus jostii* RHA1 (26), *Rhodococcus opacus* B4 (27), *Segniliparus rotundus* DSM 44985 (28), *Segniliparus rugosus* ATCC BAA-974 (29) and *Tsukamurella paurometabola* DSM 20162 (30).

Table S2. Pairwise protein identities between PE35 protein homologs of *Mycobacterium* and their related species.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	100																									
2	78.0	100																								
3	77.0	99.0	100																							
4	36.7	36.7	36.7	100																						
5	31.8	26.7	26.7	50.5	100																					
6	31.8	26.7	26.7	50.5	100	100																				
7	33.9	ND	ND	45.6	35.7	35.7	100																			
8	33.3	35.0	35.0	55.6	59.6	59.6	38.8	100																		
9	41.1	43.3	41.7	52.2	42.3	42.3	55.7	45.4	100																	
10	38.1	43.3	41.7	46.4	45.6	45.6	63.9	44.3	74.2	100																
11	38.1	43.3	41.7	46.4	45.6	45.6	63.9	44.3	74.2	100	100															
12	38.1	43.3	41.7	46.4	45.6	45.6	63.9	44.3	74.2	100	100	100														
13	33.9	ND	ND	45.6	35.7	35.7	100	38.8	55.7	63.9	63.9	63.9	100													
14	36.7	36.7	36.7	100	50.5	50.5	45.6	55.6	52.2	46.4	46.4	46.4	45.6	100												
15	36.7	36.7	36.7	100	50.5	50.5	45.6	55.6	52.2	46.4	46.4	46.4	45.6	100	100											
16	36.7	36.7	36.7	100	50.5	50.5	45.6	55.6	52.2	46.4	46.4	46.4	45.6	100	100	100										
17	36.7	36.7	36.7	100	50.5	50.5	45.6	55.6	52.2	46.4	46.4	46.4	45.6	100	100	100	100									
18	36.7	36.7	36.7	100	50.5	50.5	45.6	55.6	52.2	46.4	46.4	46.4	45.6	100	100	100	100	100								
19	33.3	35.0	35.0	55.6	59.6	59.6	38.8	100	45.4	44.3	44.3	44.3	38.8	55.6	55.6	55.6	55.6	55.6	100							
20	41.1	36.7	35.0	45.6	37.8	37.8	73.5	38.8	64.9	71.1	71.1	71.1	73.5	45.6	45.6	45.6	45.6	45.6	38.8	100						
21	29.2	30.2	30.2	42.1	37.3	37.3	29.6	32.4	45.7	75.0	75.0	75.0	29.6	42.1	42.1	42.1	42.1	42.1	32.4	34.5	100					
22	29.2	30.2	30.2	42.1	37.3	37.3	29.6	32.4	45.7	75.0	75.0	75.0	29.6	42.1	42.1	42.1	42.1	42.1	32.4	34.5	100	100				
23	27.3	75.0	75.0	30.8	80.0	80.0	29.3	32.5	30.9	32.8	32.8	32.8	29.3	30.8	30.8	30.8	30.8	30.8	32.5	75.0	51.1	51.1	100			
24	70.0	45.5	45.5	ND	66.7	66.7	30.0	71.4	27.6	44.4	44.4	44.4	30.0	ND	ND	ND	ND	ND	71.4	ND	83.3	83.3	ND	100		
25	58.8	62.9	61.9	ND	ND	ND	34.5	36.4	33.3	50.0	50.0	50.0	34.5	ND	ND	ND	ND	ND	36.4	34.6	41.4	41.4	28.6	ND	100	
26	74.5	71.0	70.0	40.0	ND	ND	66.7	39.7	39.3	41.4	41.4	41.4	66.7	40.0	40.0	40.0	40.0	40.0	39.7	33.7	38.2	38.2	29.5	42.1	55.9	100

ND: Not Detected. **Species ID:** *Mycobacterium abscessus* ATCC 19977 (1), *Mycobacterium avium*104 (2), *Mycobacterium avium* subsp. *Paratuberculosis* K-10 (3), *Mycobacterium bovis* AF2122/97 (4), *Mycobacterium bovis* BCG str. Pasteur 1173P2 (5), *Mycobacterium bovis* BCG str. Tokyo 172 (6), *Mycobacterium gilvum* PYR-GCK (7), *Mycobacterium marinum* M (8), *Mycobacterium smegmatis* str. MC^2 155 (9), *Mycobacterium* sp. JLS (10), *Mycobacterium* sp. KMS (11), *Mycobacterium* sp. MCS (12), *Mycobacterium* sp. Spyr1 (13), *Mycobacterium tuberculosis* CDC1551 (14), *Mycobacterium tuberculosis* F11 (15), *Mycobacterium tuberculosis* H37Ra (16), *Mycobacterium tuberculosis* H37Rv (17), *Mycobacterium tuberculosis* KZN 1435 (18), *Mycobacterium ulcerans* Agy99 (19), *Mycobacterium vanbaalenii* PYR-1 (20), *Rhodococcusequi* 103S (21), *Rhodococcus equi* ATCC 33707 (22), *Rhodococcus erythropolis* SK121 (23), *Rhodococcus opacus* B4 (24), *Segniliparus rotundus* DSM 44985 (25) and *Segniliparus rugosus* ATCC BAA-974 (26).

Table S3. Accession numbers for ancestral PE35/PPE68 homologs.

Species	PPE Accession Number	PE Accession Number
Mycobacterial Species		
M. abscessus ATCC 19977	YP_001701421.1	YP_001700800.1
M. avium 104	YP_879414.1	YP_883991.1
M. avium subsp. *Paratuberculosis* K-10	NP_962699.1	NP_962715.1
M. bovis AF2122/97	NP_857540.1	NP_857539.1
M. bovis BCG str. Pasteur 1173P2	YP_979128.1	YP_979884.1
M. bovis BCG str. Tokyo 172	YP_002646085.1	YP_002646845.1
M. gilvum PYR-GCK	YP_001132041.1	YP_001132042.1
M. leprae Br4923	YP_002502795.1	-
M. leprae TN	NP_301164.1	-
M. marinum M	YP_001853710.1	YP_001853545.1
M. smegmatis str. MC^2 155	YP_884482.1	YP_884481.1
M. sp. JLS	YP_001068363.1	YP_001068362.1
M. sp. KMS	YP_936087.1	YP_936086.1
M. sp. MCS	YP_637247.1	YP_637246.1
M. sp. Spyr1	YP_004074632.1	YP_004074631.1
M. tuberculosis CDC1551	NP_338541.1	NP_338540.1
M. tuberculosis F11	YP_001289835.1	YP_001289834.1
M. tuberculosis H37Ra	YP_001285263.1	YP_001285262.1
M. tuberculosis H37Rv	YP_178022.1	YP_178021.1
M. tuberculosis KZN 1435	YP_003033925.1	YP_003033924.1
M. ulcerans Agy99	YP_906673.1	YP_907832.1
M. vanbaalenii PYR-1	YP_950934.1	YP_950933.1
Mycobacterial-Related Species		
Rhodococcus equi 103S	YP_004005014.1	YP_004005015.1
Rhodococcus equi ATCC 33707	ZP_08154865.1	ZP_08154864.1
Rhodococcus erythropolis SK121	ZP_04387847.1	ZP_04387845.1
Rhodococcus jostii RHA1	YP_703992.1	-
Rhodococcus opacus B4	YP_002781107.1	YP_002781243.1
Segniliparus rotundus DSM 44985	YP_003659713.1	YP_003658055.1
Segniliparus rugosus ATCC BAA974	ZP_07967135.1	ZP_07964203.1
Tsukamurella paurometabola DSM 20162	YP_003645309.1	-

Shown are accession numbers for closest protein homologs (highest identity to the MTB strain CDC1551 proteins) of the ancestral PE35 and PPE68 from the respective species.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.