# 4. Diversity of the PE and PPE gene families from *Mycobacterium tuberculosis*

Part of this analysis has been published in:

J. M. Bryant, S. R. Harris, J. Parkhill, R. Dawson, A. H. Diacon, P. van Helden, A. Pym, A. A. Mahayiddin, C. Chuchottaworn, I. M. Sanne, C. Louw, M. J. Boeree, M. Hoelscher, T. D. McHugh, A. L. C. Bateson, R. D. Hunt, S. Mwaigwisya, L. Wright, S. H. Gillespie and S. D. Bentley (2013). "Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study." The Lancet Respiratory Medicine. DOI: 10.1016/S2213-2600(13)70231-

Statement of contribution

I performed all bioinformatic analyses and interpretation. PacBio sequencing and assembly was performed by Paul Coupland (WT Sanger Institute).

## 4.1. Introduction

The first genome sequence of *M. tuberculosis* (Cole, Brosch *et al*. 1998) revealed two novel gene families characterized by conserved Proline-Glutamate (PE) and Proline-Proline-Glutamate (PPE) residues at their N termini. These genes have been of significant interest due to their number, diversity and cell surface localization. Despite this, an understanding of these genes remains elusive and a clear function is yet to be determined.

A total of 168 PE and PPE genes have been described for *M. tuberculosis*, which make up almost 10% of its genomic coding capacity. The families are unique to Mycobacteria, found in varying numbers across the genus and are commonly cited to be particularly abundant in the more pathogenic species (Sampson 2011). However this observation may be merely anecdotal, as it appears that the number of genes seems to be more related to the traditional genus division into slow and rapid growers rather than their propensity to cause disease (Table 3).

**Table 3 - Numbers of PE and PPE genes annotated in the genomes of Mycobacterial species.** Genome annotations acquired from NCBI (NC_008596, NC_008726, NC_009338, NC_009077, NC_008703, NC_008146, NC_010394, NC_010604, NC_005916, NC_000962,NC_008595, NC_002944, NC_002677). * 27 PE/PPE pseudogenes not counted.

| Species | PE | PPE | Pathogenic to humans? | Growth phenotype |
|---|---|---|---|---|
| *M. smegmatis* | 6 | 4 | No | Rapid |
| *M. vanbaalenii* | 1 | 22 | No | Rapid |
| *M. gilvium* | 2 | 11 | No | Rapid |
| *M. sp JLS* | 2 | 12 | No | Rapid |
| *M. sp KMS* | 3 | 13 | No | Rapid |
| *M. sp MCS* | 2 | 11 | No | Rapid |
| *M. abscessus* | 3 | 6 | Yes- opportunistic | Rapid |
| *M. marinum* | 170 | 105 | Yes- opportunistic | Slow |
| *M. ulcerans* | 70 | 46 | Yes- opportunistic | Slow |
| *M. tuberculosis* | 99 | 69 | Yes- obligate | Slow |
| *M. avium 104* | 7 | 35 | Yes- opportunistic | Slow |
| *M. avium paraTB K10* | 6 | 36 | Yes- opportunistic | Slow |
| *M. leprae* | 1* | 2* | Yes- obligate | Slow |

The N termini of these proteins, which contain the PE or PPE motifs, are highly conserved in comparison to their C termini, which vary in both sequence and length (Cole, Brosch *et al*. 1998). This variable C terminus is either completely unique or contains specific repeat sequences allowing the families to be further subdivided (Figure 19). The largest subfamily consists of the PE-PGRS genes, which are typically long, extremely GC rich and full of tandem repeats. Other members have a simpler arrangement, and may only be composed of the conserved PE or PPE N termini domain. Only one crystal structure has been solved (Strong, Sawaya *et al*. 2006), which revealed that PE25 and PPE41 (which are found adjacent in the *M. tuberculosis* genome) form a 1:1 heterodimeric complex. It is unknown how representative this complex is of the proteins, but it is interesting that 18 other contiguous pairs of PE and PPE genes are found across the *M. tuberculosis* genome, suggesting that they too may form heterodimers as proteins.

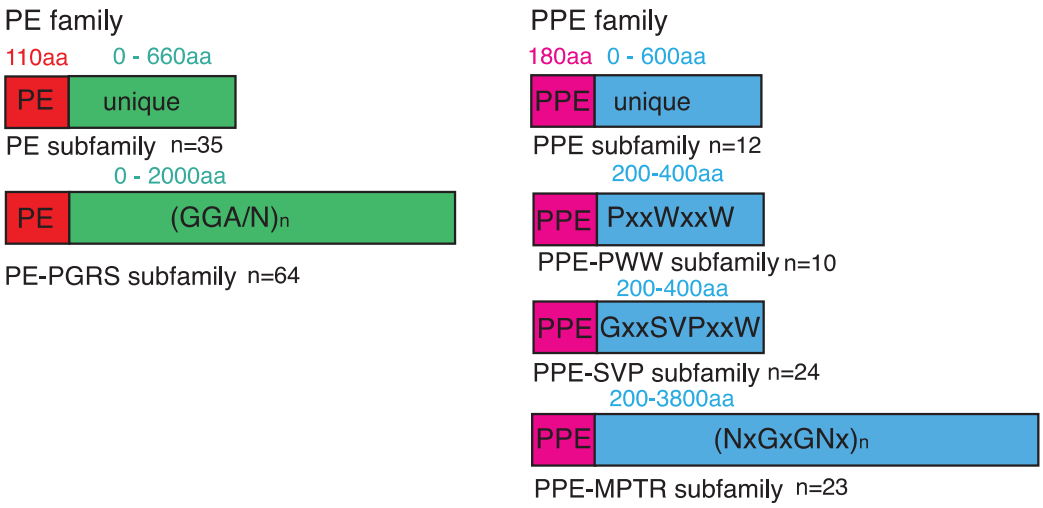


**Figure 19 - Subfamilies and structure of the PE and PPE genes.** The families are characterised by a conserved N terminal domain (red and pink) with variable C termini (green and blue). The amino acid motifs found in the C termini are stated with n indicating that they are tandem repeated. The approximate lengths of these regions in amino acids are indicated. PGRS = polymorphic GC-rich-repetitive sequence, MPTR = major polymorphic tandem repeat. Diagram adapted from Pittius *et al* (Gey van Pittius, Sampson *et al*. 2006).

Our functional knowledge of the PE and PPE genes remains poor, and is limited to a number of disparate observations and studies on individual genes. One of the first observations was that many of the members are found in close proximity to the

ESAT-6 genes, suggesting they may have co-expanded via duplication (Gey van Pittius, Sampson *et al*. 2006). These genes encode proteins that form part of the type VII secretion apparatus responsible for the secretion of the immuno-potent and antigenic ESAT-6 proteins (Gey Van Pittius, Gamieldien *et al*. 2001) and experimental evidence from *M. marinum* suggests that it is also likely to mediate the secretion of the PE/PPE proteins (Abdallah, Verboom *et al*. 2009). A second key observation is that many of the proteins are thought to be cell surface associated or exported, as demonstrated experimentally for 22 of the members (reviewed by Sampson 2011). In addition many members are expressed specifically during macrophage invasion (Rachman, Strong *et al*. 2006) or granuloma formation (Ramakrishnan, Federspiel *et al*. 2000). Collectively, these findings suggest that in *M. tuberculosis* at least, the PE and PPE genes are intimately involved with pathogenicity and/or interactions with the human immune system. However, we are yet to determine a more specific role.

There is a great deal of evidence supporting the observation that there is a high level of diversity between isolates, for both individual genes (Talarico, Cave et al. 2005, Talarico, Zhang et al. 2008) and the gene families as a whole (McEvoy, Cloete *et al*. 2012). The PE-PGRS and PPE-MPTR subfamilies are thought to contribute most of this diversity, mainly through mutations in their long repeat-rich C terminal domains (Sampson 2011) (Figure 19). This has led to speculation that these genes may contribute to evasion of host immune responses through antigenic diversification (Cole, Brosch et al. 1998, Karboul, Mazza et al. 2008). However, there has been no direct evidence demonstrating within patient antigenic variation (Sampson 2011). One of the reasons our understanding of these genes is very much incomplete is that they are hard to both sequence and to analyse, due to their high GC content (up to ~85%) and their repetitive nature. Although the presence of diversity through SNPs, insertions/deletions and intra-genomic recombination has been observed by many studies, the parameters of this diversity have not been quantified. It is unknown at what rate it is generated and similarly at what level: whether at the lineage, cluster or within patient level. Understanding these processes has important implications for our understanding of these gene families.

Here, two analyses of these gene families are presented. The first utilises sequences generated for the ReMoxTB project (Chapter 3) where pairs of isolates were collected from patients with multiple disease episodes. By looking at the PE/PPE genes in relapse cases, this study aimed to determine if there is variability generated within patients. The second analysis utilised sequence data from a study on the Beijing lineage in Samara, Russia (Casali, Nikolayevskyy *et al*. 2014). The Beijing clone is an important lineage which is thought to have recently spread across Asia and Eastern Europe, and has gained much interest due to its rapid global dissemination and high drug resistance. The aim of this work was to determine how much variation had been generated in the PE/PPE gene families and relate this back to the phylogeny.

## 4.2. Methods

Previously high-throughput sequence analysis of the PE and PPE genes was impossible due to inherent GC bias in the sequencing technology. This was overcome by the sequencing development team (WT Sanger Institute) who found that library preparation with the Kapa Hifi enzyme greatly improved coverage in GC rich regions (Quail, Otto *et al*. 2011). For this reason all the library preparation for Illumina sequencing in these analyses was carried out with Kapa Hifi (Kapa Biosystems, MA USA).

Sequencing data from 48 isolates were used for the first part of this study, which are composed of 24 pairs of isolates from baseline and relapse disease episodes collected during the ReMoxTB clinical trial. Velvet (Zerbino and Birney 2008) was used to *de novo* assemble the reads with scaffolding enabled (Methods 8.4). Raw reads were mapped back to the assembly to mask or correct possible assembly errors. Sequences of the 160 PE and PPE genes annotated in H37Rv, were extracted from the assemblies using an in-house script that uses a simulated PCR approach where upstream and downstream ‘primer’ sequences are specified. Alignments were made using Muscle (Edgar 2004) for each gene for the 48 isolates. SNPs, insertions and deletions were identified between the pairs using a custom script. The few differences identified were manually checked assessed using raw mapping data.

For the second part of the study, the sequencing data from 186 isolates available from a study in Samara, Russia were used (Casali, Nikolayevskyy *et al*. 2014), which all belong to the Beijing lineage. PacBio sequencing was carried out on one of these samples by Paul Coupland in the sequencing development team (WT Sanger Institute). The reads were *de novo* assembled and corrected using software provided as part of the SMRT Analysis software package (HGAP and Quiver - https://github.com/ PacificBiosciences/SMRT-Analysis). The short-read Illumina data from the same sample was then mapped onto the reference (Methods 8.2), in order to correct any possible single-base assembly errors, of which there were only ten. Members of the PE and PPE gene families were identified using a simulated PCR approach, and manually annotated. Low complexity regions were identified using DustMasker (Morgulis, Gertz *et al*. 2006) on default settings. The raw sequencing data from the 186 isolates were then mapped onto this final reference using GATK to realign insertions and deletions (Methods 8.2). Variants were called and trees built as described in Methods 8.3 and 8.6.

## 4.3. Results

### 4.3.1. Diversity of PE/PPE genes in relapse cases

Overall 86% of genes (82% PPE, 97% PE, 84% PE-PGRS) were assembled which enabled 82% of comparisons to be made between the 24 pairs. A minority of genes could not be assembled at all (see Appendix 9.3). Many of these can be attributed to common deletions such as PPE57, PPE58 and PPE59 which are found in the RD6 region and are flanked by IS sequences (Brosch 2002). Similarly PPE38 is a hotspot for IS6110 integration (McEvoy, Cloete *et al*. 2012); PPE54 contains tandem repeats and PPE46-47 are near identical so could not be assembled.

When comparing the assembled genes between the ReMoxTB pairs, no SNPs could be identified. A few single base pair deletions were detected, however, these were considered untrustworthy as they were found in homopolymeric tracts, which are susceptible to a common assembly error caused by collapsed repeats. In addition they were not supported by mapping data when mapped to the H37Rv reference.

### 4.3.2. Diversity of PE/PPE genes across the Beijing lineage

#### 4.3.2.1. PacBio reference sequence

Although an assembly approach has its advantages, it can be very labour intensive and not suitable for high throughput analyses. Mapping is quick and requires less manual curation, as it is not prone to artifacts that are the result of miss-assembly. Mapping does however, suffer from the problems of mismapping especially when the reference is too genetically disparate from the isolate in question. Thus in order to use a mapping approach to investigate PE/PPE diversity across the Beijing lineage, a suitable high quality reference is required, as genetically similar as possible to the samples under investigation. PacBio sequencing was carried out on a sample from the Samaran collection, 2535G, which was chosen on the basis of its high coverage (via Illumina sequencing) and central phylogenetic position within the lineage. An assembly of the raw PacBio reads resulted in 11 contiguous sequences.

Members of the PE/PPE genes on this sequence were identified using an *in silico* PCR approach using primers designed from the H37Rv sequence. In total, 157 of the 159 genes found in H37Rv were successfully annotated. All had assembled completely, except for PE_PGRS6 which was situated on a contig boundary. Only small fragments of PE_PGRS9, wag22 and PE_PGRS43 were found at the ends of contigs, and so were excluded from further analyses. A number of genes were identified as pseudogenes due to truncation or frameshifts outlined in Table 4.

**Table 4 - Pseudogenes identified in reference sequence, 2535G.** *Undetermined due to the requirement of high quality assemblies to assess whether this IS element is present at this position in all the strains.

| Gene | Type of mutation | Shared by all isolates? |
|---|---|---|
| PE_PGRS13 | frameshift | All |
| PE_PGRS57 | frameshift | just S535G |
| PPE34 | truncation due to IS element | Undetermined* |
| PPE38 | truncation due to IS element | Undetermined* |
| PPE39 | truncation due to IS element | Undetermined* |
| PPE56 | frameshift | All |
| PPE57 | frameshift | All |
| PPE58 | frameshift | All |
| PPE66 | stop codon | All |

#### 4.3.2.2. Rate of SNP accumulation in PE/PPE genes

186 isolates from the Samaran collection were mapped to the reference sequence, and an average of 95.1% of the base calls in the PE and PPE genes passed quality filters. The resultant phylogenetic tree (Figure 20) revealed a clonal structure as described previously (Casali, Nikolayevskyy *et al.* 2012). Two outlying isolates (non-Beijing but belong to the East Asian clade) were excluded from further analysis but were used as out-groups to correctly root the tree.

In total, 3931 sites across the genome were found to have a high quality SNP, of which 6.3% were within the PE and PPE genes (n=249). This is extremely close to what would be expected by chance (the PE and PPE genes make up 6.2% of the genome), suggesting that there is not an excess number of SNPs in these genes. When the number of PE/PPE vs. non-PE/PPE SNPs per branch was plotted there was a similar relationship (Figure 21), where the proportion of PE/PPE SNPs was found to be 7%.

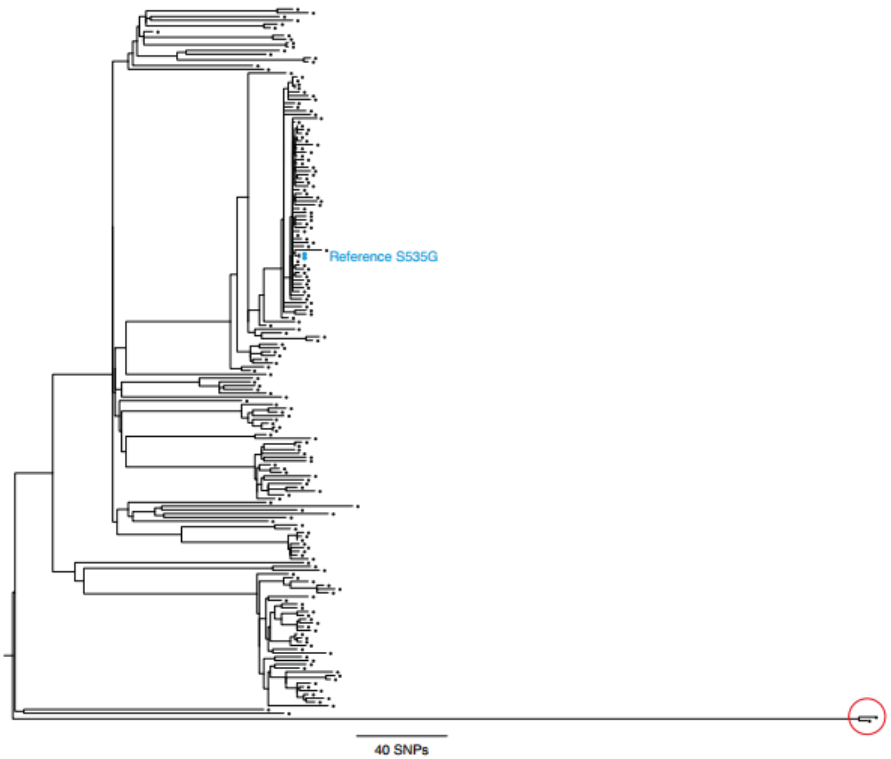


**Figure 20. Maximum likelihood phylogeny of Beijing isolates from Samara.** Isolates circled in red were excluded from further analyses but were used as an outgroup. The reference strain (both PacBio assembly and the mapping of original Illumina reads) are indicated in blue.

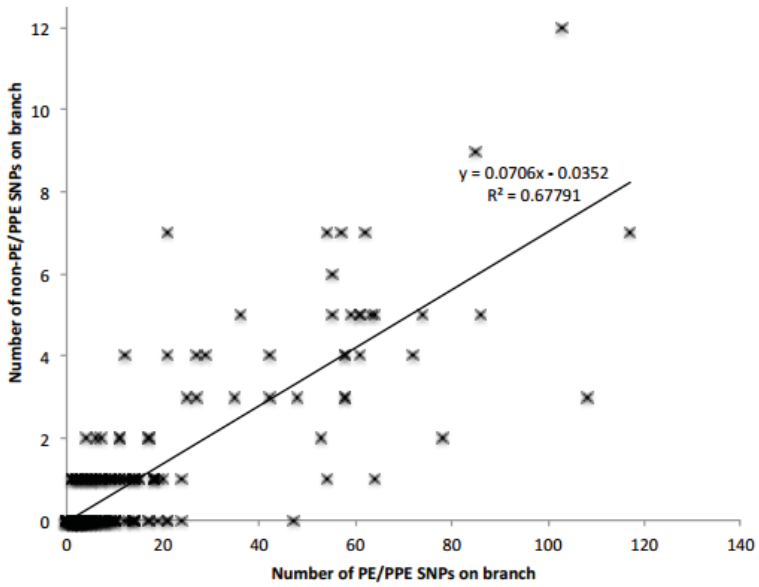


**Figure 21 – Plot of number of SNPs found in PE/PPE genes vs. rest of genome.** Each point represents a branch on the maximum likelihood phylogeny. Linear regression model p-value: < 2.2e-16.

The total dN/dS of all the PE/PPE SNPs was 0.77, which is slightly higher than the value calculated for the rest of the genome (0.66). When using the raw numbers of non-synonymous and synonymous sites, this was not significantly different using Fishers exact test (P=0.35). Overall this SNP analysis suggests that within this lineage there is no difference in the rate or type of SNP accumulation in this gene family in comparison to the rest of the genome.

#### 4.3.2.3. Insertions and deletions

Using a mapping approach, 134 insertions and deletions were detected within the PE/PPE genes, which were reconstructed onto the SNP-based phylogeny. Events that were found to be homoplasic (n=26), were inspected manually. 17 of these were found within two tandem homopolymeric tracts of PPE13:

GTGCCCCCCCCCCCAAAAAAAAGTA,

Homopolymeric tracts are highly susceptible to frameshift mutations, which were found to occur within these tracts 80 times, making these events extremely frequent. This region is positioned close to the C terminus of the gene (within last 12 amino acids), with a stop codon at a similar position in each alternative frame, making it less likely this frameshift could be deleterious. The other nine homoplasic deletions all occur in a small number of strains, close to the tips of the tree, and appear to be the

result of some variants passing quality filters in some strains but not in others. All these homoplasic variants (including the PPE13 frameshifts) were excluded from further analyses as they might obscure interpretation. An additional six insertions found close together within one strain were also excluded, which were likely due to a combination of low coverage and low level contamination from another bacterium. This left a total of 102 high quality insertions and deletions to interpret.

Overall when correcting for the number of sites, approximately four fold more insertions and deletions were detected in the PE/PPE genes than the rest of the genome (Figure 22). The statistical significance of this difference could not be assessed here but could possibly be investigated with more sophisticated evolutionary analyses implemented in a package such as BEAST ((Drummond and Rambaut 2007).

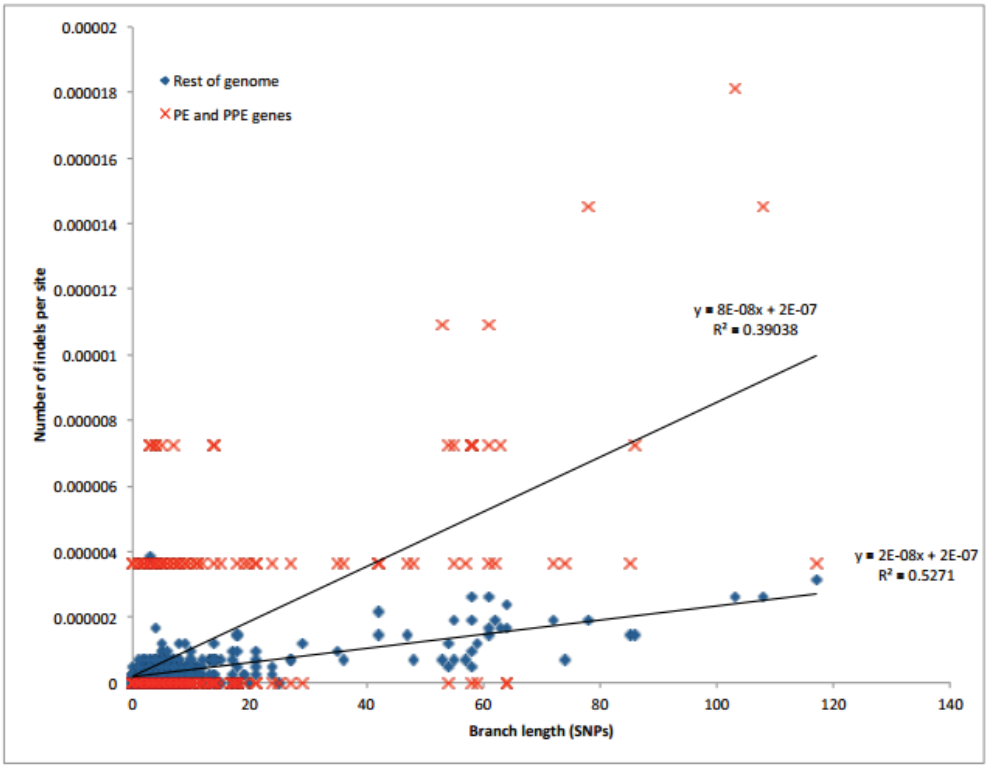


**Figure 22 – Number of insertions and deletions per branch found in PE and PPE genes (red) and rest of genome (blue) corrected for number of sites.** The PE/PPE genes had a total size of 275650bp, and the rest of the genome was 4176827bp. Linear regression R squared values and equations are shown.

Insertions and deletions were found in the PE-PGRS and PPE subfamilies, but were completely absent from the PE genes (Figure 23). When considering the total length

of these gene families, the number of PE-PGRS variants (0.02 per nucleotide) outweighs those found in PPE genes (0.002), by 10 fold. The majority of the variants (n=58) maintained the frame of the gene (length was a multiple of 3), so are less likely to be deleterious. Interestingly, the majority of these in-frame variants were found in the PE-PGRS subfamily (**Table 5**), and this was significantly more than would be expected by chance (Fishers exact test P = 0.0002). By chance, a third of variants would be expected to be in-frame; in the PE-PGRS genes the opposite was found, where 71% maintained the frame.

**Table 5 – Number of insertions and deletions found in the PE and PPE genes.**

| | PE_PGRS subfamily | PPE family |
|---|---|---|
| **In frame** | 54 | 4 |
| **Out of frame** | 22 | 14 |
| **Whole gene deletions** | 3 | 7 |

2 of the PPE and 1 of the PE_PGRS whole gene deletions were a single event.

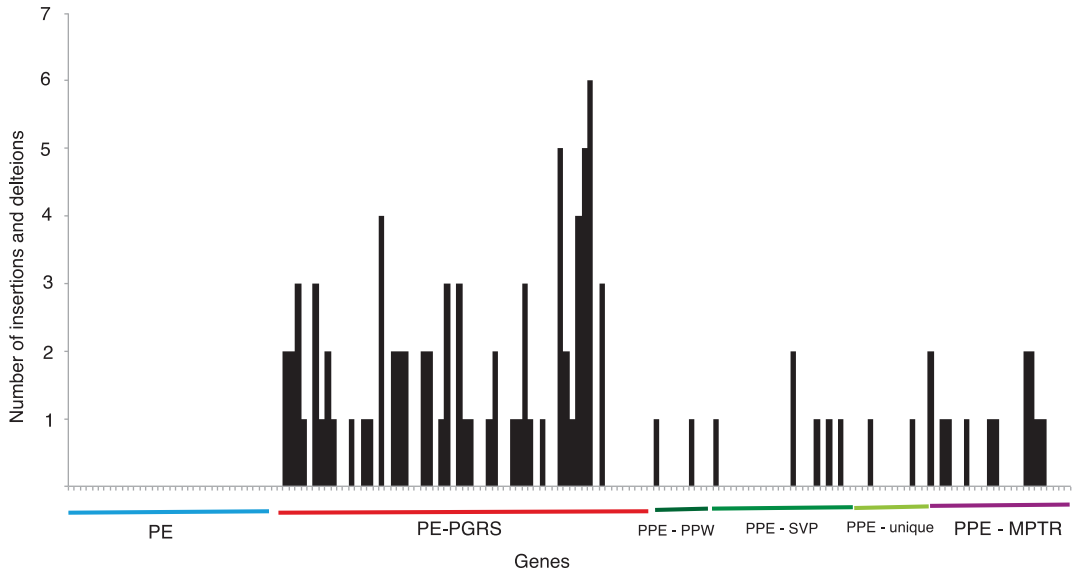


**Figure 23 – Numbers of insertions and deletions found in the genes – excluding homoplasies and PPE13 homopolymer frameshifts.**

Next, in order to investigate the sequence context of these insertions and deletions, DustMasker (Morgulis, Gertz *et al*. 2006) was used to predict the co-ordinates of low complexity or repetitive sequences. For the in-frame variants, the majority (51/58) were found in these repetitive regions, which was significantly more than the out-of-frame variants (17/36; Fishers exact test P value <0.001).

#### 4.3.2.4. Intra-genomic recombination

There is plenty of anecdotal evidence supporting the occurrence of intra-genomic recombination between PE and PPE genes (Liu, Gutacker et al. 2006, Karboul, Mazza et al. 2008, McEvoy, van Helden et al. 2009), however these events are yet to have been placed in a phylogenetic context. In order to identify such events within the Beijing lineage, SNPs were identified that occurred within 200bp of one another within the PE and PPE genes that were reconstructed onto the same phylogenetic branch. After manual inspection to exclude any events that were likely due to mapping error, seven events were identified and the contextual sequence extracted in order to look for a possible donor (Table 6). An exact match in another gene in the same genome was found for three of the sequences, two of them in a highly similar adjacent gene. None of the donor genes had a reciprocal event suggesting that gene conversion without crossover was the most likely mechanism. All three occurred on terminal branches, perhaps suggesting that these events aren't readily fixed in the population. The underlying basis for the other four events is unknown, but they could possibly represent the normal *de novo* generation of SNPs, inter-genomic recombination with another genome or mapping error.

**Table 6 - Possible intra-genomic recombination events identified.** These events represent the occurrence of more than one SNP within 200bp of one another on the same phylogenetic branch. For three of them a potential donor was identified.

| Branch (node ID->node ID) | No. of SNPs | Gene | Donor identified? |
|---|---|---|---|
| 17->ERR230004 | 2 | PPE5 | |
| 293->ERR227999 | 2 | PPE47 | Exact match in PPE46 |
| 293->ERR227999 | 4 | PE-PGRS50 | Exact match in PE-PGRS7 |
| 12->ERR228068 | 2 | PPE56 | |
| 355->ERR227984 | 4 | PPE56 | 3/4 found in PPE55 |
| 240->ERR227996 | 3 | PE-PGRS55 | |
| root->42 | 2 | PE-PGRS22 | |

## 4.4. Discussion

Although it has long been appreciated that there is a high level of genetic diversity in the PE and PPE genes, this evidence has been mainly anecdotal, with a few notable exceptions where studies have attempted to quantify the diversity over sub-sets of the family (McEvoy, Cloete et al. 2012, Copin, Coscolla et al. 2014). This work represents one of the most comprehensive analyses of the two families to date, over the most samples and the most genes. This has been achieved in part because of advances in sequencing technologies developed here at the WT Sanger Institute, which significantly reduced the effects of GC bias. However, despite this the data still presented methodological difficulties, meaning that the analysis was confined to short evolutionary distances and lacked a species-wide view. However, this did allow a more focused approach where the diversity could be quantified over two specific evolutionary scales.

The first study aimed to investigate diversity generated on the within-patient scale, by comparing two samples from two disease episodes. A striking and singular conclusion of this work was that there was none; there were no genetic differences detected within these genes between disease episodes. This supports previous observations made using RFLP analysis of PE-PGRS genes within patients, where no diversity in gene structure was detected (Richardson, van der Spuy *et al*. 2004). This result strongly suggests that these genes are not a source of antigenic diversity within patients. However this analysis was limited to around the 80% of the genes that could be assembled, so there could be missing variation generated in the remaining 20%. For this reason, in the second study a mapping rather than *de novo* assembly approach was used.

Using a mapping approach, the genetic diversity was captured across more genes and more isolates, as the method has the advantage of being automated without the possibility of miss-assembly. This enabled the investigation of PE and PPE gene diversity within a phylogenetic context across the Beijing lineage. It was found that the rate of SNP accumulation in these genes, and the corresponding dN/dS was indistinguishable from that of the rest of the genome. This suggests that any unusual levels of diversity observed in the PE and PPE genes are not being generated by *de*

*novo* SNPs, at least on this evolutionary scale. This is counter to previous observations that the PE and PPE genes had 3 fold more non-synonymous variants than the rest of the genome (McEvoy, Cloete *et al*. 2012). However, the variation in that study wasn't put in a phylogenetic context, which meant that they could not distinguish *de novo* SNPs from those introduced via recombination. In this study, there was strong evidence supporting three intra-genomic gene conversion events, which over larger evolutionary distances may be more numerous and a greater contributor to diversity, and thus dN/dS. A species-wide analysis will be required to investigate the contribution of intra-genomic gene conversion to diversity over larger evolutionary distances.

Instead of SNPs, an excess of insertions and deletions were detected: particularly within the PE-PGRS subfamily. These were dominated by in-frame variants located within repetitive or low complexity regions. The excess of in-frame variants was higher than would be expected by chance, suggesting that either there is a strong selection pressure to retain function in these genes, or the arrangement of repeats (in multiples of three bases) promotes events which are in-frame. The repetitive regions of the PE-PGRS genes are dominated by Gly-Ala repeats, which appear to be to be extremely similar to those found in the Epstein-Barr virus nuclear antigens (Cole, Brosch *et al*. 1998). In EBV, these repeats are involved in immune interference of cytotoxic T cells, so it is speculated that they may play a similar role in *M. tuberculosis*. In support of this, it's been found that the PE region alone is able to elicit a cell-mediated protective immune response, but when the PGRS domain is included this effect is lost (Delogu and Brennan 2001). If these genes were involved in immune interference, it's unclear what role if any a high level of genetic diversity would have. A recent study found few predicted T cell epitopes in the variable C termini of PE-PGRS genes (Copin, Coscolla *et al*. 2014), suggesting that this diversity is independent of T cell recognition. It's possible that the high level of diversity generated could be a side effect of the repeat structure required for immune interference, and that most of the in-frame insertions and deletions generated are neutral. However, even if the inter-genomic variation has little apparent function, it cannot be avoided that on the intra-genomic level these gene families are extremely diverse and that this diversity has been generated and maintained in *M. tuberculosis*.

Despite the extensive work carried out on the PE and PPE genes, they remain enigmatic, and yet more genetic diversity and experimental studies are probably required to gain further ground. However the work presented here has provided greater clarity on the underlying processes that are generating diversity within these genes. Furthermore, many in the field have been concerned that current mapping approaches may be missing a significant amount of diversity generated during transmission or outbreaks by excluding the PE and PPE genes. This work confirms that is unlikely to be the case as little variation seems to be generated over those time scales.