Abstract

Mycobacterium tuberculosis is a world-wide health crisis. CS1 is a relatively recent *M. tuberculosis* strain found in New Zealand which belongs to clade L4.4.1.1 (comprised of New Zealand and Canadian strains with a particular deletion). Formerly known as Rangipo, CS1 has been found to have both a high transmission rate and many mutations as compared with the reference strain. As such, these mutations could explain the transmission rate of CS1, and could expand our knowledge of the characteristics of this and other strains. This project looked at genetic anomalies in CS1 and found several different categories of interest for further research. A notable type were dense mutated regions inside and between genes, dubbed VRIs. Many of the genes impacted by VRIs were involved in virulence. It was also found that protein structures were affected by the VRIs, making VRIs significant and promising targets for further research into their involvement and impact. Ultimately, research into genetic anomalies found in this project will broaden our understanding of *M. tuberculosis* for improved prevention and treatment, and could lead to the eradication of CS1.

Introduction

In 2018, active infections of *Mycobacterium tuberculosis* (*Mtb*) affected 10 million people globally, killing 1.4 million. There were also around 1.6 billion more who had an infection without disease (latent *Mtb*) (1). In New Zealand the same year, around 300 people developed active infections, most of them aged 25-34 but three percent were children (2). The mortality rate in New Zealand is 4% (2). Globally, *Mtb* has been the leading cause of death by infectious agent since 2007, above HIV/AIDS (3).

The only current vaccine against *Mtb* is Bacillus Calmette–Guérin (BCG), which is moderately successful (1). In children it provides roughly a 20% reduction in infection, and a 71% reduction in infections becoming active (4), although these results vary significantly between regions and populations (3). It is more effective protection against severe forms of disease than against all *Mtb*, with protection lasting 15 years and longer depending on age of vaccination and environmental factors (3). Treatment of active non-drug resistant *Mtb* involves six months of a multi-drug rifampicin-based regimen (5), with a 3.1% chance of relapse and 50-94% risk of developing multi-drug resistance (6). New drugs and vaccines are a priority, and require discovering drug targets and learning more about *Mtb* and its strains (1).

There are several lineages of *Mtb* strains, the most globally-widespread lineage being L4, which has regionally-restricted sublineages and is known to be more virulent than other lineages (7). While *Mtb* is associated with residents in and immigrants from Africa, Asia, and the Pacific (8), (9), New Zealand has its own clusters of *Mtb* strains (10), and 43% of New Zealand-born cases of L4 are sub-lineage L4.4 (7). In 2016, New Zealand-born cases were 41.4% Māori, 13.8% Pacific Islanders, and 36.2% European/other ethnicity, with 60.9% cases residing in socio-economically deprived areas. Māori and Pacific Islanders are almost twice more likely than average to have a non-unique molecular-typed strain belonging to a cluster (10), like those found within clade L4.4.1.1. This clade contains two of the three New Zealand strains (Otara and CS1) and the Canadian strains, and has a deletion similar to but independently evolved from that found in L2 (which the Beijing strain belongs to), called RD152 in L2 & $DS6^{Quebec}$ in L4.4.1.1 (11). As the Beijing strain and its relatives are very virulent (12), this is of interest.

A particular New Zealand strain of interest from the L4.4.1.1 clade is Colonial Strain 1 (CS1, formerly known as Rangipo), an important *Mtb* source in Māori (11). It likely arose from a 1980's clonal expansion combined with social changes and urbanisation affecting Māori in particular (11). It has been responsible for some unusual prolonged outbreaks (13), and is noted for its high transmission rate (14). It has been noted that CS1 has extra virulence genes and mutations not found in the reference *Mtb* strain H37Rv, with which it otherwise shares high similarity (14). These genes and mutations could provide an advantage for CS1 and explain the high transmission rate, in particular the mutations could change the structure of existing proteins.

The aim of this project was to examine potentially significant genetic abnormalities found in CS1 compared with other strains of *Mtb* using bioinformatic techniques. These abnormalities could provide new valuable targets for drugs and vaccines against the highly contagious CS1. They can also give us insight into the characteristics of this strain as compared with Beijing and H37Rv, without needing to manipulate it in a laboratory environment.

Methods

Initial Alignment

The full genomes of CS1 (NZ_CP044345), Beijing strain W148 (NZ_CP012090), and reference strain H37Rv (NC_000962) were obtained from NCBI. W148 was selected as a member of L2, to identify the position of the RD152/DS6^{Quebec} deletion. The three sequences were aligned with the Mauve alignment plugin (15) in Geneious Prime (16) for comparison, using the default settings of the progressiveMauve algorithm with automatically calculated minimum locally colinear block (LCB, an orthologous segment) score and seed weight.

Comparison and anomalies

The analysis of the Mauve Alignment was done manually, by inspecting the whole alignment for genetic anomalies between the three strains, and tabulating the anomalies found. Anomalies were recorded, categorised, and appraised for significance (i.e. large mutations, insertion-deletions larger than three base pairs inclusive). The similarity of the strains was calculated from Geneious. W148 was ultimately excluded from the count due to a large number of W145-specific anomalies found in the three-way alignment.

3D structural predictions for mutation-affected proteins

To assess potential significance on protein products, one category of anomaly was further investigated. Two proteins with relatively large anomalous regions were selected and run through different protein prediction tools to analyse the impact of these regions. Results were compiled and amalgamated into a 3D structural prediction for each sequence. Basic Local Alignment Search Tool (BLAST) (17) was used to find similar proteins and identify any existing structures. An overview of protein features and domains was provided by InterPro (18), while the secondary structure was predicted by Porter 5.0 (19). Helices of interest were checked for hydrophobicity and amphipathicity using HeliQuest (20). For the membrane protein, transmembrane regions were found with TMPred (21), Phobius (22), and TMHMM 2.0 (23), and visualised with Protter (24). For the cytoplasmic protein, a 3D model was created from a template using SwissModel (25).

Results & Discussion

Initial Alignment

Mauve was used to align the genomes and show large-scale rearrangements, seen in Figure 1. Mauve found six LCBs, some of which had boundaries in the middle of genes for one or more strains. From this view, the main difference between CS1 and H37Rv is where the circular genome was opened, which was between LCB-6 & LCB-1 for H37Rv but between LCB-3 & LCB-4 for CS1. W148 was different, as the positions of LCBs 2 and 5 were swapped and both inverted compared with CS1 and H37Rv, although it had the same opening position as H37Rv.

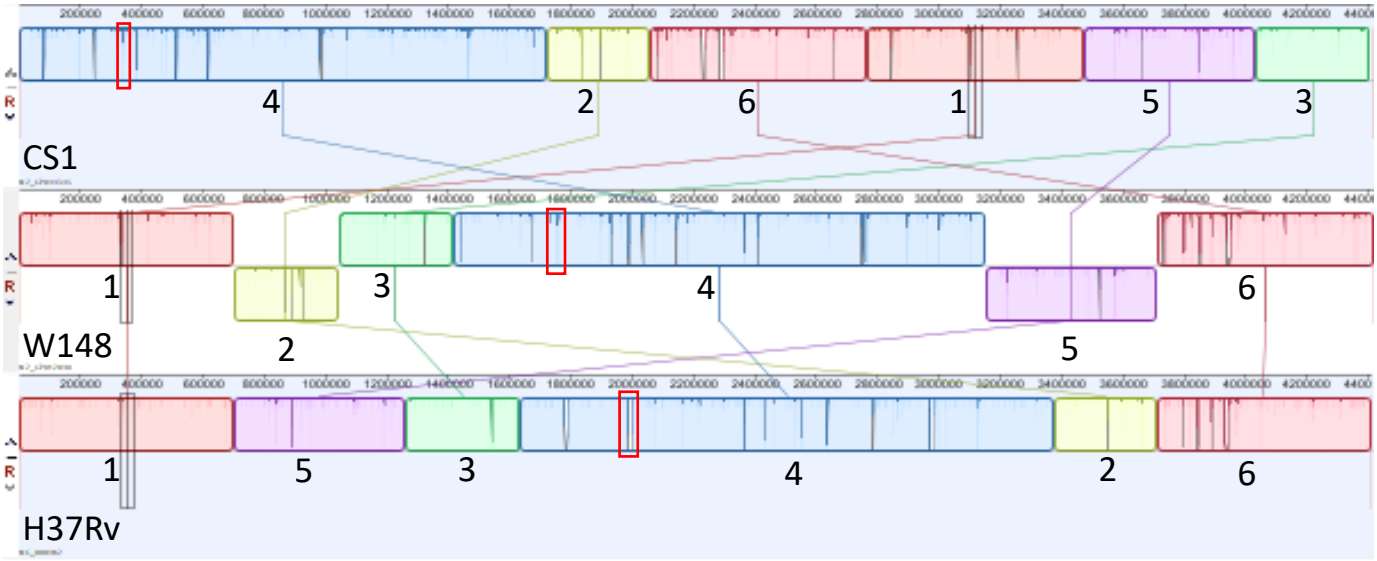

Figure 1: The Mauve alignment of the three strains with labelled LCBs, showing the different opening points for H37Rv/W148 and CS1, and the rearrangements found in W148. Dropped sections show that section is inverted in that strain. The positions affected by the RD152/DS6^{Quebec} deletions are shown in the red boxes.

The identity of the sequences was also examined to look at similarity at the base-pair level, to give a level of expectation for mutations. Table 1 shows the pairwise identity between all three strains, as calculated in Geneious. The least similar LCB was 6, which had a high amount of significant deletions. LCBs 3 & 5 were the second and third smallest sections, and were 99% identical. This is even more of interest when the strains are compared in pairs.

Table 1: The pairwise identity and percentage of identical sites for the three-way whole-genome alignment created in Geneious plugin Mauve.

Alignment of all three	**LCB-1**	**LCB-2**	**LCB-3**	**LCB-4**	**LCB-5**	**LCB-6**	**Total**
Pairwise identity	98.0%	97.7%	99.0%	96%	99.0%	92.5%	**96.5%**
Identical sites	97.0%	96.6%	98.5%	94.0%	98.5%	89.1%	**94.8%**

To examine sequence similarity more closely, the strains were compared in pairs and this is seen in Table 2. From this, the most similar strains appear to be CS1 and W148, but only by a mere 0.6% over the least similar strains – H37Rv and W148. W148 was noted to be quite divergent when the entire alignment of all three was looked at closely, which was not expected with the values of similarity seen. It should be noted that the LCB's for each alignment are calculated separately and do not correspond. This could be related, as during manual inspection it was found that almost all mutations had two strains in agreement with the third changed. While by numbers CS1 and W148 seem similar, CS1 and H37Rv have less LCBs which indicate less rearrangements and could be why W148 was found to be more divergent than expected on inspection.

Table 2: The pairwise identity and percentage of identical sites for the pair-alignment created in Geneious plugin Mauve.

CS1 & W148	**LCB-1**	**LCB-2**	**LCB-3**	**LCB-4**	**LCB-5**	**LCB-6**	**Total**
Pairwise identity	96.5%	97.1%	95.6%	98.4%	98.9%	99.4%	**97.3%**
Identical sites	96.5%	97.1%	95.6%	98.4%	98.9%	99.4%	**97.3%**

H37Rv & W148	**LCB-1**	**LCB-2**	**LCB-3**	**LCB-4**	**LCB-5**	**LCB-6**	**LCB-7**	**Total**
Pairwise identity	98.3%	19.3%	99.8%	99.1%	96.1%	97.5%	94.1%	**96.7%**
Identical sites	98.3%	19.3%	99.8%	99.1%	96.1%	97.5%	94.1%	**96.7%**

CS1 & H37Rv	**LCB-1**	**LCB-2**	**Total**
Pairwise identity	95.5%	99.0%	**96.8%**
Identical sites	95.5%	99.0%	**96.8%**

Comparing Tables 1 and 2, we see the percentage of identical sites decreases more than expected in the alignment of all three strains, and this could be because some LCBs had more mutations than others. For example, in the H37Rv/W148 alignment, the second LCB was part of the PE_PGRS4 gene, part of a variable family discussed later, and affected by large insertions, deletions, and base changes, and consequently the identity score decreases. However, that same gene is found inside the fifth LCB between CS1 and W148, and is only mildly affected by small SNPs and insertion-deletions. This could then be linked to the problem above where W148 was more divergent than expected, because the number of LCBs are as much indication of similarity (less LCBs being more similar),as the identity score.

Comparison

During manual comparison base-for-base between the strains in each LCB, several types of anomaly were found. W148 contained the highest number of anomalies relative to CS1 and H37Rv, but the current project is focussed on CS1, so analysis has focused on differences in this strain with H37Rv. These anomalies and their potential significance will be discussed in this section.

SNPs, small insertion-deletions, and annotation errors

A large number of single nucleotide polymorphisms (SNPs) and 1- or 2-base pair insertion-deletions were seen. The largest number was seen in W148, but they were seen frequently in H37Rv and CS1. These mutations can change the protein's primary structure, cause early or late stop codons, or impact expression (26), but this project was

identifying larger mutations. The author agrees with Stucki and Gagneux (27) that a database of SNPs in *Mtb* is needed, and small insertions should also be included due to the impact of frameshifts on a gene product. It was noted that a substantial amount of genes were affected by annotation errors, for example in succinate dehydrogenase subunit A (*sdhA*, Rv3318/ F6W99_RS10130). In *sdhA*, CS1 has a 1-base pair deletion at position 300, which at the amino acid level causes multiple new stop codons, but the gene is annotated as full length and functional. This annotation error could be masking an interesting change in CS1. Another type of error involved the shifting, elongation, or creation of gene annotations, and there were 451 of these found between CS1 and H37Rv. It is possible that not all of these were in fact errors, but it would need to be researched experimentally and through looking at the amino acid changes caused by small insertion-deletions. It would be worth exploring all of these in the future, the SNPs because they can have a big impact and CS1 isn't well-researched yet, and the annotation errors because they could indicate sequencing mistakes and aged annotation software or human error. Both SNPs and the annotation errors highlight how greatly research into all *Mtb* strains is needed.

Considered anomalies

Several other anomalies were found that were deemed to be significant or of interest. These were categorised as run-on genes, Whole Protein Insertion-Deletions (WPIDs), Partial Protein Insertion-Deletions (PPID), and Variable Regions of Interest (VRIs). A complete list of these can be found in Appendix 1.

Run-on genes occur when a gene fully or partially covers the same space as 2 or more proteins in the other strain. This was usually caused by another mutation such as a small insertion-deletion, as shown in Figure 2. This anomaly was the second least common, likely because of the risk of creating non-functional proteins and pseudogenes (28). This was not investigated further in this project, but is significant and of great interest for future research.

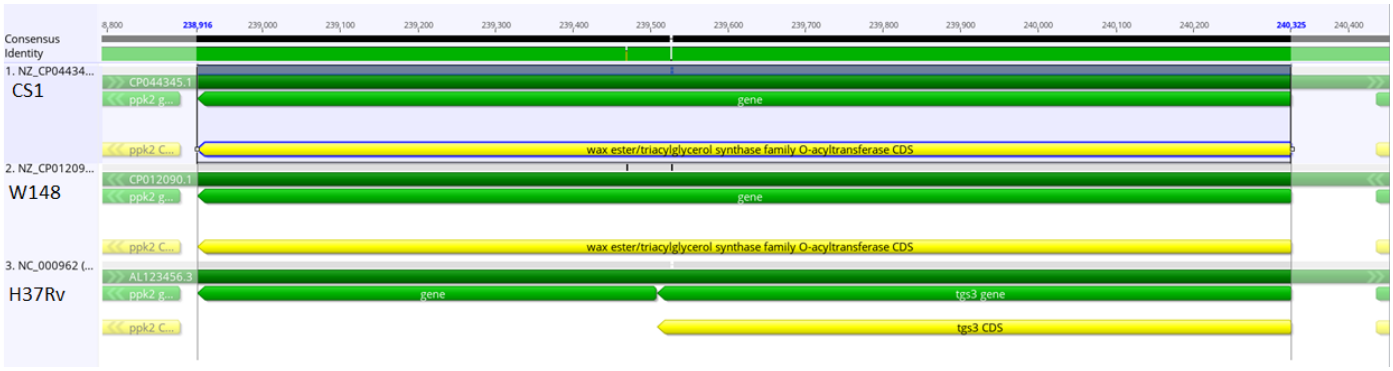

Figure 2: Example of a run-on gene, where a CS1 wax-ester/triacylglycerol synthase O-acyltransferase (F6W99_RS09685) is over two genes in H37Rv.

WPIDs are where an entire sequence has been inserted or deleted. These are often transposases and phage remnants, one of which is shown as an example in Figure 3. These created much of the genome length difference between CS1 and H37Rv, and often interrupted genes which could have an impact on the gene product. Another affect these can have is that some transposases, like IS6110, are known to act as promoters (29).

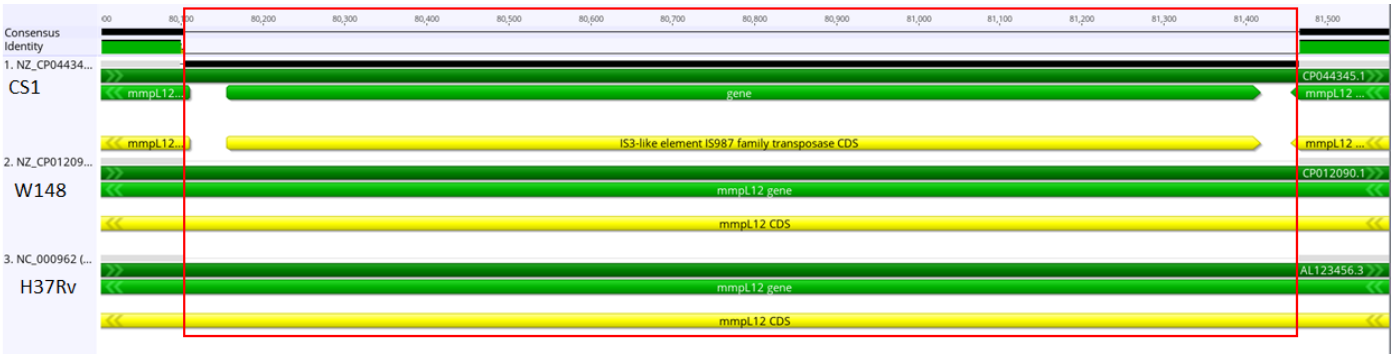

Figure 3: Example of a WPID, an IS3-like element IS987 family transposase (F6W99_RS00400) interrupting *mmpL12* in CS1. H37Rv and W145 are unaffected and have an intact *mmpL12*. The insertion is highlighted in the red box.

PPIDs, such as those shown in Figure 4 are sequences with insertion-deletions inside larger than 3 base pairs (usually substantially larger than this) as well as those with insertion-deletions affecting the beginnings or ends and their

length. Those affected at the beginning or end of the sequence were often affected by the ends of a phage or transposase insertion. These produced a good proportion of in-gene length differences between CS1 and H37Rv, and are significant as they produce frameshifts and change the amino acid sequence. This could have an effect on the function of the final product, which was out of the scope of this project.

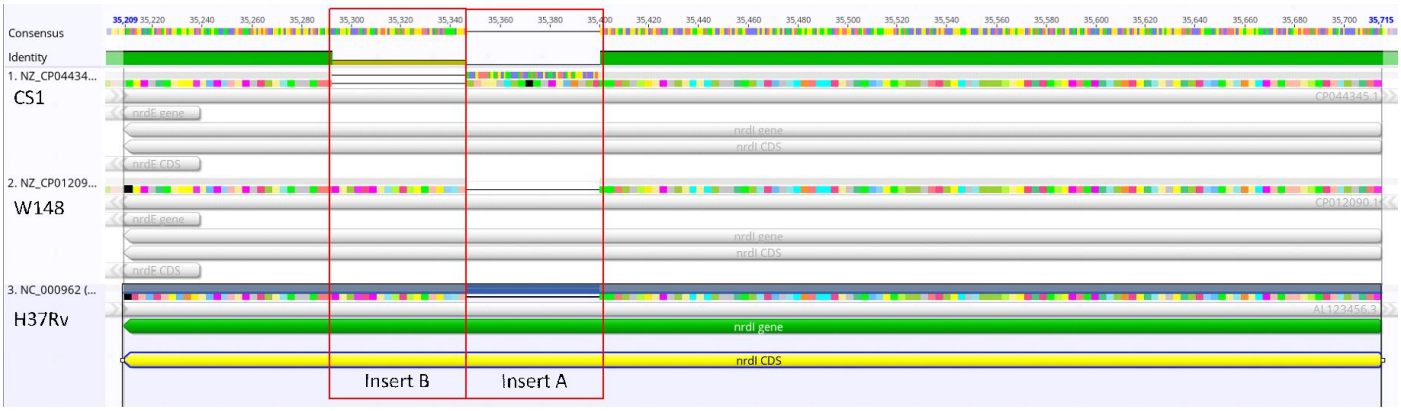

Figure 4: Example of a PPID, *nrdI* (Rv3052c), showing two sequential insertion-deletions. The gene runs from right to left, the larger coloured blocks are amino acids, with black indicating a STOP codon. Insert A is a 54-nucleotide insert in the CS1 sequence, causing an early stop codon. Insert B is also 54 nucleotide in length, and is present in both W145 and H37Rv but not CS1.

VRIs are sections of high-density mutations, often within protein-coding genes such as in Figure 5. The mutations can be base changes and insertion-deletions, and the amino acid sequence and final gene product structure could be changed. Some VRIs were also noted between genes, where they may change protein expression.

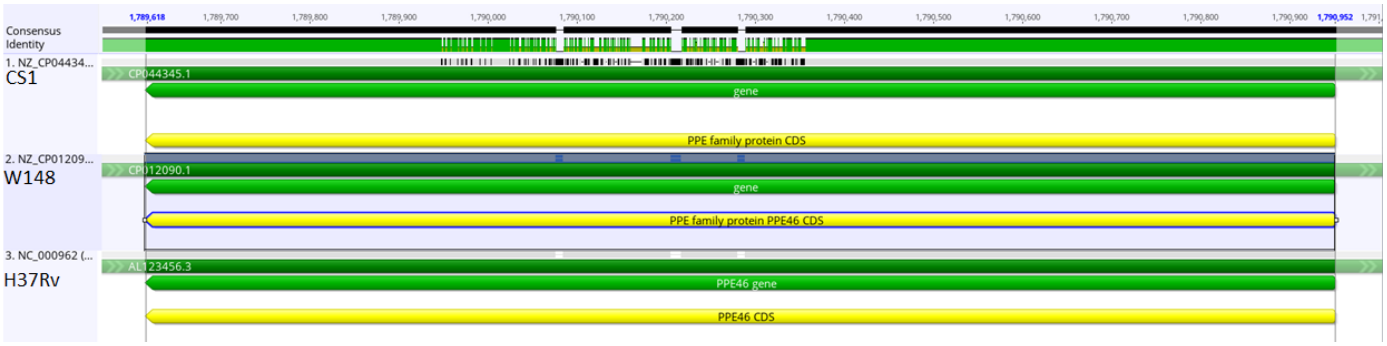

Figure 5: Example of a VRI inside PPE46 (Rv3018c). The black lines and blocks above CS1 show where it varies from the consensus sequence. The consensus sequence is the uppermost green and brown line; where it is white, CS1 has an insertion not seen in W148 or H37Rv, while short brown segments show mutations in one strain only (CS1 in this case) and green shows identical sections. The brown and white patches are closely packed together, which is the VRI.

We can look at the distribution of the different anomalies in Table 3, which shows the count of each type for each LCB using the three-way comparison, but excludes W148. PPIDs are the most common anomaly, with VRIs the least common. LCB-6 was noted in Table 1 having the least similarity, and this is seen here. Despite being roughly 42% of the size of the largest LCB (4), it has over half the anomalies that 4 has and 70% of 4's VRIs, which means the mutations are slightly more dense for this LCB. Similarly, LCB-3 had the most inter-strain similarity, and less anomalies than the smallest LCB, LCB-2. LCBs 1 and 5 are an interesting case, as LCB-1 is larger in size than LCB-5 and scored lower in the alignment similarity, but has less anomalies. This could indicate that these are not equally distributed, and could provide some kind of positive or negative impact (i.e. mainly found in non-functional or redundant genes, or mainly found in genes where the anomaly provides an advantage to the cell).

Table 3: Number of anomalies found in CS1 and H37Rv using the three-way comparison, sorted by Mauve LCB and anomaly type.

Anomaly	**LCB-1**	**LCB-2**	**LCB-3**	**LCB-4**	**LCB-5**	**LCB-6**	**Total**
Run-on	7	2	3	15	6	3	**36**
WPID	1	5	1	51	3	31	**92**
PPID	5	6	4	34	11	19	**79**

VRI	3	3	1	10	2	7	**26**
Total	**16**	**16**	**9**	**110**	**22**	**60**	**233**
Length of section in bp	716,582	343,170	377,804	1,793,095	564,083	757,321	4,418,548

VRIs

All the anomalies found could affect gene products, ultimately conferring advantages to CS1 and giving us information on its characteristics. This is reason to explore all of the above anomalies, however it is even more so for VRIs. These are regions, usually within proteins, with concentrations of mutations. This was the least common anomaly, and hit specific areas of proteins. Table 4 shows their size and where they are found.

Table 4: List of genes affected by VRIs, the size and density of the VRI, and information on the protein encoded.

Accession	**Size (mutations/span)**	**Annotation CS1**	**Annotation Reference**	**Function of region/gene (if known)**
Affecting F6W99_RS14965 & 70	69/126	N/A, affects length of fadD2		Between genes
Rv0279c	16/30 with outside 5 SNPs	PE family protein	PE PGRS4	PE
Rv0387c	11/143	PPE family protein	[no annotation]	PPE
Rv3021c	14/81 with outside 1 indel	PPE family protein	PPE47	PPE
F6W99_RS09600	48/78 with outside indel of 3	ATPase	Two component sensor kinase	
F6W99_RS09845	65/122	NTP transferase domain-containing protein	manB	May phosphorylate mannose (30)
Rv1452c	31/957 with two clusters of density	PE family protein	PE PGRS28	PE
Rv1587c	6 mutations close to a deletion	DUF222 domain-containing protein	Hypothetical protein	Unknown
F6W99_RS01550	11/348 & 248/662 with outside 2	PPE domain-containing protein	PPE24	PPE
F6W99_RS03140	28/172 with outside 3 SNPs and 2 insertion-deletions	Type I polyketide synthase	pks12	Antibiotic family, mannose 2° metabolite (31)
Spanning Rv2353c & IS6110-8	36/66	Affects length of PPE39		Between genes
F6W99_RS04825	83/147	Glycine--tRNA ligase	glyS	Adds glycine to its tRNA (30)
F6W99_RS07360	58/106 with outside mutation of 9 and 1 SNP	cas10	CRISPR-associated protein Cas10/Csm1	CRISPR-associated protein
Rv2931	11/33 with outside 3 SNPs and 1 indel	ppsA	ppsA	Polyketide synthase (32)
F6W99_RS08130	34/68	Permease	Integral membrane protein	
F6W99_RS08445	185/408	PPE family protein	PPE46	PPE
F6W99_RS17435 / Rv0720	47/93	Hypothetical protein x2	rplR	rRNA, 50S (33). Not found in CS1 (2 genes instead)
Rv0978c	51/147 with outside SNP	PE domain-containing protein	PE PGRS17	PE
F6W99_RS11140	51/76	MCE family protein	lprN	Virulence factor (34)
F6W99_RS11230 / Rv3508 / Rv3514	Multiple large complex regions	PE family protein	PE PGRS54 / PE PGRS57	All PE

	spanning the proteins			
Affecting F6W99_RS11335 & F6W99_RS11340	171/240	N/A		Between genes
F6W99_RS11360	118/220	3-oxosteroid 1-dehydrogenase	kstD	Cholesterol degradation pathway (35)
F6W99_RS12680 / Rv3778c	49/103 flanked by 4 large indel sections	Aminotransferase class V-fold PLP-dependent enzyme	Aminotransferase	
F6W99_RS13015	24/48	PheA		Aromatic amino acid pathway (36)

As shown, 23 of the 26 VRIs are inside a protein, and 16 of those are within genes which have an annotated function in at least one strain. Three VRIs are within run-on genes, and may be the cause of that run-on. One VRI was found to be directly downstream of a known hypervariable region (37).

PE/PPE proteins with VRIs

It is of note that the PE/PPE family is represented in this list, with regions found in five PE and four PPE proteins with a further PPE affected by an upstream region. PE/PPEs all share the Proline-Glycine (PE) or Proline-Proline-Glycine (PPE) motif which can be used as a signal peptide (38), but there are also sub-groups distinguished by extra motifs (for example PPE46 belongs to the PPW subgroup) and different regulation (39), (40). PE/PPEs are exclusive to the *Mycobacterium* genus (41) and several, especially PPE46 & 24, are promising drug targets (42). Their functions are largely unknown, although they are highly immunogenic and could be a part of antigenic variation (41). They have been shown to be the most consistently expressed protein types throughout the stages of infection (32). This family, particularly PPE46, has been implicated in disease pathogenesis, with the non-pathogenic relative *Mycobacterium smegmatis* having only a few PE/PPEs limited to a few subgroups (39). Many are associated with the ESAT-6 secretion system, which appears to be the origin of the protein family (39). It is also of note that PPE24 is predicted to be an essential protein for *in vitro* growth (39).

Non-PE/PPE known proteins with VRIs

Because PE/PPEs are known to play roles in virulence, the remaining nine genes with annotated functions were explored to look at function. If these gene products also play a role in virulence, they could explain the high transmission rate of CS1. It was found in Table 5 that seven are associated with virulence in some way, with the remaining two essential for cell function.

Table 5: A list of the nine functionally-known non-PE/PPE proteins affected by VRIs and their descriptions.

Gene	**Description**
manB	manB is known to be overexpressed in *Mtb*, causing greater association with human macrophages which is important for pathogenicity (43). It is a phosphomannomutase and the product is used for cell wall lipoarabinomannan, which is important for processes like phagocytosis as well as regulating human dendrocytes (43). The protein product was studied further to investigate the structural effect of the VRI.
pks12	pks12 is the largest ORF in the genome, producing two enzymes with different substrate selectivities (44), (45). It is involved in making a phospholipid which activates CD1c-mediated human T cell responses in pathogenic *Mtb* but not its relative *M. smegmatis* (44), and in permeability and virulence (45).
glyS	Despite being a tRNA synthase found in the cytoplasm, glyS is a known antigen which is upregulated *in vivo* for pulmonary patients (46).
Cas10/csm1	cas10/csm1 is part of anti-viral defence system *Mtb* Class 1 type III-A CRISPR-Cas complex, which is commonly found in thermophiles (of which *Mtb* is not) (47), (48), (49).
ppsA	ppsA is a polyketide synthase (32) involved in making phthiocerol derivatives (30). It has been identified as a drug target (42) and linked with antibiotic resistance (50). It is also upregulated in RpoB mutants, which are rifampin-resistant (of note as rifampin is the current front-line drug) (51).

rplR	rplR is a ribosomal protein (33) which is not a drug target, although other rpl genes are (42). Due to rRNA binding role, the changes in CS1 could affect the ribosome complex.
lprN	lprN is an immunogenic lipoprotein coded by a highly polymorphic gene inside the mammalian cell entry operon 4 (52). Sensitisation to lprN causes higher T-cell proliferation rates, increased cytokines like nitric oxide and TNF-a, more Colony Forming Units in lungs and spleen, and greater lung damage (52), so this virulence lipoprotein is not suitable for immunisation use (53). It is involved in maintaining late-stage disease (52).
kstD	kstD is part of cholesterol degradation pathway. It is essential to utilise cholesterol as a carbon source, so if this is disrupted, growth in a cholesterol-based medium is inhibited while in other media cholesterol will accumulate in the membrane (35). Cholesterol accumulation negatively effects rifampin permeability, which allows antibiotic persistence (35). The VRI may not make the enzyme inactive, and if it does then we could expect to see more cholesterol accumulation, rifampin insensitivity, and antigen-masking in CS1 (35). The AD/ADD pathway this enzyme is a part of is used for cholesterol in all *Mycobacterium* (54), (55).
pheA	The gene product of pheA is prephenate dehydratase (PDT). PDT catalyses the reaction forming phenylpyruvate, which needs an extra aminotransferase to become phenylalanine (36). Often PDT and catalyst for the preceding step (chorismate mutase (CM)) are together in one protein, but not in the case of *Mtb* (36). It has some unusual environmental preferences for a bacterial PheA (36). This is a drug target (42) which catalyses the second step after the branch point for Tryptophan and Phenylalanine/Tyrosine (56). PheA is predicted to be regulated by iron-dependent regulator IdeR (57).

Proteins

Some predictions were run on two proteins, manB and PPE46, to explore the effect of VRIs on protein structure. PPE46 was chosen due to having a large VRI and drug-target significance. It has not been characterised and was predicted to be a membrane protein by InterPro, so more programs were used to predict its structure than for manB. ManB is a cytosolic enzyme which has been characterised in other bacteria, so unlike PPE46, SwissModel was able to predict the structure using a template, with other programs supporting that prediction. The importance to virulence and species-wide commonality of manB made it a good choice for investigations.

PPE46 Predictions

To look at the variability of this gene, PPE46 was examined at the amino acid and DNA sequence levels. A list of similar strains was found using BLAST, and their PPE46 sequences compared in Geneious and shown in Figure 6. Five of them were identical (Korean vaccine strain BCG Korea 1168P(58), Beijing-like strain 36918, Guatemalan L4 strain GG-134-11 (59), hypervirulent Shanghai strain H107 (60), S3, and TBDM2444), with H37Ra identical except for a late start. CS1 and H37Rv shared more similarity to each other than the other six, both in mutations at the C and N termini and in insertion-deletion regions. This is interesting as, unlike the other strains, BCG Korea 1168P is a *Mycobacterium bovis* strain and therefore not expected to be similar. H37Rv and H37Ra are related strains, the latter being the avirulent attenuated form of the former, and differences between them can show where genes important to virulence are located (61), (62). Mutations within PE and PPE genes have previously been noted between H37Rv and H37Ra (62). We can gather from this that the H37Rv and CS1 versions could confer improved virulence for their respective strains.

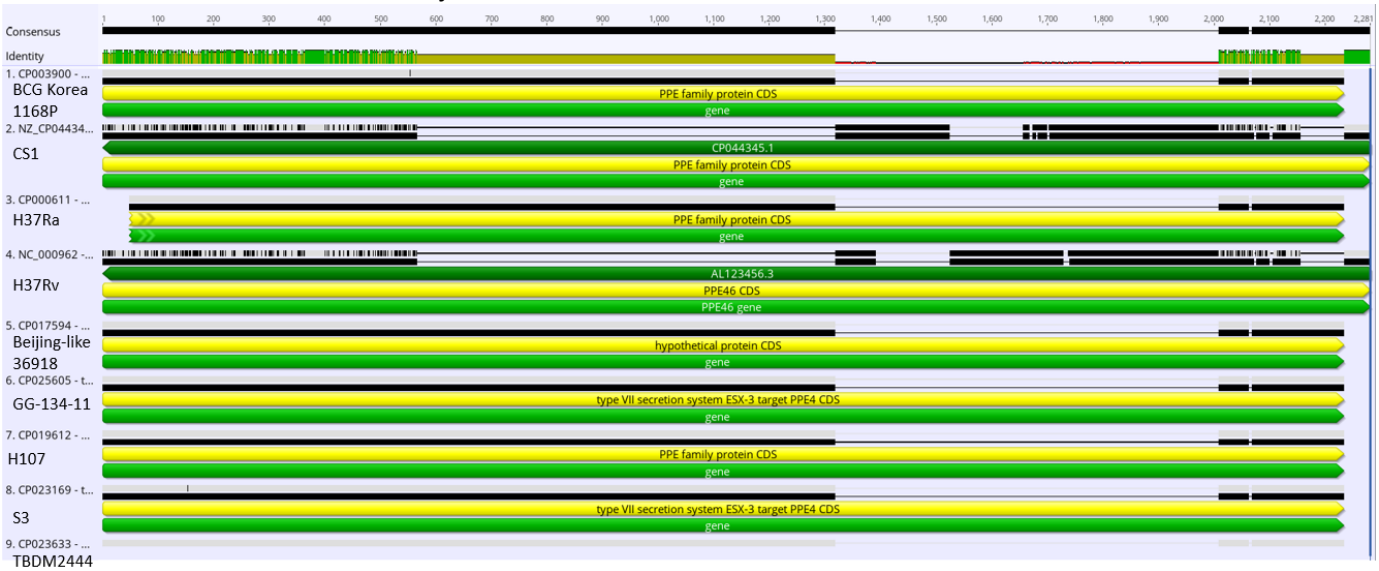

Figure 6: Geneious alignment of PPE46 genes in eight different strains of the *Mtb* cluster using the Mauve plugin. The top multicoloured bar shows the similarity – green is 100% identical, brown is over 30% identical, red is under 30% identical, and no colour shows that most strains have a deletion in this position. The pair of black bars over the yellow CDS annotations show the mutations – for the uppermost, black indicates base-pair mutation in that position for that strain; for the lower, a black line instead of the block shows a deletion.

To look at the relationships between the VRI, secondary folding, domains, and protein features, the predictions from Porter 5.0, InterPro, TMHMM, Phobius, and TMPred were compiled in Figures 7 and 8 for H37Rv and CS1 respectively. Porter 5.0 gave predictions for secondary structure using SS3 and SS8. TMPred gives a primary (TMPredP) and an alternative (TMPredA) prediction for transmembrane helices, with values over 500 considered significant. Red shows the variable region, italics and underlined residues are those used in HeliQuest predictions. All support InterPro's prediction of PPE46 as a membrane protein, predicting nine (CS1) and ten (H37Rv) helices in total, with two to three of these being transmembrane (but three transmembrane regions). However, TMPred gave between six and eight transmembrane helices, with the four predictions varying as to whether the N terminus was outside or inside the membrane. Where any TMPred prediction overlapped with other models, it was often of a different length (however, did confirm that the transmembrane helices predicted by Phobius and TMHMM likely existed). Phobius and TMHMM tended to match in transmembrane helices position and lengths, predicting three transmembrane helices with the N terminus outside. All programs agreed that the C terminus was inside the cell. We can be confident with having 9-10 helices and 3 transmembrane helices, with the N terminal outside the cell and in C terminus inside.

Number	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	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
Amino	M	T	A	P	V	W	L	A	S	P	P	E	V	H	S	A	L	L	S	A	G	P	G	P	G	S	L	Q	A	A	A	A	G	W	S	A	L	S	A	E	Y	A	A	V	A	Q	E	L	S	V	V	V
Porter3	C9	C9	C8	C5	C3	C2	C3	C5	C8	C7	H3	H5	H5	H6	H7	H7	H7	H6	H4	C2	C9	C9	C8	H1	H6	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
Porter8	C9	C9	C4	C5	C2	C1	C0	C0	S0	C6	H3	H5	H6	H6	H7	H8	H8	H7	H4	H0	S1	C0	C0	H5	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	Non-Cytoplasmic	PPE family IPR000030, PF00823 (8-164); PPE homologous superfamily IPR038332, G3DSA:1.20.1260.20 (5-178); Unintegrated superfamilies PTHR46766 (unnamed 9-183), 1SSF140459 (PE/PPE-dimer like 5-177); NC domain																																																		
TMHMM	Outside																																																				
Phobius	Non-Cytoplasmic																																																			
TMPredP	Inside																																																		In to Out (929)	
TMPredA	Outside																																																		

Number	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104
Amino	A	A	V	G	A	G	V	W	Q	G	P	S	A	E	L	F	V	A	A	Y	V	P	Y	V	A	W	L	V	Q	A	S	A	D	S	A	A	A	A	G	E	H	E	A	A	A	A	G	Y	V	C	A	L
Porter3	H8	H7	H3	C0	C5	C7	C8	C8	C8	C7	H6	H7	H9	H9	H9	H9	H9	H9	H9	H7	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
Porter8	H8	H7	H4	H1	T1	T0	T0	C4	C4	S5	H7	H7	H9	H9	H9	H9	H9	H9	H6	H7	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	PPE family IPR000030, PF00823 (8-164); PPE homologous superfamily IPR038332, G3DSA:1.20.1260.20 (5-178); Unintegrated superfamilies PTHR46766 (unnamed 9-183), 1SSF140459 (PE/PPE-dimer like 5-177); Non-Cytoplasmic domain																																																		
TMHMM	Outside																																																				
Phobius	Non-Cytoplasmic																																																			
TMPredP	In to Out (929)																											Outside																								
TMPredA	Outside									Out to In (771)																		Inside																								

Number	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156
Amino	A	E	M	P	T	L	P	E	L	A	A	N	H	L	T	H	A	V	L	V	A	T	N	F	F	G	I	N	T	I	P	I	A	L	N	E	A	D	Y	V	R	M	W	V	Q	A	A	T	V	M	S	A
Porter3	H6	C0	C9	C9	C9	H4	H6	H7	H7	H6	H6	H5	H9	H9	H9	H9	H9	H9	H9	H9	H7	H1	C4	C3	C3	C3	C5	C5	H1	H8	H9	H9	H9	H8	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
Porter8	H6	H1	S0	C8	C9	H5	H6	H7	H7	H6	H6	H6	H9	H9	H9	H9	H9	H9	H9	H9	H7	H0	C0	T1	T1	C0	T0	T0	H5	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	PPE family IPR000030, PF00823 (8-164); PPE homologous superfamily IPR038332, G3DSA:1.20.1260.20 (5-178); Unintegrated superfamilies PTHR46766 (unnamed 9-183), 1SSF140459 (PE/PPE-dimer like 5-177); Non-Cytoplasmic domain																																																		
TMHMM	Outside																																																				
Phobius	Non-Cytoplasmic																																																			
TMPredP	Outside																																																			
TMPredA	Inside																																																			

Number	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208
Amino	Y	E	A	V	V	G	A	A	L	V	A	T	P	H	T	G	P	A	P	V	I	V	K	P	G	A	N	E	A	S	N	A	V	A	A	A	T	I	T	P	F	P	W	H	E	I	V	Q	F	L	E	E
Porter3	H9	H9	H9	H9	H9	H8	H8	H8	H6	H3	C2	C9	C9	C9	C9	C9	C9	C9	C9	C6	C4	C5	C8	C8	C7	C1	H0	H3	H5	H5	H5	H6	H6	H6	H4	H2	C2	C6	C9	C7	C5	C0	H7	H8	H9	H9	H9	H9	H8	H9	H9	H9
Porter8	H9	H9	H9	H9	H9	H8	H8	H8	H7	H4	H2	C1	C5	C5	C5	C4	C6	C8	C8	C5	C2	C2	C5	T1	T2	H1	H3	H5	H5	H6	H7	H7	H7	H7	H6	H4	H1	C0	C1	C1	C2	C0	H6	H8	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	PPE family (164); PPE superfamily (178); Unintegrated superfamilies unnamed (183) & PE/PPE-dimer like (177); NC domain																											Non-Cytoplasmic domain																											
TMHMM	Outside																																																			
Phobius	Non-Cytoplasmic																																																			
TMPredP	In to Out (684)									Outside																																											
TMPredA	Out to In (811)																											Inside																								

Number	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
Amino	T	F	A	A	Y	D	Q	Y	L	S	A	L	L	S	E	L	P	A	V	A	W	V	W	F	Q	L	F	V	D	I	L	G	F	N	I	I	G	F	I	I	T	L	A	S	N	A	Q	L	L	T	E	F
Porter3	H8	H8	H7	H8	H8	H8	H8	H8	H9	H8	H8	H7	H4	C0	C2	H0	H4	H7	H8	H9	H9	H9	H9	H9	H9	H9	H8	H7	H7	H5	H4	H1	H2	H4	H6	H8	H8	H8	H8	H8	H8	H7	H6	H5	H2	H0	H6	H7	H7	H7	H7	H7	H7
Porter8	H9	H9	H8	H8	H9	H9	H9	H9	H9	H8	H8	H5	T0	H0	H3	H4	H7	H8	H9	H9	H9	H9	H9	H9	H9	H8	H8	H7	H7	H6	H4	H5	H6	H7	H8	H8	H8	H8	H8	H8	H7	H7	H6	H3	H2	H6	H7	H7	H7	H7	H7	H8
InterPro	Non-Cytoplasmic domain																											TMHMM & Phob (helix nested in 1 aa either side)																									Cytoplasmic domain										
TMHMM	Outside																											Transmembrane Out to In																									Inside										
Phobius	Non-Cytoplasmic																											Transmembrane Out to In																									Cytoplasmic										
TMPredP	Outside																											Out to In (1922)																									Inside										
TMPredA	Inside																											In to Out (1847)																									Outside										

Number	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312
Amino	A	I	N	A	S	Y	V	A	V	G	L	L	Y	A	I	A	G	V	I	D	I	V	V	E	W	V	I	G	N	L	F	G	V	V	P	L	L	G	G	P	L	L	G	A	L	A	A	A	V	V	P	G
Porter3	H7	H7	H7	H8	H8	H8	H7	H6	H7	H7	H7	H8	H8	H7	H7	H6	H5	H5	H7	H8	H8	H8	H8	H9	H8	H8	H7	H5	H4	H3	H2	C1	C1	H1	H3	H3	H0	C0	C1	H0	H3	H5	H5	H4	H5	H5	H4	H1	C0	C1	C1	C0
Porter8	H8	H8	H8	H8	H8	H8	H7	H7	H8	H8	H8	H8	H8	H7	H8	H7	H6	H7	H8	H8	H9	H9	H9	H9	H9	H9	H8	H6	H6	H6	H2	T0	H1	H2	H3	H2	H3	H2	H2	H4	H6	H7	H7	H7	H7	H7	H5	H3	H3	H0	H1	H2
InterPro	Cytoplasmic	TMHMM & Phob (helix ends 1 early)																												Non-Cytoplasmic domain												TMHMM & Phob (helix nested in 2 aa either side)										
TMHMM	Inside	Transmembrane In to Out																												Outside												Transmembrane Out to In										
Phobius	Cytoplasmic	Transmembrane In to Out																												Non-Cytoplasmic												Transmembrane Out to In										
TMPredP	Inside	In to Out (1280)																												Outside												Out to In (1700)										
TMPredA	Outside																																									Out to In (1700)										

Number	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364
Amino	V	A	G	L	A	G	V	A	G	L	A	A	L	P	A	V	G	A	A	A	G	A	P	A	A	L	V	G	S	V	A	P	V	S	G	G	V	V	S	P	Q	A	R	L	V	S	A	V	E	P	A	P
Porter3	H0	H0	H0	H0	H1	H2	H2	H2	H1	C0	C2	C3	C2	C1	C1	C2	C1	C2	C4	C5	C3	C2	C1	C0	C1	C2	C4	C5	C6	C7	C7	C7	C7	C7	C6	C4	C4	C7	C5	C3	C2	C1	C1	C2	C3	C5	C5	C5	C5	C8	C9
Porter8	H3	H3	H2	H2	H3	H3	H4	H4	H3	H2	H1	C0	C1	C2	C0	H0	C0	C0	C0	T0	C0	C2	C3	C1	C1	C1	C1	C2	C2	C2	C3	C6	C3	C2	C1	C2	C2	C2	C5	C3	C2	C1	C1	C1	C1	C1	C2	C2	C3	C4	C5	C4
InterPro	TMHMM & Phob (helix nested in 2 aa either side)																											Cytoplasmic domain																											
TMHMM	Transmembrane Out to In																											Inside																											
Phobius	Transmembrane Out to In																											Cytoplasmic																											
TMPredP	Out to In (1700)																											Inside																											
TMPredA	Out to In (1700)																											Inside																											

Number	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416
Amino	A	S	T	S	V	S	V	L	A	S	D	R	G	A	G	A	L	G	F	V	G	T	A	G	K	E	S	V	G	Q	P	A	G	L	T	V	L	A	D	E	F	G	D	G	A	P	V	P	M	L	P	G
Porter3	C8	C7	C7	C6	C5	C2	C0	E0	C2	C4	C5	C5	C6	C6	C5	C5	C5	C5	C3	C4	C4	C4	C5	C7	C7	C6	C6	C7	C7	C7	C6	C4	C2	C0	E2	E2	C0	C5	C5	C4	C4	C7	C8	C9	C9	C8	C6	C7	C7	C8	C7	C6
Porter8	C2	C1	C2	C3	C3	C1	E1	E1	C0	C2	C0	T0	C0	C1	C1	C1	C0	C0	C1	C1	C0	C2	C3	C2	C0	C1	C3	C1	C1	C3	C5	C1	E0	E1	E3	E2	E0	C1	T0	C0	T1	T0	S0	C2	C6	C5	C3	C5	C5	C5	T1	T3
InterPro	Cytoplasmic domain																																																				
Annotation	GFGT Motif																																																		PPW Motif	
TMHMM	Inside																																																			
Phobius	Cytoplasmic																																																			
TMPredP	Inside																																																			
TMPredA	Inside																																																			

Number	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434
Amino	S	W	G	P	D	L	V	G	V	A	G	D	G	G	L	V	S	V
Porter3	C6	C7	C8	C5	C4	C2	C4	C5	C4	C5	C6	C7	C8	C5	E2	E4	E2	C9
Porter8	C1	C1	C4	T0	T0	C0	T0	C0	C2	C2	C0	T1	T0	C1	E4	E5	E4	C9
InterPro	Cytoplasmic domain																	
Annotatic	PPW																	
TMHMM	Inside																	
Phobius	Cytoplasmic																	
TMPredP	Inside																	
TMPredA	Inside																	

Figure 7: H37Rv PPE46 prediction compilation of amino acid sequence, annotations of PPE motifs and areas of interest, Porter 5.0, InterPro, Phobius, and TMPred.

Number	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	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
Amino	M	T	A	P	V	W	L	A	S	P	P	E	V	H	S	A	L	L	S	A	G	P	G	P	G	S	L	Q	A	A	A	A	G	W	S	A	L	S	A	E	Y	A	A	V	A	Q	E	L	S	V	V	V
Porter3	C9	C9	C8	C5	C3	C2	C3	C5	C8	C7	H3	H5	H5	H6	H7	H7	H7	H6	H4	C2	C9	C9	C8	H1	H6	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
Porter8	C9	C9	C5	C5	C2	C1	C1	C0	S0	C6	H3	H5	H6	H7	H7	H8	H8	H7	H4	H0	S1	C0	C0	H5	H7	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	Non-Cytoplasmic	PPE family IPR000030, PF00823 (8-164); PPE homologous superfamily IPR038332, G3DSA:1.20.1260.20 (5-178); Unintegrated superfamilies PTHR46766 (unnamed 9-188), 1SSF140459 (PE/PPE-dimer like 5-177); NC domain																																																		
Annotation										PPE																																											
TMHMM	Outside																																																			
Phobius	Non-Cytoplasmic																																																			
TMPredP	Inside																																																	In to Out (929)		
TMPredA	Outside																																																Out to In (768)		

Number	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104
Amino	A	A	V	G	A	G	V	W	Q	G	P	S	A	E	L	F	V	A	A	Y	V	P	Y	V	A	W	L	V	Q	A	S	A	D	S	A	A	A	A	G	E	H	E	A	A	A	A	G	Y	V	C	A	L
Porter3	H8	H7	H3	C0	C5	C7	C8	C8	C8	C7	H6	H7	H9	H9	H9	H9	H9	H9	H7	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
Porter8	H8	H7	H4	H1	T1	T0	T0	C4	C4	S5	H7	H7	H9	H9	H9	H9	H9	H9	H6	H7	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	PPE family IPR000030, PF00823 (8-164); PPE homologous superfamily IPR038332, G3DSA:1.20.1260.20 (5-178); Unintegrated superfamilies PTHR46766 (unnamed 9-188), 1SSF140459 (PE/PPE-dimer like 5-177); Non-Cytoplasmic domain																																																		
TMHMM	Outside																																																				
Phobius	Non-Cytoplasmic																																																			
TMPredP	In to Out (929)																																																Outside			
TMPredA	Out to In (768)																																																Inside			

Number	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156
Amino	A	E	M	P	T	L	P	E	L	A	A	N	H	L	T	H	A	V	L	V	A	T	N	F	F	G	I	N	T	I	P	I	A	L	N	E	A	D	Y	V	R	M	W	V	Q	A	A	T	V	M	S	A
Porter3	H6	C0	C9	C9	C9	H5	H6	H7	H7	H6	H6	H5	H9	H9	H9	H9	H9	H9	H9	H9	H7	H0	C4	C3	C3	C4	C5	C5	H1	H8	H9	H9	H9	H8	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
Porter8	H6	H1	S0	C7	C9	H5	H7	H7	H7	H6	H6	H6	H9	H9	H9	H9	H9	H9	H9	H9	H7	H0	C0	T1	T1	C0	T0	T0	H6	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	PPE family IPR000030, PF00823 (8-164); PPE homologous superfamily IPR038332, G3DSA:1.20.1260.20 (5-178); Unintegrated superfamilies PTHR46766 (unnamed 9-188), 1SSF140459 (PE/PPE-dimer like 5-177); Non-Cytoplasmic domain																																																		
TMHMM	Outside																																																				
Phobius	Non-Cytoplasmic																																																			
TMPredP	Outside																																																	Inside			
TMPredA	Inside																																																	Out to In (811)			

Number	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208
Amino	Y	E	A	V	V	G	A	A	L	V	A	T	P	H	T	G	P	A	P	V	I	V	K	P	G	A	N	E	A	S	N	A	V	A	A	A	T	I	T	P	F	P	*F*	*G*	*E*	*L*	*A*	*K*	*F*	*L*	*E*	*M*
Porter3	H9	H9	H9	H9	H9	H8	H8	H8	H6	H3	C2	C9	C9	C9	C9	C9	C9	C9	C9	C6	C4	C5	C8	C8	C7	C1	H1	H3	H5	H5	H5	H6	H7	H6	H4	H1	C2	C6	C9	C8	C7	C3	H3	H5	H8	H9	H9	H9	H9	H9	H9	H9
Porter8	H9	H9	H9	H9	H8	H8	H8	H7	H4	H2	C1	C5	C5	C5	C4	C6	C8	C8	C5	C2	C2	C5	T1	T2	H1	H3	H5	H5	H7	H7	H8	H8	H7	H6	H3	H1	C0	C1	C2	C2	C1	H4	H6	H8	H9	H9	H9	H9	H9	H9	H9	H9
InterPro	PPE family (8-164); PPE homologous superfamily (5-178); Unintegrated superfamilies unnamed (9-188) & PE/PPE-dimer like (5-177); NC domain																																		Non-Cytoplasmic domain																				
TMHMM	Outside																																																				
Phobius	Non-Cytoplasmic																																																			
TMPredP	In to Out (684)																																																				
TMPredA	Out to In (811)																																																				

Number	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
Amino	*A*	*A*	*Q*	*A*	*F*	*T*	*E*	*V*	*G*	*E*	*L*	*I*	*M*	*K*	*S*	*A*	*E*	*A*	*W*	*A*	*V*	*G*	*F*	*V*	*E*	*L*	*I*	*T*	*G*	*L*	*V*	*N*	*F*	*E*	*P*	*W*	*L*	*V*	*L*	*T*	*G*	*M*	*I*	*D*	*M*	*F*	*F*	*A*	*T*	*V*	*G*	*F*
Porter3	H9	H8	H8	H7	H7	H7	H7	H7	H7	H8	H8	H7	H6	H5	H5	H7	H8	H9	H9	H8	H8	H8	H8	H9	H9	H9	H8	H6	H3	H1	C2	C4	C0	H5	H6	H7	H7	H6	H5	H2	C0	H1	H4	H6	H7	H8	H8	H8	H8	H8	H8	H8
Porter8	H9	H9	H9	H8	H8	H8	H8	H8	H8	H8	H8	H8	H7	H7	H7	H8	H8	H9	H9	H9	H9	H9	H9	H9	H9	H9	H8	H6	H4	H2	T0	T1	H2	H5	H6	H6	H7	H7	H6	H3	H1	H3	H5	H6	H8	H8	H8	H8	H8	H8	H8	H8
InterPro	Non-Cytoplasmic domain																																																			
TMHMM	Outside																																																			
Phobius	Non-Cytoplasmic																																																		
TMPredP	Inside																																																		
TMPredA	Inside																																																		

Number	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312
Amino	A	L	G	V	F	V	L	V	P	L	L	E	F	A	V	V	L	E	L	A	I	L	S	I	G	W	I	I	S	N	I	F	G	A	I	*P*	*V*	*L*	*A*	*G*	*P*	*L*	*L*	*G*	A	L	A	A	A	V	V	P
Porter3	H8	H8	H7	H6	H4	H1	H3	H3	H5	H6	H7	H7	H7	H7	H7	H7	H7	H7	H8	H8	H8	H7	H7	H7	H7	H8	H8	H7	H5	H5	H4	H2	C0	C0	H0	H3	H3	H2	H1	H2	H3	H5	H5	H4	H4	H4	H4	H3	H1	C0	C2	C2
Porter8	H8	H8	H7	H7	H5	H0	H2	H4	H6	H6	H7	H7	H7	H7	H7	H7	H7	H8	H8	H8	H8	H8	H7	H8	H8	H8	H8	H8	H7	H6	H6	H4	H1	H2	H2	H4	H5	H5	H4	H5	H6	H7	H7	H7	H7	H6	H5	H4	H3	H0	H1
InterPro	Phob & TMHMM							Cytoplasmic domain						TMHMM (273-295) helix; Phob transmembrane 275-299																													Non-Cytoplasmic								Phob & TMHMM					
TMHMM	Transmembrane Out to In							Inside						Transmembrane In to Out																											Outside								Transmembrane Out to In					
Phobius	Transmembrane Out to In									Cytoplasmic																																			Non-Cytoplasmic								Transmembrane Out to In					
TMPredP	In to Out (2101)												Outside																																								Inside			
TMPredA	In to Out (2101)												Outside																																								Inside			

Number	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364
Amino	G	V	A	G	V	T	G	L	A	G	L	A	A	V	P	A	V	G	A	A	A	G	A	P	A	A	L	V	G	S	V	A	P	V	S	G	G	V	V	S	P	Q	A	R	L	V	S	A	V	E	P	A
Porter3	C1	C0	C0	C0	C0	H0	H1	H1	H1	H1	H0	C1	C3	C4	C4	C3	C2	C3	C2	C2	C3	C5	C3	C2	C1	C1	C1	C3	C4	C5	C6	C7	C7	C7	C7	C7	C6	C4	C4	C7	C5	C3	C2	C1	C1	C1	C2	C5	C5	C5	C5	C8
Porter8	H1	H2	H2	H1	H1	H2	H3	H3	H3	H2	H1	H0	C1	C1	C4	C2	C1	C1	C1	C1	T0	C0	C2	C3	C1	C0	C1	C1	C2	C2	C2	C3	C6	C3	C1	C1	C2	C2	C2	C5	C3	C2	C1	C0	C0	C0	C1	C2	C2	C3	C4	C5
InterPro	Phob & TMHMM (1aa diff, helix ends inside Cytoplasmic Domain)																																																				
TMHMM	Transmembrane Out to In																																																				
Phobius	Transmembrane Out to In																																																				
TMPredP	In to Out (1496)																																																				
TMPredA	In to Out (1496)																																																				

Number	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416
Amino	P	A	S	T	S	V	S	V	L	A	S	D	R	G	A	G	A	L	G	F	V	G	T	A	G	K	E	S	V	G	Q	P	A	G	L	T	V	L	A	D	E	F	G	D	G	A	P	V	P	M	L	P
Porter3	C9	C8	C7	C7	C6	C5	C2	C0	C0	C2	C5	C6	C6	C6	C6	C5	C5	C5	C3	C4	C4	C4	C5	C7	C7	C6	C6	C6	C7	C7	C6	C4	C3	C0	E2	E2	C1	C5	C5	C4	C4	C8	C8	C9	C9	C8	C6	C6	C6	C8	C7
Porter8	C4	C2	C2	C2	C4	C3	C2	E0	E0	C1	C2	C0	T0	C0	C1	C1	C1	C0	C0	C1	C1	C0	C1	C2	C2	C0	C1	C2	C1	C1	C2	C4	C1	E0	E1	E2	E2	E0	C1	T0	T0	T1	C0	S0	C2	C6	C5	C2	C4	C5	C5	T1
InterPro	Cytoplasmic Domain																																																			
Annotation																			GFGT Motif																																			PPW Motif		
TMHMM	Inside																																																			
Phobius	Cytoplasmic																																																			
TMPredP	Outside																																																			
TMPredA	Outside																																																			

Number	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435
Amino	G	S	W	G	P	D	L	V	G	V	A	G	D	G	G	L	V	S	V
Porter3	C6	C6	C7	C8	C5	C4	C2	C4	C5	C4	C5	C6	C7	C8	C5	E2	E3	E2	C9
Porter8	T2	C1	C1	C4	T0	T0	C0	T1	C0	C2	C2	C0	T1	T0	C1	E3	E5	E3	C9
InterPro	Cytoplasmic Domain																		
Annotatic	PPW Motif																		
TMHMM	Inside																		
Phobius	Cytoplasmic																		
TMPredP	Outside																		
TMPredA	Outside																		

Figure 8: CS1 PPE46. prediction compilation of amino acid sequence, annotations of PPE motifs and areas of interest, Porter 5.0, InterPro, Phobius, and TMPred.

InterPro gave domain predictions that could not all be placed in Figures 7 and 8, so the domain and feature predictions it gave for H37Rv and CS1 are in Appendix 2. The VRI is on the C-terminal side of the middle, affecting the transmembrane domains. The N terminal domains are unaffected, being the PPE family domain between residues 8-164 (InterPro IPR000030, Pfam PF00823), the PPE superfamily domain at residues 5-178 (InterPro IPR038332, CATHGene3D G3DSA:1.20.1260.20), and PE/PPE dimer-like superfamily domain at residues 5-177 (InterPro SSF140459). The length of N-terminal unnamed PANTHER family domain (PTHR46766) is five residues longer in CS1 than H37Rv (residues 9-183 for H37Rv). The non-cytoplasmic domain (the part of the protein outside the cell) is larger in CS1, running from the N terminus to residue 227 in H37Rv but to residue 243 in CS1. TMHMM and Phobius are used by InterPro to predict the transmembrane regions, this region of transmembrane helices and cytosolic/non-

cytosolic loops are the same as seen in Figures 7 and 8. The C-terminal cytosolic region is the same size in both strains, at 326-434 for H37Rv and 327-435 to CS1.

To visualise the structural changes better, they were run through Protter and annotated using secondary structure prediction from Porter 5.0 and domain information from InterPro and PredictProtein (with extra information on folds provided by SCOP (63)). Figures 9 and 10 show those images of H37Rv and CS1 PPE46 respectively. Protter used InterPro to create the images, so agrees with Figures 7 and 8 above (except for TMPred as discussed earlier). It becomes evident here that the VRI created a difference in the sizes of the features more than in the secondary or tertiary folding.

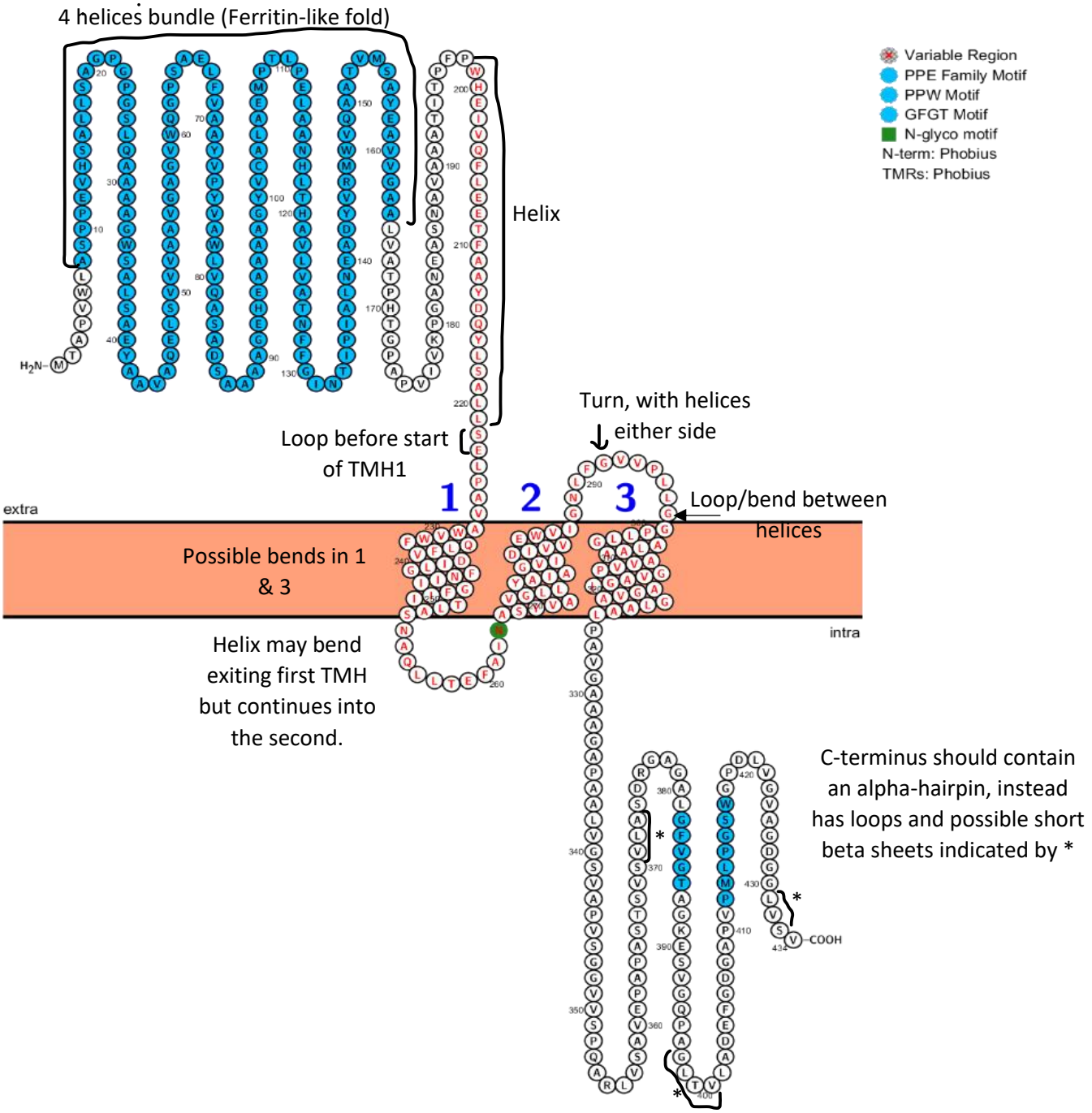

Figure 9: H37Rv PPE46 in Protter, annotated with information from SCOP and Porter 5.0.

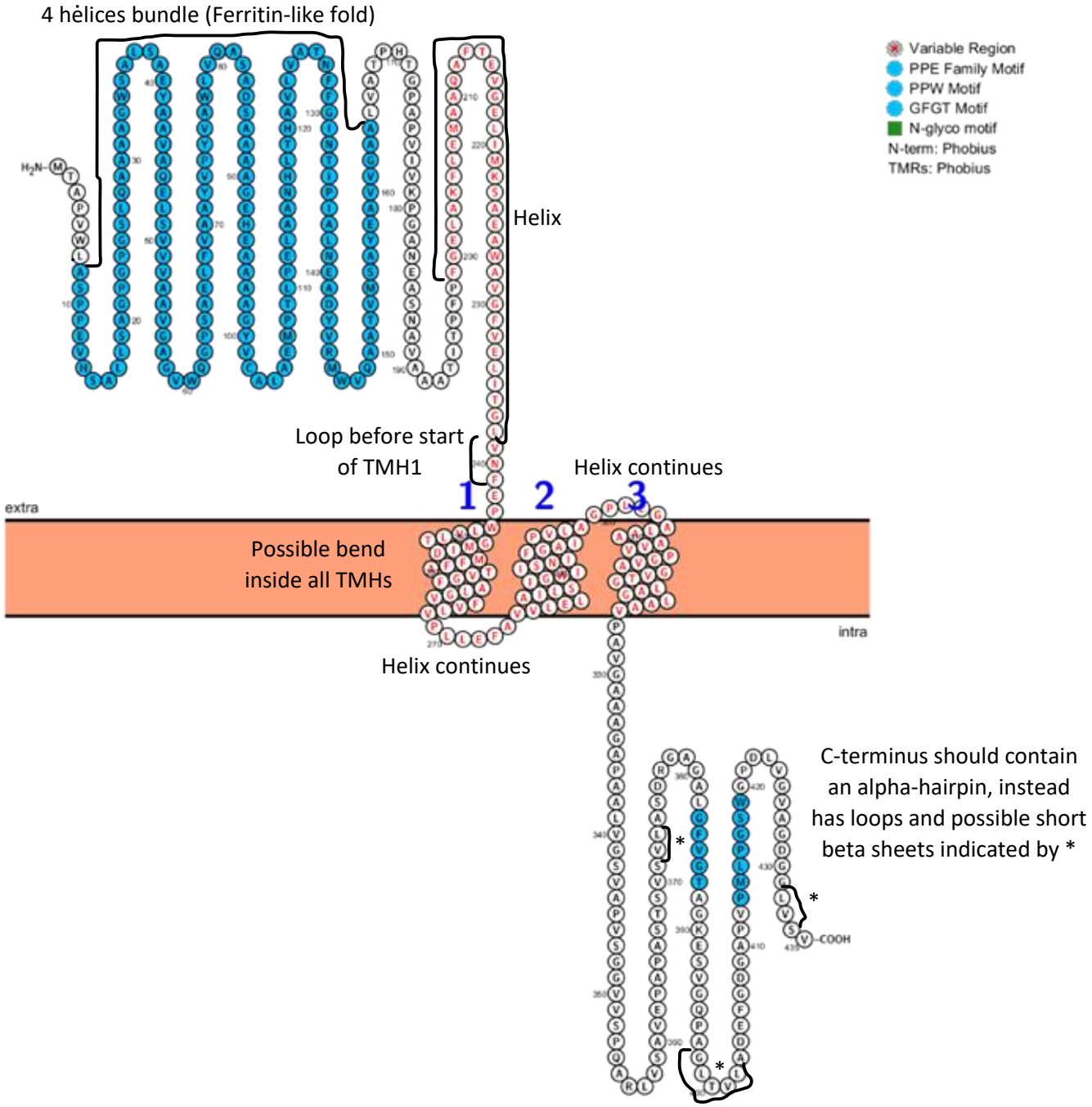

Figure 10: CS1 PPE46 in Protter, annotated with information from SCOP and Porter 5.0.

HeliQuest was used to investigate four helices in the variable region, highlighted in Figures 7 and 8. The aim was to see if the trans- and extra-membrane helices were amphipathic. Amphipathic transmembrane helices could indicate interactions to protect the hydrophilic sides, while amphipathic extra-membrane helices could indicate that the helix is half-buried in the membrane or has its own interactions. HeliQuest took all sequences and analysed them in an 18-residue sliding window (except for the outer helix looping between TMH 2 and 3, which was shorter than this). Residue ranges for helices analysis was decided using Porter 5.0 and InterPro. Table 6 shows a basic overview of what was found, with the image outputs available in Appendix 3.

Table 6: Summary of HeliQuest simulations of selected alpha-helical regions as predicted by Porter 5.0 and InterPro.

Helix	**H37Rv amino acid range**	**H37Rv results**	**CS1 amino acid range**	**CS1 results**
VRH1	199-221 [23]	More charged than CS1 version but smaller hydrophobic face	199-238 [40]	Charged and aromatic residues opposite large hydrophobic face

TMH1	224-253 [30]	1-3 hydrophobic faces which twist, overall hydrophobic 1 acidic residue	242-268 [27]	1 hydrophobic face, mostly hydrophobic with 2 acidic residues.
TMH2	264-287 [24]	2 acidic residues but overall hydrophobic, 1 hydrophobic face which gets larger	275-295 [21]	1 hydrophobic face, 1 acidic residue, overall hydrophobic
Helix bridging 2 & 3	288-298 [11]	Four hydrophobic residues together, 1 proline, 1 aromatic	290-304 [9]	2 proline residues and maximum of two neighbouring residues of any type
TMH3	301-323 [23]	Almost entirely hydrophobic	305-327 [23]	Mainly hydrophobic

In summary, the VRI region in PPE46 has affected the structure in more ways than adding an amino acid. While the basic shape (number of transmembrane helices, extra- or intra-cellular positioning of the N and C termini) did not change between CS1 and H37Rv in any one model, the length and hydrophobicity of features were altered and domains differed between the two. While the function and true structure of PPE46 in any strain is unknown, the predictions here give a good indication that the structure does change in a way that could change function in some way. As PPE46 is involved in virulence which requires interactions outside the cell, the shortening of the extra-cellular helix between the transmembrane helices and the lengthening of the extra-cellular N-terminal helix bundle could impact those interactions.

manB Predictions

The VRI in manB is found in the C-terminal end of the protein. To explore the effect of the VRI on protein structure, predictions for manB was formed using SwissModel, with support from the secondary structure and domain predictions of InterPro, PredictProtein, and Porter 5.0.

SwissModel provided fifty potential templates for each variant of manB. The best fitting template for H37Rv (shown in Figure 11) was another *Mtb* protein called GlmU (2QKX), involved in peptidoglycan and lipid A synthesis, as well as being essential for optimal growth (64). As the sequences are not the same, the fit to this template was not perfect (as seen by the red residues, which are poorly fitting, while blue sections fit well on the template), but within acceptable ranges, and considering this and the fact that manB proteins in other species were found to have the same basic shape, we can assume that H37Rv manB looks similar to this.

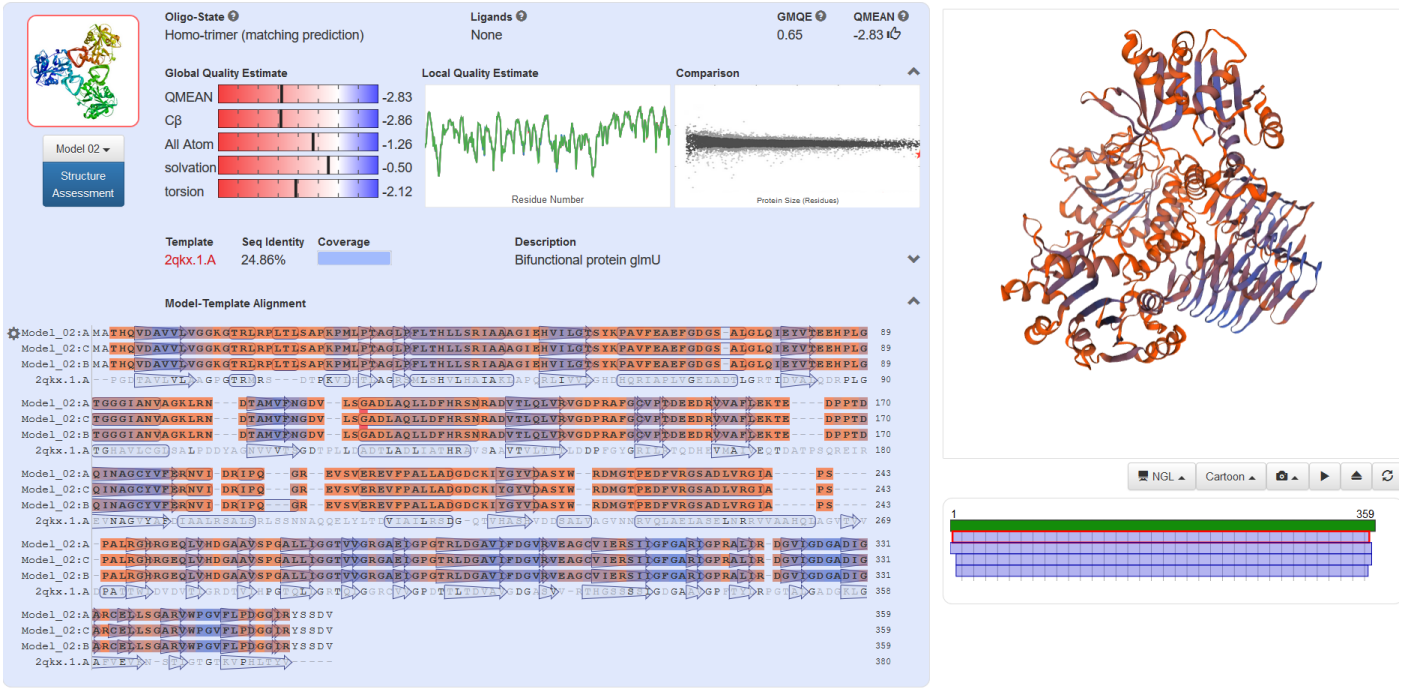

Figure 11: H37Rv manB best-fitting template in SwissModel – quote the templates in the text.

Out of the fifty templates for CS1, there were many different predictions including monomers, trimers, and tetramers. Even the best fits were not calculated to be acceptably certain, so two predictions are shown. In Figure

12, a second variant of glmU (3D98) (64) is used as the template as it was the best fit of the fifty templates given, and is a trimer. In Figure 13, the tetrameric ADP-glucose phosphorylase found in potatoes to make starch (65) is the template, as it was the best tetrameric fit. The tetrameric model is a slightly worse fit (most obviously seen by the proportion of red poorly-fitting residues), but it is unlikely that either represents the structure of CS1 manB. However, a brief look at similar proteins in a range of species found the same basic trimer shape with many small variations, so the author would tend to agree with the basic predicted shape of manB, and theorise that CS1 manB is a trimer. The increased difficulty in fitting the CS1 manB to a model compared with H37Rv shows there are structural differences caused by the VRI.

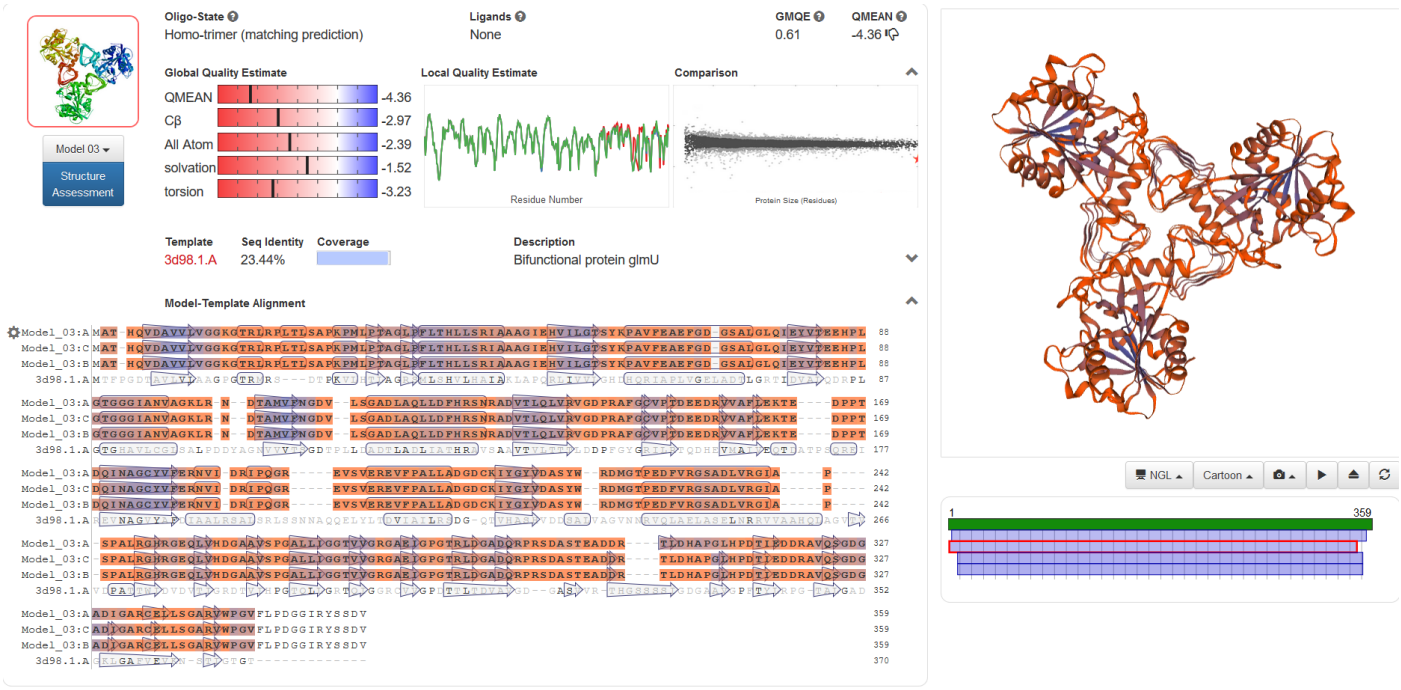

Figure 12: CS1 manB best-fitting trimer template in SwissModel.

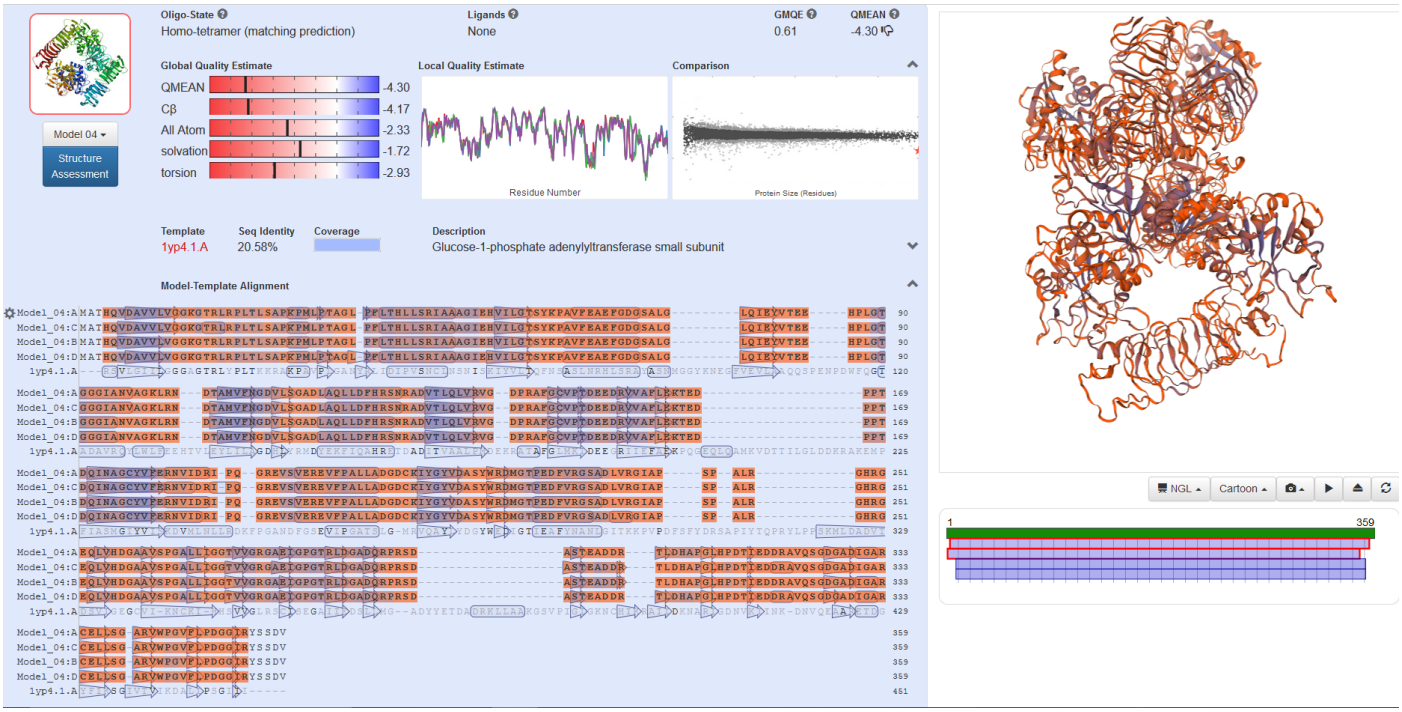

Figure 13: CS1 manB best-fitting tetramer template in SwissModel.

To look more closely at the changes in secondary folding created by the VRI, Porter 5.0 was used to compare the two. Appendix 4 shows a comparison between the SS3 and SS8 models for both versions of manB. The largest change is CS1 appears to have absent and shortened b-peated sheets compared with H37Rv. The rest of the differences are in certainty that the feature is present. There is some variation between SS3 and SS8 predictions for

both manB versions, more than seen in PPE46, but that could be because PPE was mainly a-helices while manB is more complex.

To investigate more clearly the domain and feature changes caused by the VRI, InterPro was used. The domain and feature predictions are shown for H37Rv and CS1 in Appendix 5. The VRI is near the C-terminal, affecting the transmembrane domains. Some N-terminal domains are unaffected, such as the NTP transferase family domain at residues 2-235 (InterPro IPR005835, Pfam PF00483), the CATHgene3D superfamily at residues 2-240 (G3DSA:3.90.550.10), and the NTP transferase catalytic domain at residues 8-224 (InterPro cd04181). The length of the longest N-terminal domains are shorter for the CS1 version, being nucleotide diphosphosugar transferase superfamily (IPR029044 H37Rv residues 2-308 and CS1 residues 2-241, SSF53448 H37Rv residues 8-301 and CS1 residues 7-241), and PANTHER sugar-1-phosphate guanyl transferase domain (PTHR22572, H37Rv 7-332 and CS1 7-289). The C-terminal hexapeptide repeat region is longer in H37Rv as it has two repeats (InterPro IPR001451, Pfam PF00132, H37Rv residues 255-335 and CS1 residues 255-286). The CATHgene3D hexapeptide superfamily region at residues 245-354 is one residue shorter in CS1 (G3DSA:2.160.10.10). CS1 has some N-terminal domains not present in H27Rv, being a trimeric lpxA-like enzyme superfamily domain at residues 252-353 (IPR011004, SSF51161), and two overlapping disorder regions at residues 278-324 and 288-324. The active, substrate-binding, and Mg^{2+} binding sites are unaffected, which is expected as they are in the N-terminal which was not affected by the VRI. The presence of the trimerization and disordered regions are interesting, as this could explain the difficulty SwissModel

In summary, manB was affected by the VRI region, although in what exact way is not clear. The active and substrate-binding sites are changed slightly, folding is altered, domain sizes changed and new domains added. This enzyme has been characterised in other species, and H37Rv was matched by SwissModel to a suitable (although not perfect) model, but this could not be done for CS1 which indicates some structural changes. Future research focussing on characterising *Mtb* proteins like manB and those affected by VRIs would be worthwhile, particularly as it would elucidate the changes predicted here. For manB, a structural change could affect activity, which could explain the high transmission rate of CS1.

Conclusion

In conclusion, this project investigated the genetic anomalies present in New Zealand *Mtb* strain CS1 as compared with the reference strain, H37Rv. CS1 was chosen as it is an unusually contagious New Zealand strain which has not been well-characterised. Several different types of anomalies were found which could explain the virulence of CS1 or lead us to a better understanding of this and other strains in general, including the discovery of potential annotation errors that highlight the need for more research. One type of anomaly was found with high densities of insertions, deletions, and base changes in a region of a gene. That type of anomaly was dubbed a VRI and investigated further. Of the genes affected by VRIs, most were encoding PE/PPE proteins or virulence factors, with only a few known to be polymorphic and many known to be potential drug targets. The high transmission rates of CS1 could be explained by mutations changing virulence factor structure and expression, making the presence of VRIs a good lead for learning about and combatting this strain. The VRI-affected genes were themselves of interest, due to their involvement in virulence and vital cell processes. Two proteins affected by VRIs and involved in virulence, PPE46 and manB, were selected to investigate the structural impact. While the software predictions may not be accurate reflections of the real structures, they confirmed strongly that the VRIs were affecting protein structure, which could affect function too. The predictions are also valuable as these two proteins, and many of the others affected by VRIs, have not been characterised, and this would be a valuable focus of future study and could lead to drugs targeting these virulence and survival proteins. There is much to be learned by looking at the effects of all the anomalies found in this paper, all of which could be useful for improving health outcomes for *Mtb* patients both in New Zealand and with other strains around the world.

References

1. WHO. Global tuberculosis report 2019. Geneva: World Health Organization; 2019.
2. WHO. New Zealand Tuberculosis Profile 2018 (web archive, accessed 10th February 2020) 2018 [Available from: https://web.archive.org/web/20200209214738/https://extranet.who.int/sree/Reports?op=Replet&name=%2FWHO_HQ_Reports%2FG2%2FPROD%2FEXT%2FTBCountryProfile&ISO2=NZ&LAN=EN&outtype=html

3. WHO. BCG vaccines: WHO position paper – February 2018. Weekly Epidemiological Record. 2018;93(8):73-96.
4. Roy A, Eisenhut M, Harris RJ, Rodrigues LC, Sridhar S, Habermann S, et al. Effect of BCG vaccination against *Mycobacterium tuberculosis* infection in children: systematic review and meta-analysis. BMJ : British Medical Journal. 2014;349:g4643.
5. WHO. Guidelines for treatment of drug-susceptible tuberculosis and patient care, 2017 update. Geneva: World Health Organization; 2017.
6. WHO. Treatment of Tuberculosis: Guidelines. 4th edition. Geneva: World Health Organization; 2010.
7. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. *Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. Nature Genetics. 2016;48(12):1535-43.
8. Littleton J, Park J, Thornley C, Anderson A, Lawrence J. Migrants and tuberculosis: analysing epidemiological data with ethnography. Australian and New Zealand Journal of Public Health. 2008;32(2):142-9.
9. Park J, Littleton J, Chambers A, Chambers K. Whakapapa in anthropological research on tuberculosis in the Pacific. SITES: New Series 2011;8(2):6-31.
10. ESR. Tuberculosis in New Zealand Annual Report 2016. Porirua: The Institute of Environmental Science and Research Ltd (ESR). 2019.
11. Mulholland CV, Shockey AC, Aung HL, Cursons RT, O'Toole RF, Gautam SS, et al. Dispersal of *Mycobacterium tuberculosis* Driven by Historical European Trade in the South Pacific. Frontiers in Microbiology. 2019;10(2778).
12. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, et al. Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. Nature Genetics. 2015;47(3):242-9.
13. De Zoysa R, Shoemack P, Vaughan R, Vaughan A. A prolonged outbreak of tuberculosis in the North Island. New Zealand Public Health Report. 2001;8(1).
14. Gautam SS, Mac Aogáin M, Bower JE, Basu I, O'Toole RF. Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of *Mycobacterium tuberculosis*. Infectious Diseases. 2017;49(9):680-8.
15. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one. 2010;5(6):e11147-e.
16. Geneious Prime 2020.0.3.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.
18. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. Nucleic acids research. 2009;37(Database issue):D211-D5.
19. Torrisi M, Kaleel M, Pollastri G. Deeper Profiles and Cascaded Recurrent and Convolutional Neural Networks for state-of-the-art Protein Secondary Structure Prediction. Scientific Reports. 2019;9(1):12374.
20. Gautier R, Douguet D, Antonny B, Drin G. HELIQUEST: a web server to screen sequences with specific α-helical properties. Bioinformatics. 2008;24(18):2101-2.
21. Hofmann K, Stoffel W. TMbase - A database of membrane spanning proteins segments. Biological Chemistry Hoppe-Seyler. 1993;374.
22. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic acids research. 2007;35(Web Server issue):W429-W32.
23. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Journal of Molecular Biology. 2001;305(3):567-80.
24. Omasits U, Ahrens CH, Müller S, Wollscheid B. Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics. 2013;30(6):884-6.
25. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research. 2018;46(W1):W296-W303.
26. Coscolla M, Gagneux S. Consequences of genomic diversity in **Mycobacterium tuberculosis**. Seminars in Immunology. 2014;26(6):431-44.
27. Stucki D, Gagneux S. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. Tuberculosis (Edinb). 2013;93(1):30-9.
28. Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, Harris SR, et al. The Genome of *Mycobacterium Africanum* West African 2 Reveals a Lineage-Specific Locus and Genome Erosion Common to the *M. tuberculosis* Complex. PLOS Neglected Tropical Diseases. 2012;6(2):e1552.

29. Beggs ML, Eisenach KD, Cave MD. Mapping of IS6110 Insertion Sites in Two Epidemic Strains of *Mycobacterium tuberculosis*. Journal of Clinical Microbiology. 2000;38(8):2923-8.
30. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research. 2018;47(D1):D506-D15.
31. Chopra T, Banerjee S, Gupta S, Yadav G, Anand S, Surolia A, et al. Novel Intermolecular Iterative Mechanism for Biosynthesis of Mycoketide Catalyzed by a Bimodular Polyketide Synthase. PLoS biology. 2008;6:e163.
32. Kruh NA, Troudt J, Izzo A, Prenni J, Dobos KM. Portrait of a Pathogen: *The Mycobacterium tuberculosis* Proteome In Vivo. PLOS ONE. 2010;5(11):e13938.
33. NCBI. NCBI Gene=888457 [accessed 25th January 2020]. Available from: https://pubchem.ncbi.nlm.nih.gov/gene/rplR/Mycobacterium_tuberculosis_H37Rv.
34. Kumar A, Bose M, Brahmachari V. Analysis of Expression Profile of Mammalian Cell Entry (*mce*) Operons of *Mycobacterium tuberculosis*. Infection and Immunity. 2003;71(10):6083-7.
35. Brzostek A, Pawelczyk J, Rumijowska-Galewicz A, Dziadek B, Dziadek J. *Mycobacterium tuberculosis* Is Able To Accumulate and Utilize Cholesterol. Journal of Bacteriology. 2009;191(21):6584-91.
36. Prakash P, Pathak N, Hasnain SE. pheA (Rv3838c) of *Mycobacterium tuberculosis* Encodes an Allosterically Regulated Monofunctional Prephenate Dehydratase That Requires Both Catalytic and Regulatory Domains for Optimum Activity. Journal of Biological Chemistry. 2005;280(21):20666-71.
37. McEvoy CR, van Helden PD, Warren RM, van Pittius NCG. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. BMC Evolutionary Biology. 2009;9(1):237.
38. Daleke MH, Cascioferro A, de Punder K, Ummels R, Abdallah AM, van der Wel N, et al. Conserved Pro-Glu (PE) and Pro-Pro-Glu (PPE) protein domains target LipY lipases of pathogenic mycobacteria to the cell surface via the ESX-5 pathway. J Biol Chem. 2011;286(21):19024-34.
39. Gey van Pittius NC, Sampson SL, Lee H, Kim Y, van Helden PD, Warren RM. 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. 2006;6(1):95.
40. Fishbein S, van Wyk N, Warren RM, Sampson SL. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. Molecular Microbiology. 2015;96(5):901-16.
41. McEvoy CRE, Cloete R, Müller B, Schürch AC, van Helden PD, Gagneux S, et al. Comparative Analysis of *Mycobacterium tuberculosis* pe and ppe Genes Reveals High Sequence Variation and an Apparent Absence of Selective Constraints. PLOS ONE. 2012;7(4):e30593.
42. Raman K, Yeturu K, Chandra N. targetTB: A target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. BMC Systems Biology. 2008;2(1):109.
43. McCarthy TR, Torrelles JB, MacFarlane AS, Katawczik M, Kutzbach B, DesJardin LE, et al. Overexpression of *Mycobacterium tuberculosis* manB, a phosphomannomutase that increases phosphatidylinositol mannoside biosynthesis in *Mycobacterium smegmatis* and mycobacterial association with human macrophages. Molecular Microbiology. 2005;58(3):774-90.
44. Matsunaga I, Bhatt A, Young DC, Cheng T-Y, Eyles SJ, Besra GS, et al. *Mycobacterium tuberculosis* pks12 produces a novel polyketide presented by CD1c to T cells. J Exp Med. 2004;200(12):1559-69.
45. Sirakova TD, Dubey VS, Kim H-J, Cynamon MH, Kolattukudy PE. The largest open reading frame (pks12) in the *Mycobacterium tuberculosis* genome is involved in pathogenesis and dimycocerosyl phthiocerol synthesis. Infection and immunity. 2003;71(7):3794-801.
46. Kumar M, Khan FG, Sharma S, Kumar R, Faujdar J, Sharma R, et al. Identification of *Mycobacterium tuberculosis* genes preferentially expressed during human infection. Microbial Pathogenesis. 2011;50(1):31-8.
47. Mogila I, Kazlauskiene M, Valinskyte S, Tamulaitiene G, Tamulaitis G, Siksnys V. Genetic Dissection of the Type III-A CRISPR-Cas System Csm Complex Reveals Roles of Individual Subunits. Cell Reports. 2019;26(10):2753-65.e4.
48. Grüschow S, Athukoralage JS, Graham S, Hoogeboom T, White MF. Cyclic oligoadenylate signalling mediates *Mycobacterium tuberculosis* CRISPR defence. Nucleic Acids Research. 2019;47(17):9259-70.
49. Grüschow S, Athukoralage J, Hoogeboom T, White MF. The Type III CRISPR-Cas system of *Mycobacterium tuberculosis*. Access Microbiology. 2019;1(1A).
50. Ilin AI, Kulmanov ME, Korotetskiy IS, Islamov RA, Akhmetova GK, Lankina MV, et al. Genomic Insight into Mechanisms of Reversion of Antibiotic Resistance in Multidrug Resistant *Mycobacterium tuberculosis*

Induced by a Nanomolecular Iodine-Containing Complex FS-1. Frontiers in Cellular and Infection Microbiology. 2017;7(151).
51. Bisson GP, Mehaffy C, Broeckling C, Prenni J, Rifat D, Lun DS, et al. Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, rpoB mutant *Mycobacterium tuberculosis.* Journal of bacteriology. 2012;194(23):6441-52.
52. Pasricha R, Saini NK, Rathor N, Pathak R, Sinha R, Varma-Basil M, et al. The *Mycobacterium tuberculosis* recombinant LprN protein of mce4 operon induces Th-1 type response deleterious to protection in mice. Pathogens and Disease. 2014;72(3):188-96.
53. Becker K, Sander P. *Mycobacterium tuberculosis* lipoproteins in virulence and immunity – fighting with a double-edged sword. FEBS Letters. 2016;590(21):3800-19.
54. Nesbitt N, Yang X, Fontán P, Kolesnikova I, Smith I, Sampson N, et al. A Thiolase of *Mycobacterium tuberculosis* Is Required for Virulence and Production of Androstenedione and Androstadienedione from Cholesterol. Infection and immunity. 2009;78:275-82.
55. Bragin EY, Shtratnikova VY, Schelkunov MI, Dovbnya DV, Donova MV. Genome-wide response on phytosterol in 9-hydroxyandrostenedione-producing strain of *Mycobacterium* sp. VKM Ac-1817D. BMC Biotechnology. 2019;19(1):39.
56. Bromke M. Amino Acid Biosynthesis Pathways in Diatoms. Metabolites. 2013;3:294-311.
57. Yellaboina S, Ranjan S, Vindal V, Ranjan A. Comparative analysis of iron regulated genes in mycobacteria. FEBS Letters. 2006;580(11):2567-76.
58. Joung SM, Ryoo S. BCG vaccine in Korea. Clin Exp Vaccine Res. 2013;2(2):83-91.
59. Saelens J, Lau D, Moller A, Xet-Mull A, Medina N, Guzmán B, et al. Annotated Genome Sequences of 16 Lineage 4 *Mycobacterium tuberculosis* Strains from Guatemala. Genome Announcements. 2018;6:e00024-18.
60. Wong KC, Leong WM, Law HKW, Ip KF, Lam JTH, Yuen KY, et al. Molecular Characterization of Clinical Isolates of *Mycobacterium tuberculosis* and Their Association with Phenotypic Virulence in Human Macrophages. Clinical and Vaccine Immunology. 2007;14(10):1279-84.
61. Brosch R, Philipp WJ, Stavropoulos E, Colston MJ, Cole ST, Gordon SV. Genomic Analysis Reveals Variation between *Mycobacterium tuberculosis* H37Rv and the Attenuated *M. tuberculosis* H37Ra Strain. Infection and Immunity. 1999;67(11):5768-74.
62. Zheng H, Lu L, Wang B, Pu S, Zhang X, Zhu G, et al. Genetic Basis of Virulence Attenuation Revealed by Comparative Genomic Analysis of *Mycobacterium tuberculosis* Strain H37Ra versus H37Rv. PLOS ONE. 2008;3(6):e2375.
63. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research. 2013;42(D1):D304-D9.
64. Zhang Z, Bulloch EMM, Bunker RD, Baker EN, Squire CJ. Structure and function of GlmU from *Mycobacterium tuberculosis*. Acta Crystallographica Section D. 2009;65(3):275-83.
65. Jin X, Ballicora MA, Preiss J, Geiger JH. Crystal structure of potato tuber ADP-glucose pyrophosphorylase. EMBO J. 2005;24(4):694-704.