Verifying clustered mutation regions inside the newly-completed genome of a clinically-significant New Zealand *Mycobacterium tuberculosis* strain

A thesis
submitted in partial fulfilment
of the requirements for the degree
of
Masters of Science (Research) in Biological Sciences/Cellular & Molecular Biology
at
The University of Waikato
by

Mackenzie Olivia Steele

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2021

Abstract

In 2019, active infections of *Mycobacterium tuberculosis* (*Mtb*) affected 10 million people globally, with 1.4 million deaths, and an additional 1.7-1.9 billion people infected without symptomatic disease (latent *Mtb*). While *Mtb* is often associated with Asia and Africa, New Zealand had 360 *Mtb* infections and 21 deaths in 2019, and roughly 20% of cases were caused by three strains unique to New Zealand. One is CS1 (formerly Rangipo), which is known to be more transmissible than its relatives and has been causing outbreaks in the Waikato region for 30 years. As CS1 has genetic features linked to virulence in unrelated *Mtb* strains, having a complete and accurate record of genetic features is important, in particular those with the potential to change protein structure (Variable Regions Inside Proteins, VRIPs). Resolving VRIPs would simultaneously remove sequencing mistakes and increase knowledge about *Mtb* virulence. The aim of this project was to investigate the genetic cause of improved transmissibility in CS1 and produce a closed genome.

Sequencing data from Illumina and PacBio reads was processed, assembled, and checked for quality discrepancies which would require more sequencing to resolve. VRIPs were mapped alongside regions of low-quality data, low data-coverage, and repeat regions. Characteristics of the CS1 genome, such as repeat levels and size, were found by comparing with the reference strain H37Rv.

The read data and assembly were found to be good quality, and the genome is of high enough quality to be considered closed. As expected, some regions of low coverage and poor quality were discovered inside large repeat regions, and more work on these would be valuable. 13 of the 32 VRIPs were verified, all linked with virulence factors. 5 verified VRIPs caused genes to split into domains, accumulating mutations in the non-coding region. Further work resolving the remaining 19 VRIPs is recommended.

Acknowledgements

I would firstly like to thank my supervisors, Professor Vic Arcus and Dr Adele Williamson. Their support and enthusiasm through this project has been invaluable and inspiring.

To Daniel Schipper, Cheryl Ward, Andrea Haines, and Rose Swears, please know how much I appreciate you! Without Daniel's computational expertise and Rubber Duck ability, I would not have nearly as much data. Rose and Andrea provided invaluable help with making all the information comprehensible – I apologise to the former that this wasn't rocket science, but something far worse. And of course, I could not have finished without the help of the best librarian.

I would also like to thank C2 for all the small things. I would especially like to thank our lab mum, Judith, for all the gluten free baking and comfort. Within the Arcus group and on the other side of the corridor, from troubleshooting to genuinely caring, thank you.

Lastly, thank you to my family, chosen and blood, who listened to an unimaginable amount of babbling without complaint. Your love, practical assistance (what's dinner?), and care through my studies has been amazing. A special shout out to my sister for having listened to so much of my thesis that she can now explain it to friends who ask her what I've been doing. I deserved a Quanitferrum Gold shake many times, thank you for putting up with me! To my Scouting, church, and Latin families, gratiatis ago.

Table of Contents

List of Figures

List of Tables

List of Abbreviations & terms

CS1

SNP

L4.4.1.1

PacBio

Mtb

In silico

H37Rv

VRIP

NCBI

BBDuk

BBNorm

Kmers

SD

PHRED

Mpileup

Ir

BLAST & BLASTx

PPE

CRISPR

JSON

IGV

BRIG

Chapter 1

Introduction

1.1 Tuberculosis

Tuberculosis is a human disease caused by the bacterium *Mycobacterium tuberculosis* (*Mtb*). *Mtb* is a pathogenic bacterium within the so-called *M. tuberculosis* Complex (MTC), which includes *Mycobacterium africanum* and *Mycobacterium bovis* [1]. Different species in the MTC can infect different animals, for example *M. bovis* is mainly found in cows, however it has been found in humans and has the same disease mechanism. *Mtb* infiltrates macrophage immune cells in the lungs and is spread by coughing, although it can be found most anywhere in the body [2-4]. Typical symptoms of active disease are coughing, fever, fatigue, and weight loss [5]. Antibiotic resistant tuberculosis and comorbidity with HIV cause higher fatalities and therefore are priorities [2]. 30 countries have notable issues with Mtb outbreaks, all of them developing nations, but this disease is a global concern [2].

In 2019, active *Mtb* infections affected 10 million people globally, with 1.4 million deaths, and an additional 1.7-1.9 billion more who were infected without symptomatic disease (latent *Mtb*) [2; 6]. 2019 saw 360 *Mtb* infections in New Zealand, and 21 deaths [7]. The infection rate in New Zealand of 7.5 per 100,000 and death rate between 4-6% has held relatively steadily over the last 30 years [7]; [8]. Most New Zealand *Mtb* patients were aged between 25 and 34 years, however 3% were children [7]. In 2016, New Zealand-contracted 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 [8]. Tuberculosis disproportionately affects deprived communities. This is because factors which affect deprived communities, such as poverty, smoking and alcohol use, over-crowding, poor access to health care,

food insecurity and malnutrition, and diabetes, increase susceptibility to tuberculosis [9]. These inequalities affect Māori and Pasifika populations disproportionately more than other groups in New Zealand, and there is a lack of research on strains prevalent in these two populations and their impact [9]. Globally, *Mtb* has been the leading cause of death by infectious agent since 2007, above HIV/AIDS [10].

For prevention, currently only the Bacillus Calmette–Guérin (BCG) vaccine is approved, and its effectiveness varies by population and age group [2]. In children, it provides roughly a 20% reduction in infection and a 71% reduction in infections becoming active [11], although these results vary significantly between regions and populations [10]. This reduction is useful as latent infections can become active when the immune system is weakened, often decades after exposure, which occurs in 5-10% of latent infections, with roughly a quarter of the world's population affected by latent *Mtb* [2; 3; 6]. The BCG vaccine is more effective protection against severe forms of tuberculosis disease, which requires hospitalisation and can be life-threatening, than all latent or mild symptomatic *Mtb* infections [10]. This protection can last 15 years and longer, depending on the patient's age when vaccinated and environmental factors [10]. Antibiotic treatment for tuberculosis is long (6-18 months) with a high risk of drug resistance and relapse [12], and both treatment time and cost are even greater for drug-resistant strains [2]. Treatment of active non-drug resistant *Mtb* involves six months of a multi-drug rifampicin-based regimen [13], with a 3.1% chance of relapse and 50-94% risk of developing multi-drug resistance [12]. This treatment is also expensive, at USD860 per patient [2]. For these reasons, the discovery of new drugs and vaccines are a priority for the World Health Organisation [2]. In order to reach this goal, it is important to find drug targets and learn more about *Mtb* and its strains [2].

1.2 Tuberculosis phylogeny

There are many phenotypically and genetically distinct *Mtb* strains around the world, often linked with a population and geographic location, and these can be split into 7 related lineages, as shown in Figure 1 [1; 14]. The most common and widespread lineage is L4, while L7 is known to grow slowly, and L2 is generally more virulent [1; 14]. Transmission rate, ratio of latent to active infections, and disease severity (together referred to as virulence) differ between *Mtb* strains, clades, and lineages [1; 14].

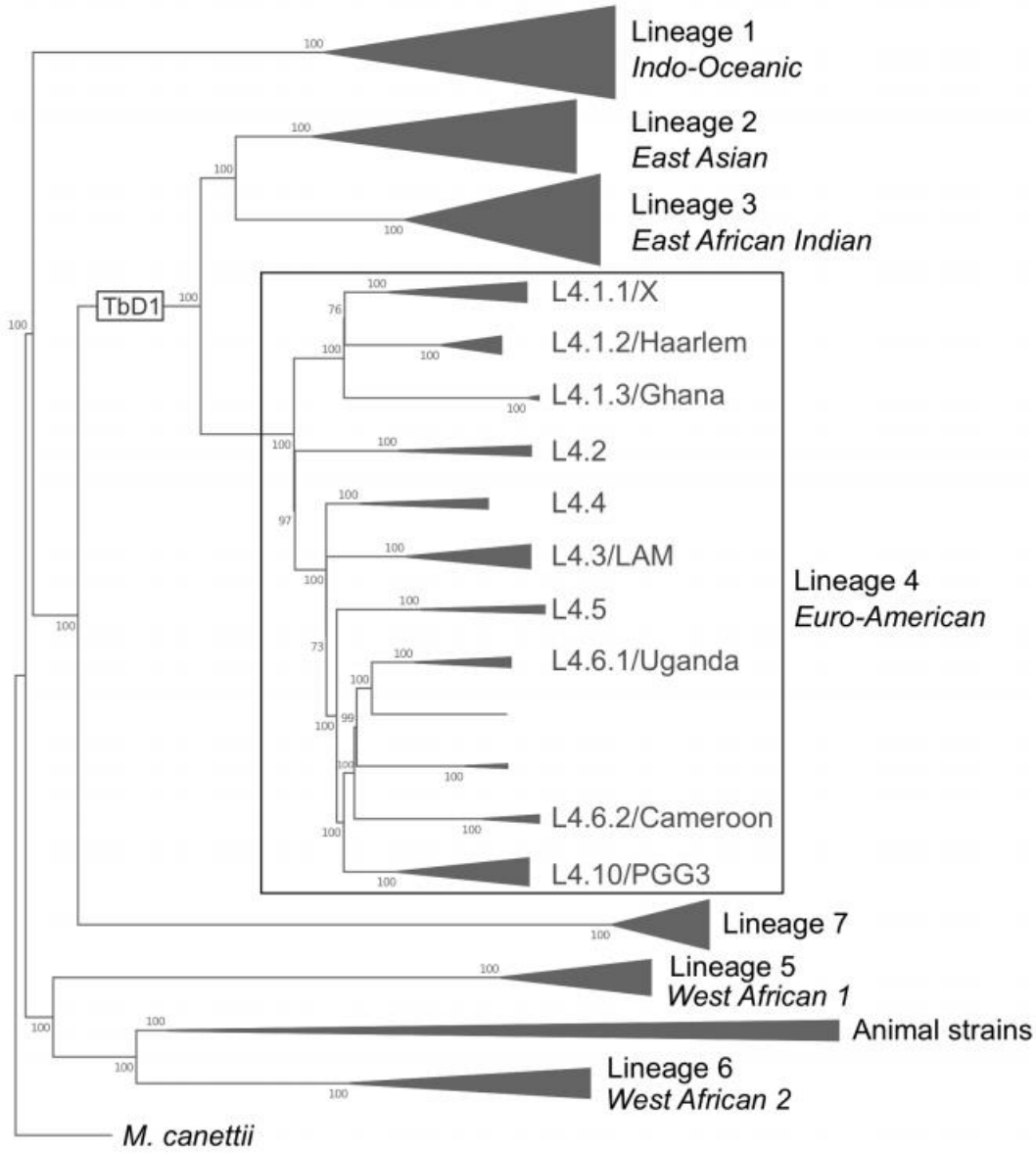

Figure 1: A figure showing the relatedness of lineages inside the MTC. Reference strain H37Rv is in Lineage 4.10, the New Zealand strain CS1 is found in Lineage 4.4, while Beijing is found in Lineage 2. Taken from Figure 1.2 in Mulholland, 2019 [14].

While *Mtb* is associated with residents in and immigrants from Africa, Asia, and the Pacific [15]; [16], New Zealand has three main strain clusters causing significant numbers of cases: Colonial Strain 1 (CS1, formerly known as Rangipo), CS2 (formerly known as Otara), and Southern Cross [1; 17]. CS1 and CS2 together make up 43% of New Zealand-born cases [14] and both belong to *Mtb* clade L4.4.1.1, which also includes the Canadian strains due to a common historical origin [1; 14; 17]. A feature of this clade is the presence of the so-called "Beijing deletion" (RD152 in L2 & $DS6^{Quebec}$ in the Canadian strains). As shown in Figure 2, this is a region present in the reference *Mtb* strain, H37Rv, but not in Beijing-family strains. This is of interest as the Beijing strains in L2 are known for high virulence [17; 18], and are not closely related to L4.4.1.1 strains, which means the deletions evolved independently and could carry fitness advantages [17].

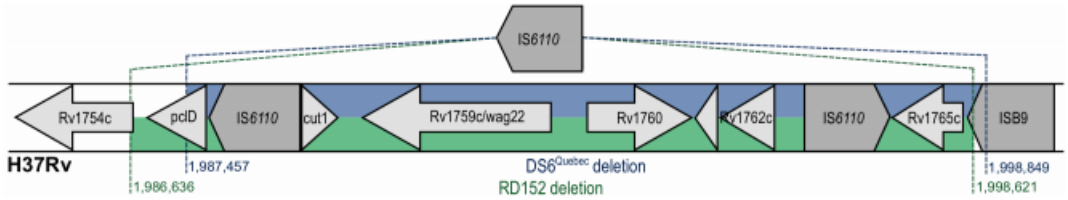

Figure 2: A schematic of the RD152 (shown in green) and $DS6^{Quebec}$ (shown in blue) deletions. Taken from Figure 2.7 in Mulholland, 2019 [14].

1.3 CS1

CS1 is the largest and most widespread New Zealand cluster, accounting for about a quarter of cases in Māori patients and still causing semi-regular outbreaks four decades after emerging [1; 17]. It likely arose from a 1980's clonal expansion, combined with social changes and urbanisation which affected Māori populations in particular [17]. CS1 has been responsible for some unusual prolonged outbreaks [19], and is noted for its high rate of active cases, being 20% of all traced contacts [20], which is not surprising considering certain deletions and other genetic features carried by CS1 are shared by strains known for their virulence [1]. It has been noted that CS1 has extra virulence

genes and genetic features (including Single Nucleotide Polymorphisms or SNPs) not found in H37Rv, with which it otherwise shares high similarity [20]. Tuberculosis strains show high genetic diversity with a multitude of SNPs [21] which can affect protein structure and gene expression [22]. CS1 has 247 SNPs compared to other sequenced Mtb strains, of which 236 are not found in H37Rv and 26 are not found in any other strain [14]. CS1 and other strains have genes which have been moved and duplicated, including IS6110 insertion sites which are linked with increased growth [23]. These extra genes and genetic features could provide an advantage for CS1 and explain the high transmission rate, in particular some of the genetic features could change the structure of existing proteins. Previous work by Mulholland using short-read Illumina sequences showed two cases where SNPs affected the biochemical properties of the encoded proteins [1].

While work focussing on CS1's SNPs and evolutionary heritage has been carried out and a draft genome is available, a completed "closed" genome has not yet been created [1; 17; 20]. A closed genome would allow further work to be done to investigate the virulence of CS1 and other L4.4.1.1 strains, confirm the work which has already been done around the SNPs and genetic features CS1 carries, and potentially inform treatment and prevention measures. A closed genome has the advantage of being accurate and complete, so strain identification is more accurate and drug targets can be found more efficiently. Instead of screening many potential compounds on a live microbe that is notoriously hard to grow and hazardous to human health, protein information from the closed genome can be utilised to develop compounds that target particular proteins *in silico*. H37Rv has a closed genome available, however most strains of *Mtb* have only incomplete or no sequences available. For CS1, there is a complete PacBio genome which has had limited detailed curation (Genbank, accession NZ_CP044345.1, henceforth referred to as the PacBio "draft") and short Illumina reads which have been used to polish the raw PacBio sequence [Aung,

Mulholland, personal communication]. Illumina sequencing produces short high fidelity reads. PacBio produces long reads with a higher error rate, but is less susceptible to errors caused by repeat regions and is easier to assemble into a contiguous draft than Illumina reads. The use of both sequencing technologies has the advantage of being complimentary, creating a genome which is accurate and less affected by the many repeat regions in a *Mtb* genome. However, further work is required to complete and close CS1 genome, due to the discovery of regions of dense genomic discrepancies between CS1 and H37Rv and poor coverage of the Illumina reads when mapped to the PacBio draft, which could indicate the need for further polishing to close the genome [24]. This is not uncommon for complete sequences created by automated methods with limited curation [25].

The dense genomic discrepancies between CS1 and H37Rv take many forms, from large insertions and deletions, to Variable Regions Inside Proteins (VRIPs) [24]. VRIPs are defined as regions where clusters of base changes and small insertion-deletions are found inside the annotation of a protein-coding gene. VRIPs can cause genes to be split into multiple annotations due to early stop codons, as well as causing potential changes to protein structure. VRIPs could be the result of errors in sequencing, assembly, or post-assembly polishing, thus the existence of each as a true feature of CS1 must be verified.

1.4 Objectives

The aim of this project was to use bioinformatic techniques to examine and resolve potentially significant genetic discrepancies found in CS1 compared with other strains of *Mtb,* such as VRIPs. Resolving these discrepancies would show the location and extent of poor sequencing in the draft genome, in order that the genome may be closed. Resolving discrepancies would also identify which of those discrepancies are genuine features of CS1, and which are artefacts of sequencing errors. Confirmed discrepancies could

provide valuable new targets for drugs and vaccines against the highly disease-causing CS1, and could also give us insight into the characteristics of this strain compared with H37Rv, without needing to manipulate it in a laboratory environment. Thus, this project investigated the data quality of the draft genome, and attempted to use previously-identified genetic discrepancies in the CS1 draft to close the genome and elucidate potential genetic mechanisms behind CS1’s high active infection rate.

Chapter 2

Methods

2.1 Genome Investigation

2.1.1 Data sources

The draft genome used in the present work was created by other projects through Single Molecule Real-Time sequencing (PacBio) [1; 26]. The resulting sequence had been polished using Illumina reads from Mulholland [Aung, Mulholland, personal communication] before uploading to NCBI. The draft genome was downloaded from NCBI on 10/04/2020. The Illumina reads were also used in this project.

2.1.2 Geneious Read Alignment

In line with a previous project conducted in 2020 [24], the Mauve alignment [27] tool was used in Geneious version 2020.0.3 [28] to map the Illumina reads to both the CS1 draft and H37Rv. The reads were trimmed with BBDuk [29] using the default settings in Geneious, then error corrected and normalised using defaults in BBNorm [29]. The Dedupe tool [29] was used despite the reads being paired, with kmers of 30 and 31 as per the default settings. The reads were then aligned to H37Rv using Geneious's assembly tool on recommended settings. This assembly was then manually inspected. Low and high coverage (in comparison with the mean) regions were searched for and annotated. The VRIPs previously found were mapped against 2 S.D. high and 1 S.D. low coverage zones. Methods and tools to graph the mapping were investigated in the literature but none were found.

Both the raw and fully-processed Illumina reads were run through a non-Geneious tool, FastQC [30], to inspect and compare their quality. An existing assembly of the reads assembled to H37Rv was loaded into Geneious, however Geneious redid the assembly step. This assembly was compared to the Genious-created assembles to

the CS1 draft and H37Rv. Known mutations were searched for in all three assemblies in order to assess potential errors in assembly. A non-Geneious tool, Qualimap [31], was then used to compare the quality for the three assemblies. This process showed a different assembly tool was required.

2.1.3 RGAPepPipe

The assembly was run using the RGAPepPipe (available from https://github.com/pepperell-lab/RGAPepPipe), as had been done above in Geneious. The reads were processed and assembled to the CS1 draft and H37Rv to default settings with BWAmem used as the assembler for both assemblies. The H37Rv assembly was used as a control to check for errors in the assembly process. After the Illumina reads were processed and deduplicated by the pipeline, FastQC was run to check the processed-read quality. The CS1 PacBio draft was also quality-checked using FastQC. The RGAPepPipe assemblies to H37Rv and the CS1 draft were run by the pipeline through Qualimap to assess the quality of the assembly. The results from FastQC and Qualimap were compared to the Geneious assemblies. As the RGAPepPipe assemblies were found to be improved, these were used for the rest of this project.

2.2 Analysis Tools

2.2.1 Quality & Coverage

In order to ascertain confidence in the assembly and discrepancies found, three tools were used to analyse the coverage and quality of the CS1 draft-Illumina assembly.

The first of these, Qualimap, investigated the quality of the CS1 RGAPepPipe alignment itself. This has been discussed above in section 2.1.3.

The second tool was a script created by Daniel A. Schipper [32], which was used to investigate the correlation between discrepancies and poor quality regions. This was done by finding

areas with low quality defined by cumulative PHRED scores of the Illumina reads to any PacBio base position under thresholds of 300 and 600. PHRED scores communicate the certainty a base call is accurate, by using a logarithmic scale where 10 is 90% certainty while 60 is 99.9999% certainty [33]. The positions in the PacBio draft where the cumulative PHRED score was below each threshold were printed into .csv and .bed files for later use and visual analysis, and manually matched with known VRIP regions. All files were manually corrected from 0-indexed to 1-indexed before use. Three files for visual analysis were created for ease of visual analysis: Very Low Quality, under 300; Low Quality, under 600; and Overlapping Quality, where a region of under-300 quality occurred within a region of under-600 quality. The overlapping points for each quality threshold were removed from their respective lists and merged to form the Overlapping Quality list.

The third tool, SAMTools [34], was used for the mpileup function to provide coverage data for each base position in the assembly. The data was stored for later visual analysis.

2.2.2 Investigation of CS1 genomic features

Known SNPs identified by Mulholland [1] and genomic discrepancies from the summer project were found in both RGAPepPipe assemblies using the Integrative Genome Viewer [35]. Lists were compiled in Excel to show the affected genes and positions.

2.2.3 Repeat Analysis

Mtb repeat regions were investigated in the literature in order to check if repeat regions could be responsible for the genomic discrepancies and anomalies in quality and coverage.

Tools were investigated to investigate the locations of repeat regions. The NCBI tool WindowsMasker [36] was installed and run on a Macintosh Operating System. WindowsMasker was used for both H37Rv and CS1, and the output modified into .csv and .tdv files. The output was also examined in Excel to investigate repeat sizes

and compare the two strains. After investigating spread of repeat lengths, only repeats longer than 99 base pairs were included in an output .tdv file used for visual analysis. The two reasons for this cut off were that most repeats were shorter than the cut off, and secondly that the length of most Illumina reads (as seen from FastQC and Qualimap) meant they would be unaffected by a repeat region shorter than 100 bases. Slippage events and sequencing errors would be more likely for repeats longer than 99 bases.

To measure the levels of repetition in *Mtb* genomes, I_r [37] was run on Linux on the whole CS1 genome, and a windowed analysis was also carried out. The output was analysed in relation with other prokaryotes previously collected by Haubold and Weihe [37].

2.2.4 Outlier Repeat Region

The investigation of the WindowsMasker data showed an outlier repeat region of significant length in CS1 which did not occur in H37Rv. BioCyc [38] and Uniprot's BLAST function [39] were used to identify the gene the outlier repeat occurs inside. The region was manually inspected in the Integrative Genome Viewer to verify the repeat and its motif. The region was also manually inspected in Geneious to compare with H37Rv, checking for genomic discrepancies which were then listed. The repeat region was annotated as overlapping both a PPE gene and a CRISPR, so CRISPR-Cas Finder [40] was used to investigate CRISPR sites. No CRISPR regions were annotated in H37Rv's NCBI entry at all, thus the CRISPR-Cas Finder was used to check why this might be the case. It was found that the annotations of CRISPR regions in CS1 were inaccurate, and the region was visually inspected to identify it. CRISPR-Cas Finder results are included in the appendix as a JSON.

2.3 Visualising Results

2.3.1 Genome Viewing

In order to analyse the CS1 genome and the assemblies, visual inspection was required. Although Geneious is a viewer, this

function cannot be separated from its in-built assembly tools, so the Integrative Genome Viewer (IGV) was used to view the RGAPepPipe assemblies. This tool has a region-of-interest mapper and can read .bed files as generated earlier by the Schipper script. The region-of-interest mapper was used to find and view low quality regions from the custom-made Schipper script, as well as to manually inspect the CS1 genome and alignment.

2.3.2 Genome and Analysis Viewing

In order to investigate the cause and reliability of CS1 genomic discrepancies, the spatial relationships between repeats, low quality regions, and genomic discrepancies (such as VRIPs) were visually analysed. The BLAST Ring Image Generator (BRIG) [41] was found to be suitable to visually display these spatial relationships.

BRIG was then installed onto a Windows Operating System with the NCBI C++ toolkit [42], and BLAST [43] downloaded separately. The final image used genomes added as genbank files, while quality and anomaly information were tab delimited files added as annotations. Due to space and legibility, Very Low Quality and the Overlapping Quality files were combined into one ring, with four corrupted and unverifiable points removed (3696137-3693141, 2079564-2078078, 1963504-1962513, and 1465758-1465334). Coverage files were processed through BRIG's inbuilt graph creation tool, then added as a ring. The settings for the final image were defaults except slot spacing was set to x-small, shading turned off, and image size adjusted to 2500 by 2500 pixels.

It is of note that the coverage graph generated by BRIG's graph creation tool does not have a scale and appears to be baselined at an unknown value; neither factor was able to be user-controlled. The blue bars on the coverage graph are anomalies which could not be removed, although this was attempted by making the final image a composite of two graphs image-edited. IR results were separately graphed using Excel and not included in the BRIG figure.

In addition, Excel, Geneious, and IGV were used separately and in combination to attempt to visualise the coverage, quality, and genomic data for the outlier repeat. However, visualisation was not found to be achievable.

2.3.3 VRIP Investigation

The VRIP regions from the summer project were separated out after analysing for confidence (see above). Regions not found to overlap with low quality and repeat regions were noted and the annotated gene name and function found in IGV. Those genes with names and known functions were researched in the literature, and the most promising candidates for virulence effect were investigated further.

Images of the VRIP regions were created in Paint using Geneious data showing the coverage information gathered earlier, and IGV data showing some of the reads in relation to each other and the PacBio draft.

The selected CS1 genes (and gene fragments in the case of genes which had multiple annotations) containing VRIPs were run through a BLASTx [44] search of the *Mtb* cluster under default settings which was capped at 100 results. The hit results were downloaded as a csv, and the query-anchored alignment saved in text files. The alignment was rerun in Clustal Omega [45] and filtered down for legibility, producing phylogenetic trees as well as alignments coloured by amino acid, both of which were saved as screenshots and edited for size in PixlrE. This was redone for comparing H37Rv with CS1 for each shortlisted gene.

Chapter 3

Results

3.1 Genome Quality Analysis

3.1.1 VRIP mapping

To view VRIP regions, composite images from Geneious and IGV were created. Figure 3 shows one of the VRIP regions by comparing H37Rv, the PacBio CS1 draft, and the CS1 Illumina reads. Typical features seen in a VRIP include lower coverage regions near read-stacks, reads with mutations, and insertion-deletion differences between H37Rv and CS1. It is of note that Geneious used its own assembler while IGV uses the RGAPepPipe assembly, however this does not greatly affect the VRI regions (the exact number of stacked reads may vary slightly).

-

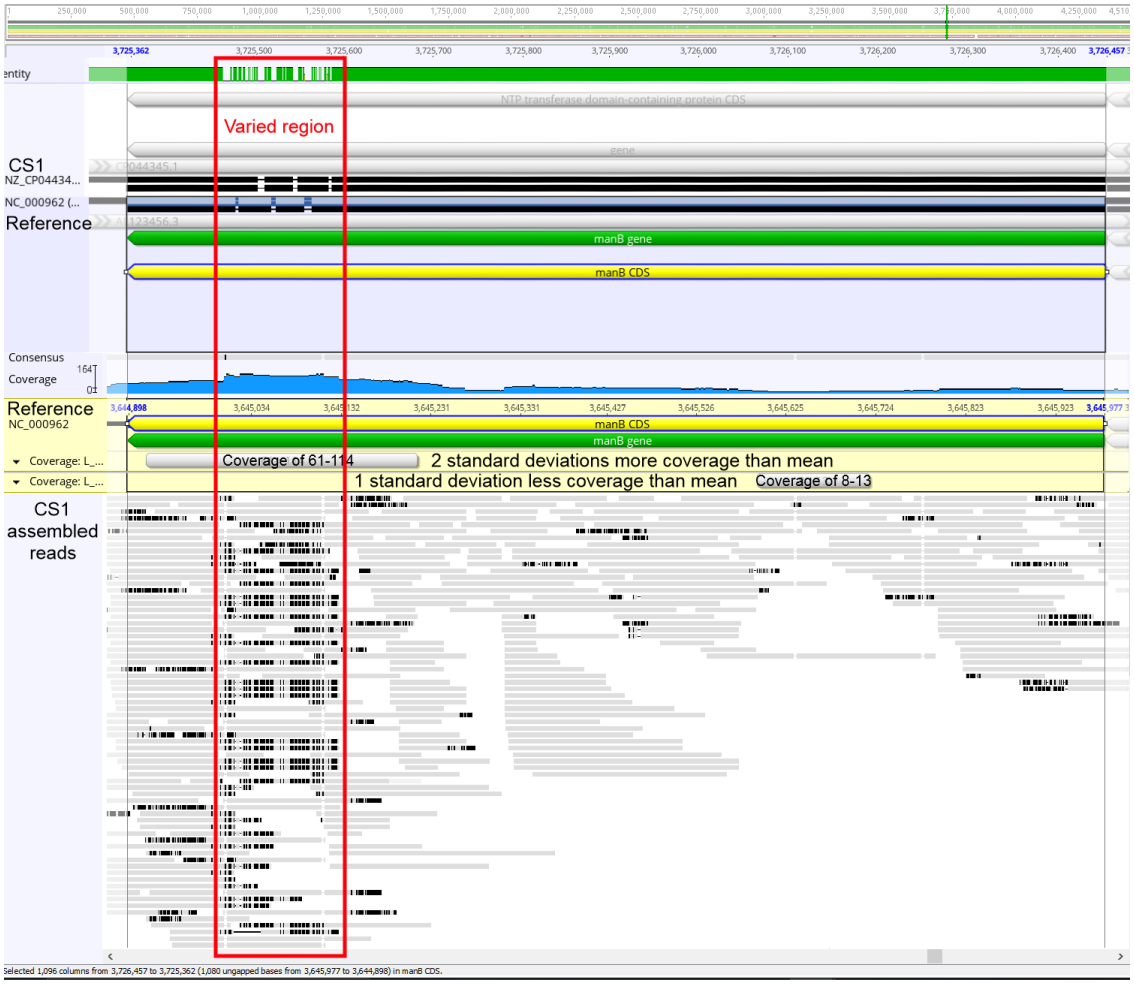

Figure 3: A composite image of manB, showing the H37Rv and CS1 alignment in Geneious (top), coverage graph from IGV (blue histogram), Illumina reads as mapped in Geneious to H37Rv with markers showing coverage above 2 SD or below 1 SD of the mean (yellow section, shown for a comparison in order to analyse any differences in read stacking or coverage gaps), and Illumina reads mapped to CS1 using the RGAPepPipe and viewed in IGV. In red is the VRIP, characterised by multiple insertions, deletions, and base changes. In the IGV reads section, black boxes show bases differing from the consensus.

3.1.2 FastQC

The following figures (4-13) illustrate FastQC quality scores for the Illumina reads before and after RGAPepPipe processing. The raw Illumina reads had very short and artificially long reads, high read duplication rates, odd GC percentage scores, and quality scores as low as a PHRED of 2 (an accuracy of 37%). After processing using the RGAPepPipe, reads are higher quality (lowest at 30, which is a base call accuracy of roughly 99.90%), deduplicated to acceptable parameters, have less variance in

length, and have a normal GC percentage distribution. The processed reads are of good quality and high-confidence, so the discrepancies in CS1 are not due to poor-quality read data.

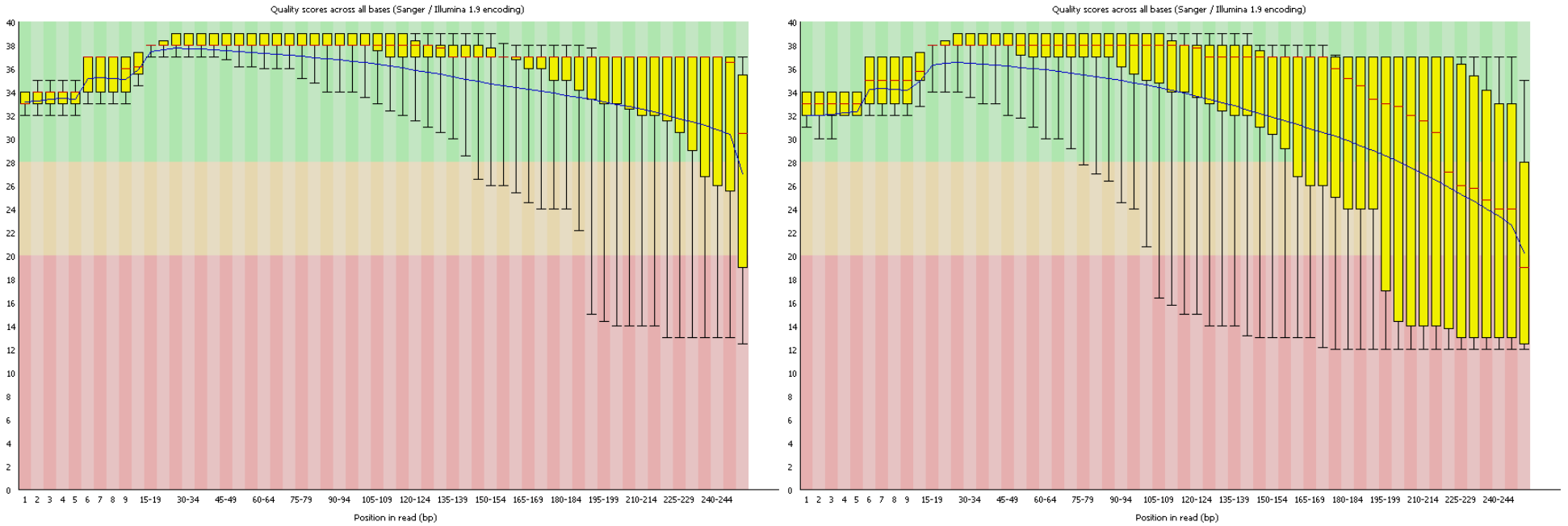

Figure 4: FastQC output showing the quality scores of bases inside the raw Illumina reads (paired reads separated). Forward-direction reads are shown on the left, and reverse-direction reads on the right. The scores are PHRED scores. The scores are inconsistent with values dropping into low-confidence areas near the end of the read (20 is an accuracy of 99%, less than this is a cause for concern due to the chance of the base being incorrectly called). Most values are high, which indicates processing will remove low-quality reads.

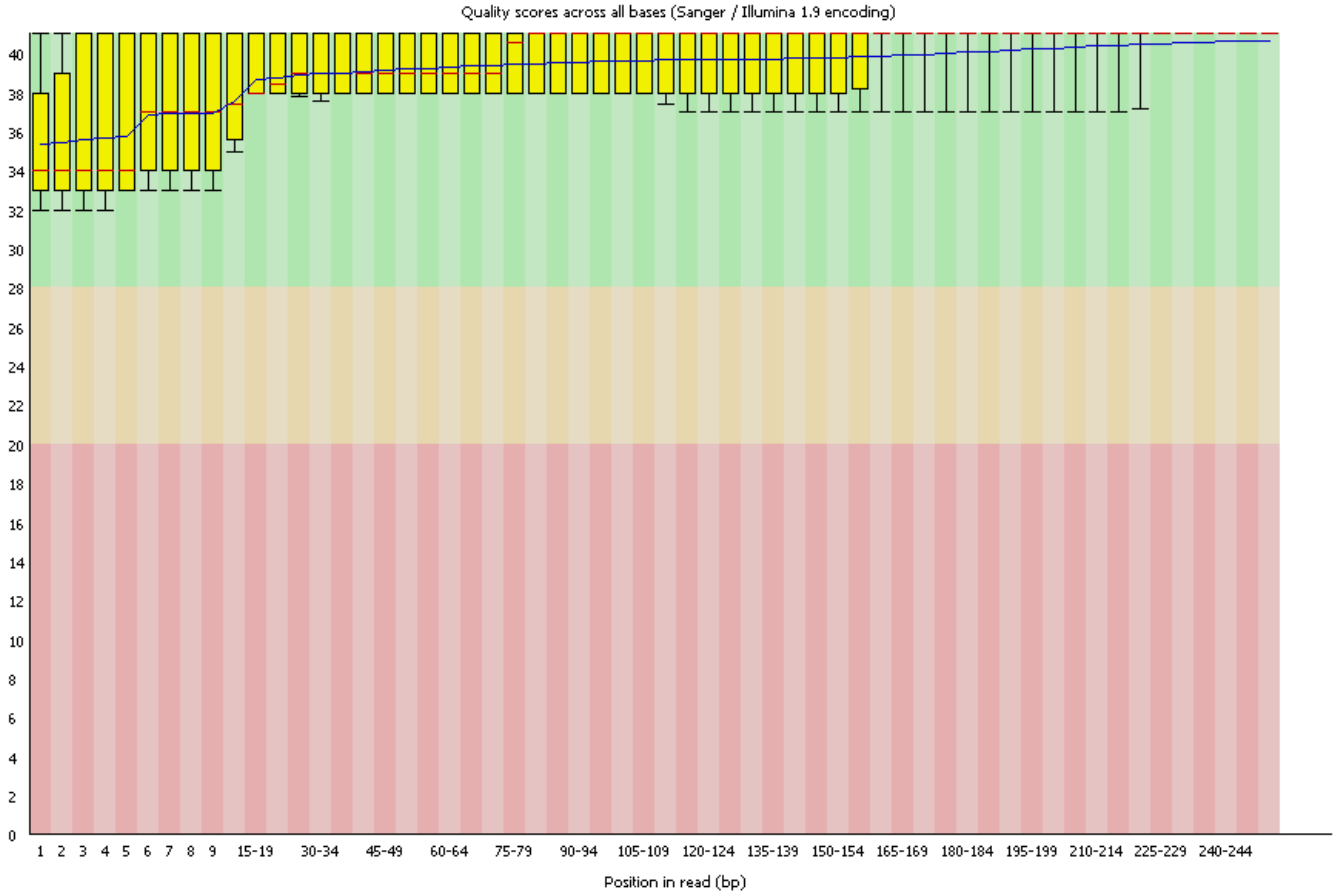

Figure 5: FastQC output showing the quality scores of bases inside the processed and deduplicated Illumina reads. The scores are PHRED scores. While the first 20 bases in a read are lower quality, this is by a small margin (32 is a call accuracy of 99.937%, 38 is a call accuracy of 99.9842%, and 40 is a call accuracy of 99.99%). The quality does not drop off at any point, which indicates low error rates and high-confidence data.

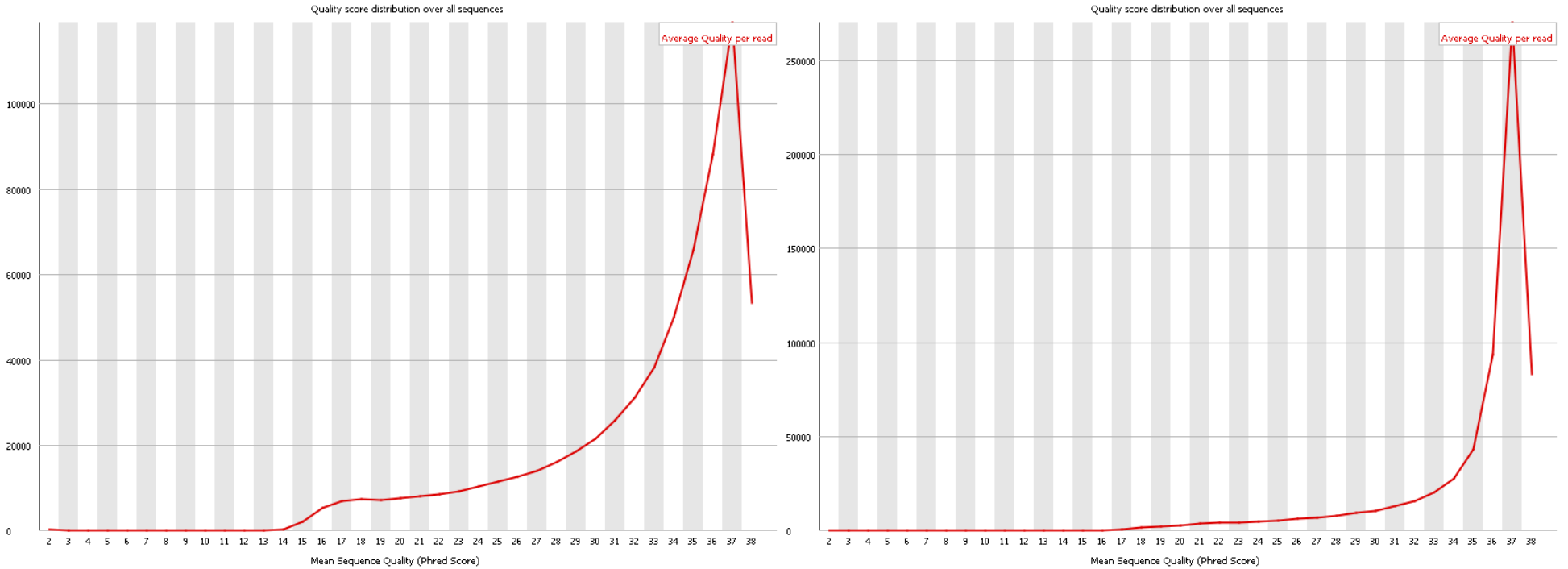

Figure 6: FastQC output showing the number of raw Illumina reads (paired ends separated) with certain quality scores, calculated by averaging the quality scores for each base inside the read. Forward-direction reads are shown on the left, and reverse-direction reads on the right. The scores are PHRED scores. The hump in the right-hand read-set around a PHRED of 17 (98% accuracy) indicates some issues, however the shape of both curves are quite good with a right-skew towards higher values.

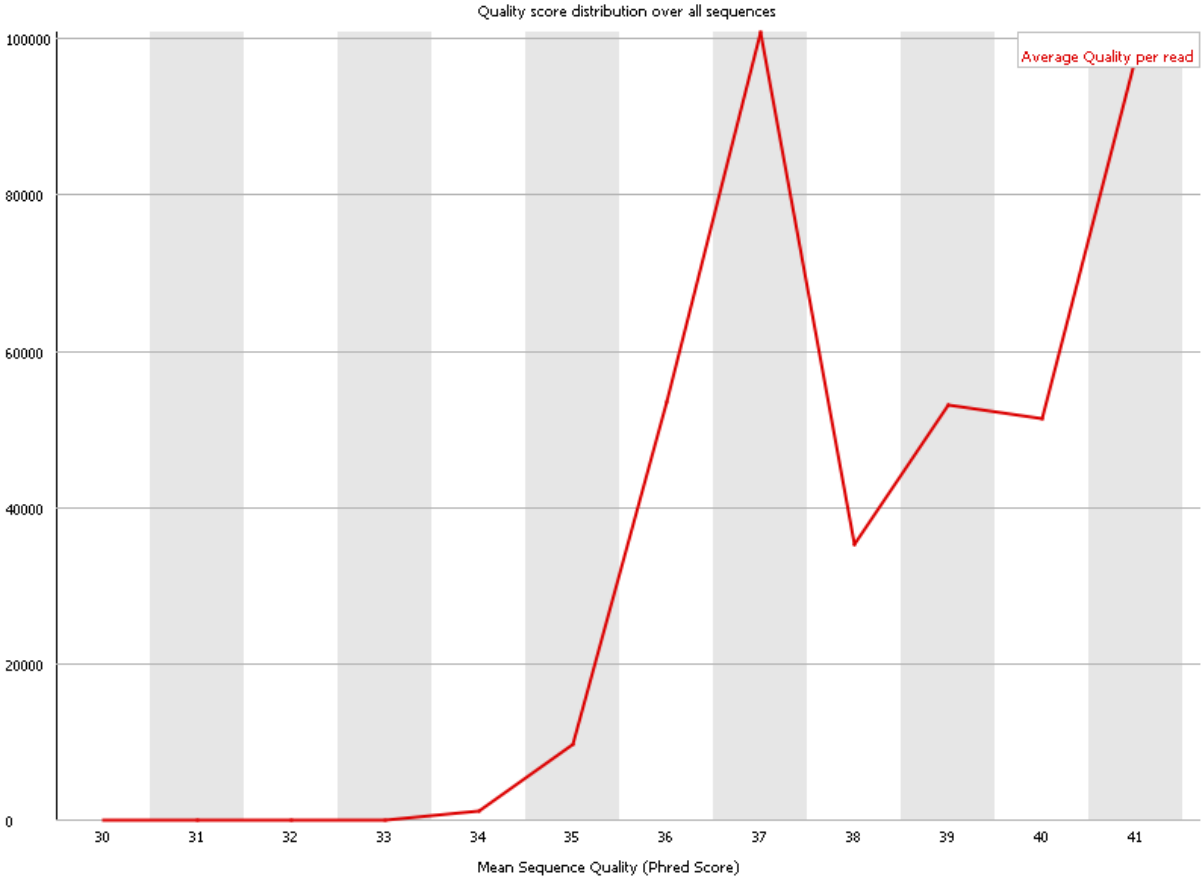

Figure 7: FastQC output showing the number of processed and deduplicated Illumina reads with certain quality scores, calculated by averaging the quality scores for each base inside the read. The scores are PHRED scores. Over 100,000 reads have a score of 37 (99.98% (2 d.p.) call accuracy). While not a smooth upward curve, the skew towards higher scores shows most of the data is of high quality with low error.

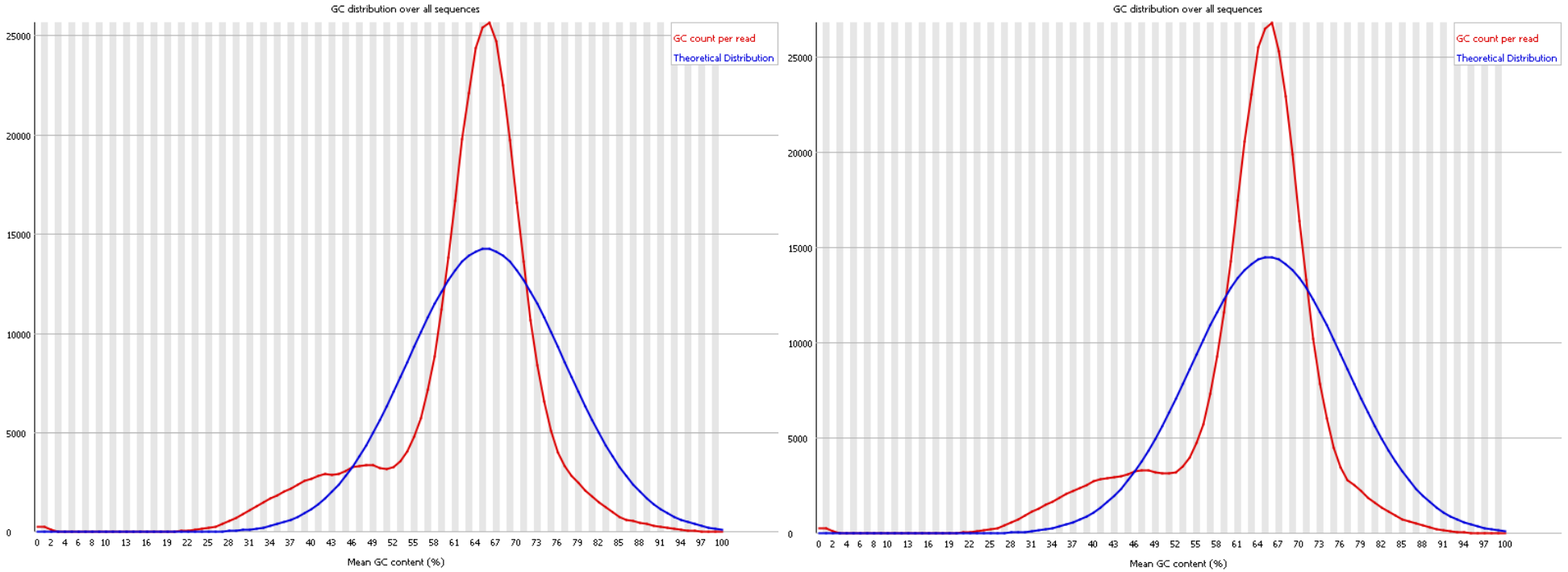

Figure 8: Graph from FastQC showing the number of raw Illumina reads (paired reads separated) with a given GC percentage. Forward-direction reads are shown on the left, and reverse-direction reads on the right. The ideal distribution is shown in blue, with the actual GC percentages shown in red. A normal distribution cannot be seen, instead there are peaks around 2% and 40% as well as around 66%, indicating poor-quality reads.

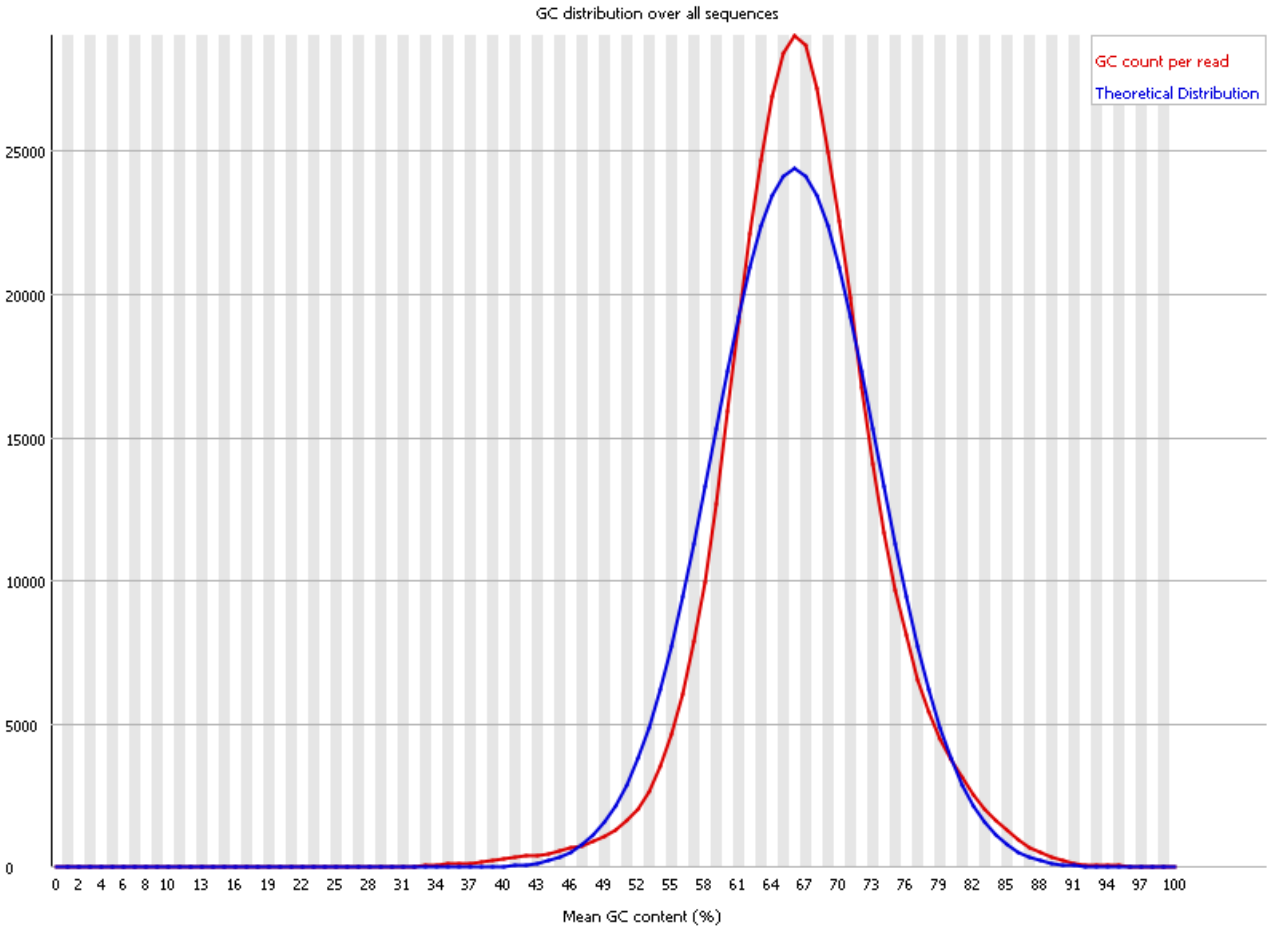

Figure 9: Graph from FastQC showing the number of processed and deduplicated Illumina reads with a given GC percentage. The ideal distribution is shown in blue, with the actual GC percentages shown in red. The data is near a normal distribution centred around 66% GC content, indicating good quality reads.

Figures 10 and 11 show read-length data before and after processing. Long reads are easier to map accurately, as the length allows more room for distinguishing features so the read is placed correctly. Short reads such as CATGAGACC, on the other hand, are ambiguous as they could have come from and could be mapped to any place in the genome. However, a large number of long Illumina reads indicates that poor-quality bases, often found at the ends and in reverse-direction reads, may not have been removed. Removing poor-quality bases improves the dataset, but will also reduce read sizes. Figure 10 shows a large peak above 250 bases for both forward and reverse-direction reads, which disappears after processing in Figure 11. This is consistent with poor-quality bases being removed from reads, increasing quality but decreasing read size.

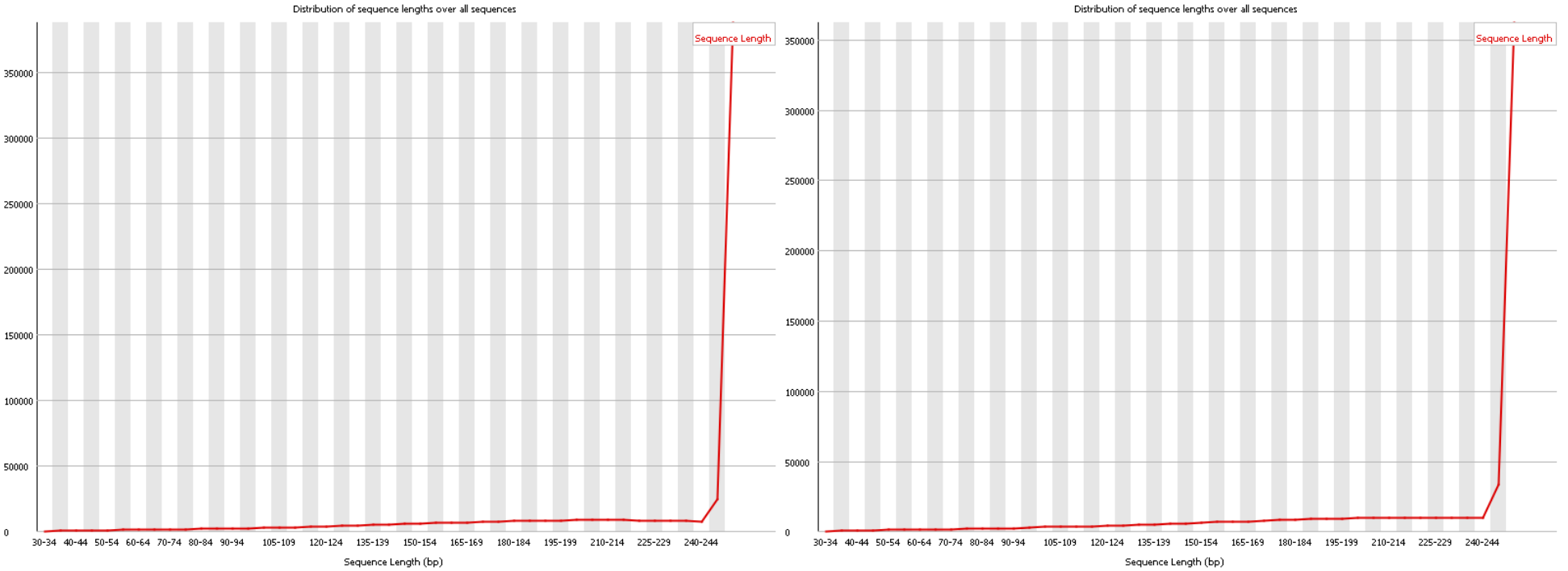

Figure 10: Graph from FastQC showing the lengths of raw Illumina reads (paired reads separated). Forward-direction reads are shown on the left, and reverse-direction reads on the right. Longer reads are more likely to be mapped accurately and without ambiguity, however the sharp peak indicates that low-quality bases may still be attached to the reads, making them artificially longer.

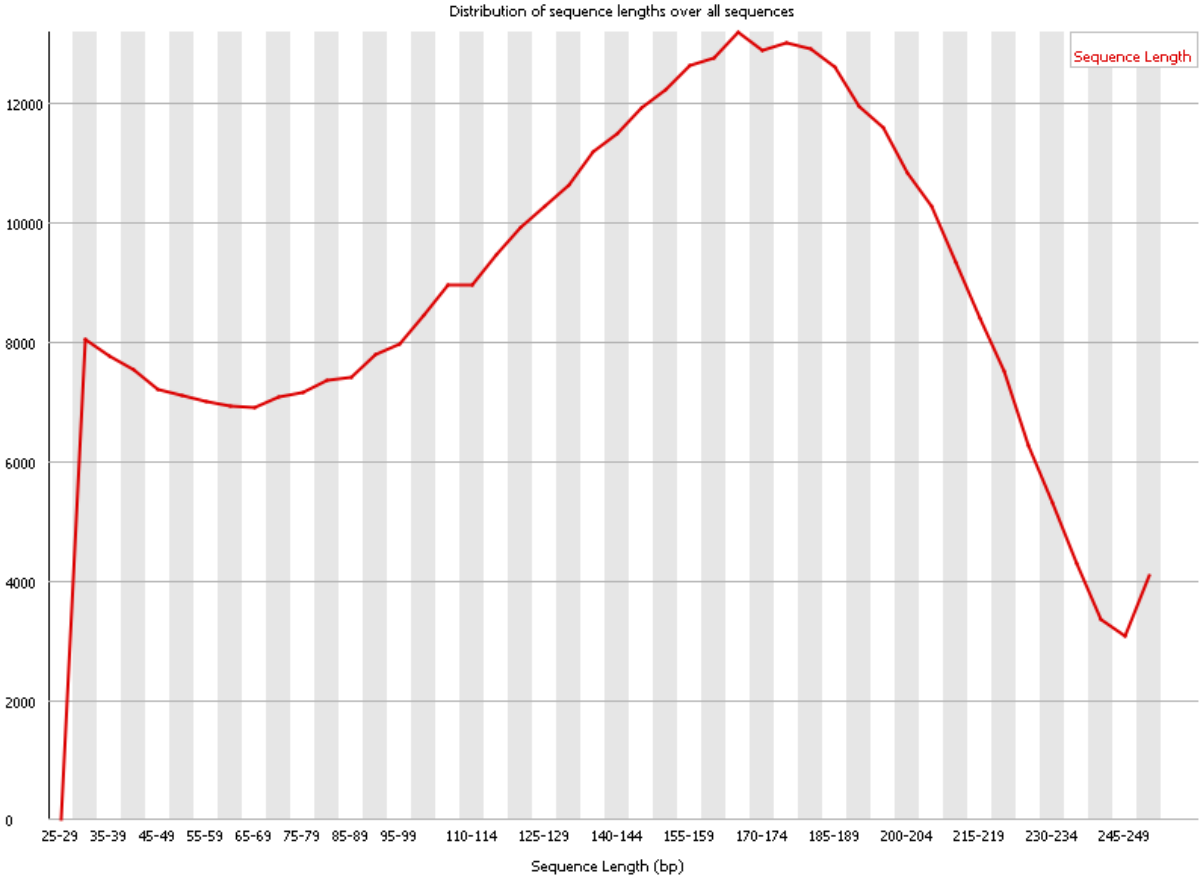

Figure 11: Graph from FastQC showing the lengths of Illumina reads used in the alignment. Longer reads are more likely to be mapped accurately and without ambiguity. Longer reads are also less likely to have errors caused by repeat regions. The modal length is around 160 bp. The shape of the distribution shows that lower-quality bases have been removed from reads, giving a variety of sizes.

Figures 12 and 13 show duplication data before and after processing. Reads which have been duplicated can cause the number of reads to be artificially high without adding any information. This can result in extra regions appearing, despite not actually existing in the genome. Another issue with duplicated reads are stacks of reads over a location which may or may not disagree with each other (making determining the true identity of bases difficult). Figure 12 shows there is some duplication in the raw Illumina reads, particularly with reads having 10-50 duplicates. Figure 13 shows the duplication reduced after processing and deduplication, but some duplication remains. Some duplication, as seen between 2 and 9 reads in Figure 13, is to be expected and can help improve coverage. The peak at 10-50 duplicates remains, although smaller, and could be due to large repeat regions giving the appearance of duplication.

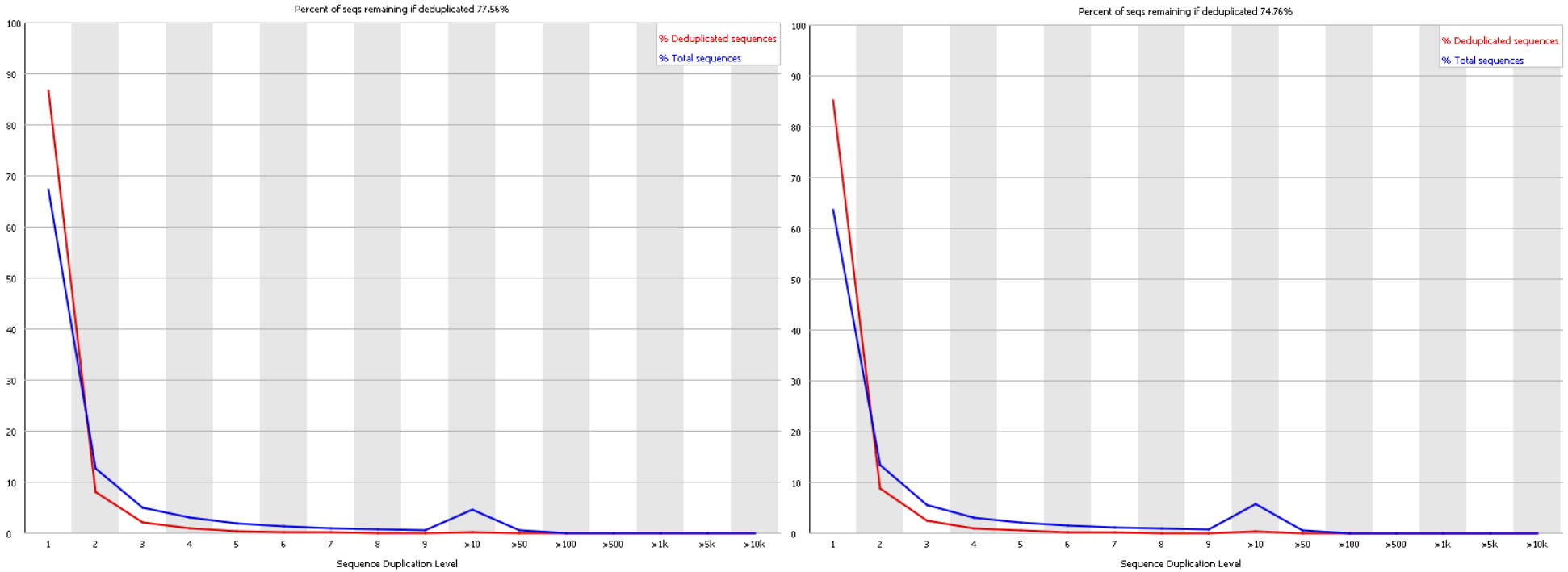

Figure 12: Graph from FastQC showing the duplication levels in the raw Illumina reads (paired reads separated). Forward-direction reads are shown on the left, and reverse-direction reads on the right. The current sequence is in blue, with red showing duplication levels if deduplicated to a given percentage (77.56% for forward-direction reads, and 74.56% for reverse-direction reads). FastQC indicated that these results were satisfactory, however the number of reads with a duplication between 10 and 50 was noted.

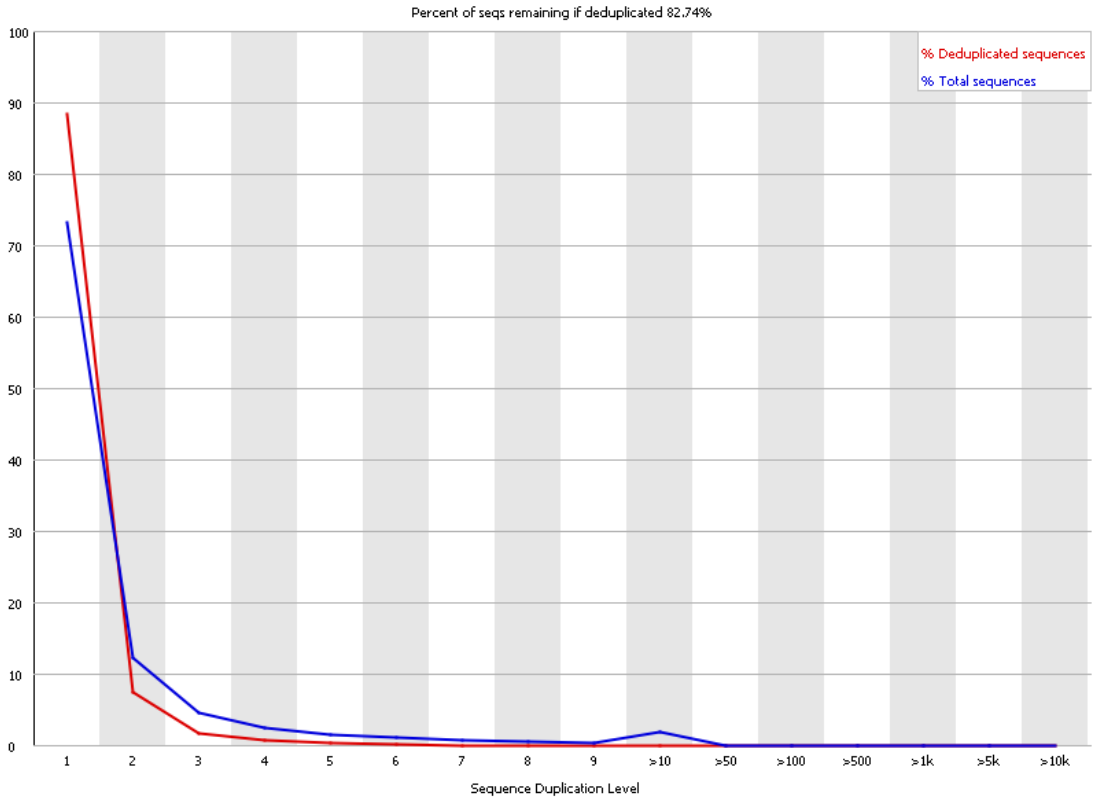

Figure 13: Graph from FastQC showing the duplication levels in the processed and deduplicated Illumina reads. The current sequence is in blue, with red showing duplication levels if deduplicated 82.74%. These reads had duplicates removed, which was successful as the data is now closer to the deduplicated line. The peak at 10-50 duplicates still remains, which may be due to repeat regions.

The FastQC output above shows that the raw Illumina reads benefitted from RGAPepPipe processing. Extremely short and artificially long reads with low-quality end-bases were removed along with duplicate reads. This improved

GC percentages, which can cause issues with gene identification when skewed, as well as overall base call accuracy.

3.1.3 Qualimap

Figure 14 and Tables 1-4 show the Qualimap output for the RGAPepPipe alignment of the Illumina reads to the CS1 draft, and show the alignment was successful and of good quality. Mean coverage appears good at 47 times, however the standard deviation is 40, meaning 34.1% of the CS1 draft is covered by between 7 and 47 Illumina reads. 0.04% of the draft has no Illumina-read coverage whatsoever, particularly on either side of where the genome was opened (incidentally, this is inside a highly-repetitive PPE gene), meaning there is no confidence in the features of those regions. Coverage of thirty or greater is preferred, however for most of the draft this is not an issue as the reads are good quality, have low error rate, and have high certainty. Furthermore, most of the CS1 draft is identical to H37Rv, which is expected as they are from the same lineage, so low coverage is not a concern for those regions. The RGAPepPipe assembly is therefore of good quality, with only small regional drops in quality and coverage which can be investigated.

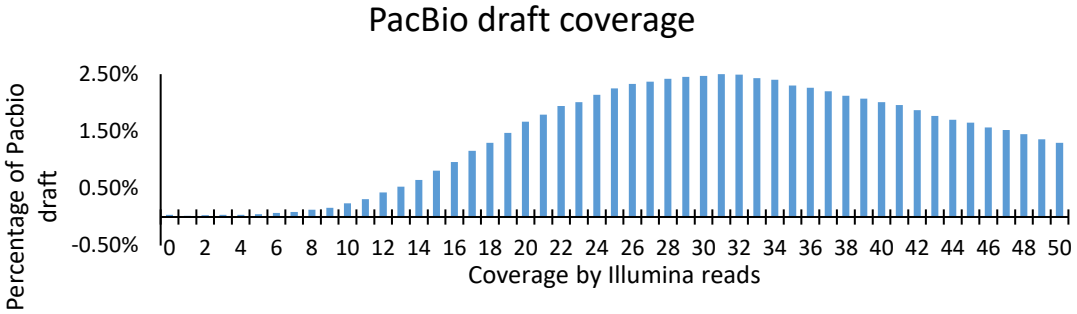

Figure 14: Graph showing the percentage of the PacBio CS1 draft covered by processed and deduplicated CS1 Illumina reads in the RGAPepPipe assembly. 0.04% of the PacBio draft has no coverage. The modal coverage is 31 reads, while the mean coverage is 47.2 reads with a standard deviation of 40.4 reads. Qualimap did not provide separated values for coverage over 50 reads.

Table 1: Quick snapshot of RGAPepPipe alignment of the processed and deduplicated Illumina reads to the PacBio draft.

Property	Mean value	Standard Deviation	Median
Mapping quality (PHRED)	57.96	Not Given	Not Given
Coverage (reads)	47.21	40.40	Not Given

Table 2: Insert sizes in the processed and deduplicated Illumina reads aligned in the RGAPepPipe to the PacBio draft.

Property	Lower Quartile	Median	Upper Quartile
Insert size (bp)	158	242	382

Table 3: Base frequency in the contig formed by the RGAPepPipe using Illumina reads mapped to the PacBio draft. A PHRED of 57.96 means the mean base-call certainty is 99.9998%.

Base	Number of bases	Percentage of bases
Adenine	34,970,206	16.78%
Cytosine	69,068,497	33.13%
Thymine	35,489,134	17.02%
Guanine	68,930,196	33.07%
N	0	0
GC percentage	N/A	**66.2%**

Table 4: Error data from the RGAPepPipe alignment of the processed deduplicated Illumina reads to the PacBio draft

General error rate per base	0.0072
Number of mismatches	1,476,725
Number of insertions	6,870
Percentage of mapped reads with insertions	0.48%
Number of deletions	12,861
Percentage of mapped reads with deletions	0.87%
Percentage of homopolymer indels*	59.09%

*This is the percentage of all insertion

3.2 Repeat Analysis

3.2.1 WindowMasker

WindowMasker (Winmasker) was used to investigate repeat regions in the CS1 draft, as large repeat regions can cause errors in sequencing, which can result in apparent discrepancies. Winmasker was also run on H37Rv as a comparison of repeat number and length. From the comparison of the two genomes, it does not appear as though there have been any significant slippage events, with one exception (a 1519 repeat in the CS1 draft). From Table 4 and Figure 15, most repeat regions are less than half of the length of the average Illumina read, making mistakes caused by repeat regions to be less likely. There is a difference in the pattern of outliers between CS1 and H37Rv, which could be associated with repeat-heavy genes (such as PE/PPE genes, involved in virulence). Winmasker also produced a list of repeat regions, which was used to verify potential involvement of large repeats in the appearance of CS1 discrepancies.

Table 5: Winmasker output for CS1 PacBio draft and H37Rv, showing the percentage of bases masked as repeats. CS1 is larger than H37Rv but also has a 0.046% increase in repeats. CS1 has more repeats by base pairs as well as individual repeat regions, but these repeat regions are marginally smaller than in H37Rv.

	Genome size	Bases masked	Repeat %	Number of repeats	Mean repeat size
CS1	4,416,671	783,296	17.735	34,327	22.819
H37Rv	4,411,532	780,336	17.689	34,168	22.838
Difference	5,139	2,960	0.046	159	-0.019

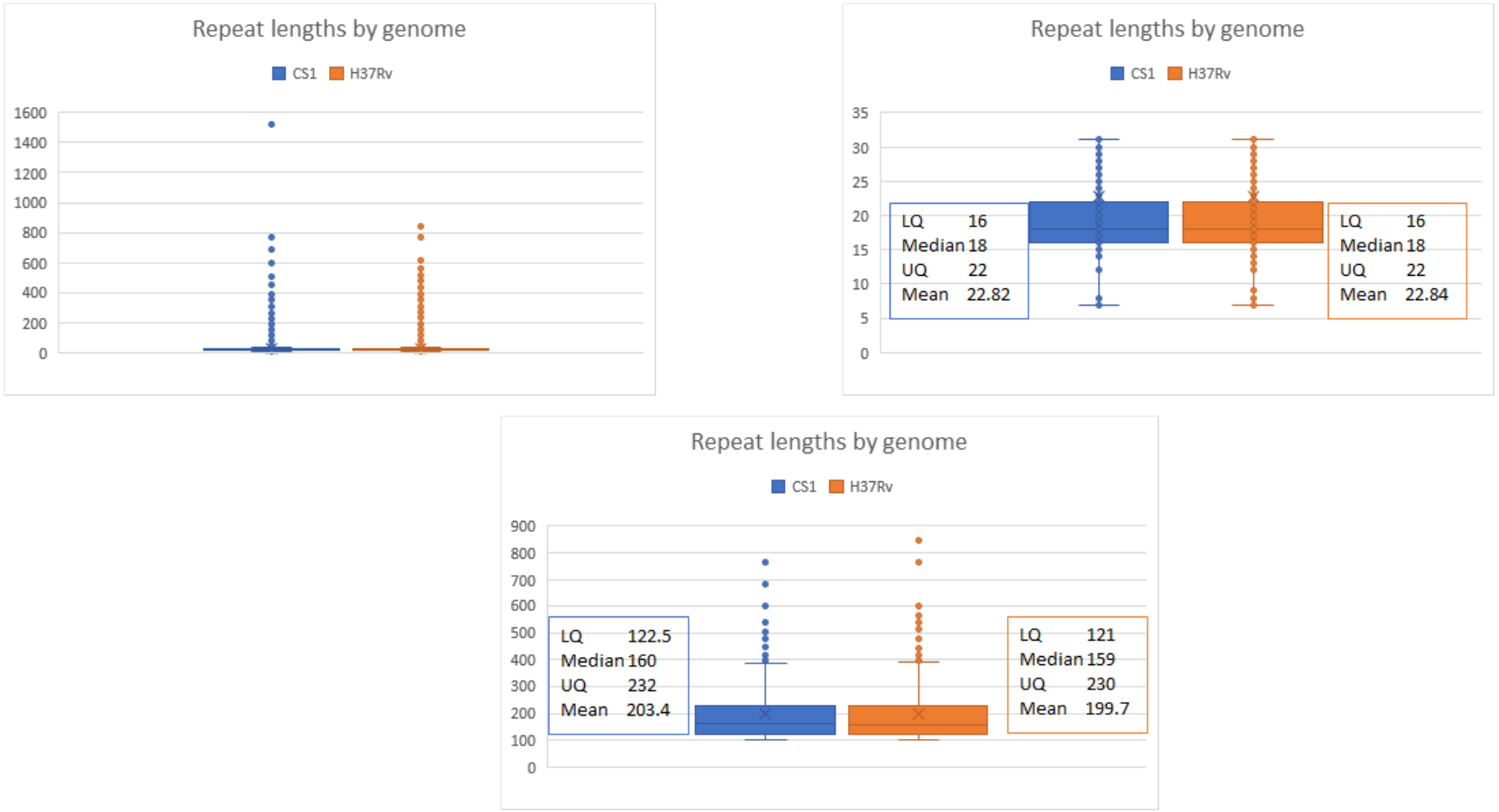

Figure 15: WinMasker output of repeat region lengths for CS1 PacBio draft and H37Rv visualised as box and whisker plots. On the upper left, outlier points are included, and the far outlier in CS1 at 1519 bases long is visible. On the upper right, the data distribution of the boxes is shown with the lower and upper quartiles, median, and mean values (calculated with outliers). The bottom plot shows the same as the upper right but only includes repeats at or above 100 bases in length (the plots exclude the far outlier in CS1, but the calculation includes it).

3.2.2 I_r

Data from Haubold and Weihe [37] was used in light of the WinMasker output to investigate where CS1 fits inside the diversity of *Mtb* repeats. Haubold and Weihe's I_r tool [37] measures how repetitious a region or genome is, rather than how many repeat regions or their location, with higher numbers indicating a more organised structure. The highly repeat-rich PE/PPE protein family is only found in *Mtb* species, suggesting that *Mtb* may have high genome repeat levels. However, Table 5 summarises the prokaryotes in the Haubold and Weihe dataset and shows that *Mtb* species were just above the lower quartile. Table 6 shows that CS1 sits close to the median value within *Mtb* species. The CS1 draft is not anomalous in its levels of repeats, further confirming that the assembly is likely accurate.

Table 6: Data from Haubold and Wiehe investigating the I_r score for 330 species of prokaryote, including five species in the *Mycobacterium tuberculosis* cluster. A comparison between the *Mtb* cluster and the entire dataset is shown. An I_r score of 0 shows a completely random sequence with no repetition. The highest I_r value was 6.337 for *Methylobacillus flagellatus KT*, showing a highly organised and repeat-rich genome. The subsection of *Mtb* cluster tested is in the lower 50% of all prokaryotes for repeat content. The addition of CS1 had minimal effect on the values.

	Maximum	Upper Quartile	Median	Lower Quartile	Minimum	Mean	Mode
All prokaryotes	6.337	1.3248	0.86275	0.561875	0.0189	1.0007	1.7159
Mtb	0.833	N/A	0.5647	N/A	0.4082	0.5767	N/A
Mtb incl. CS1	0.833	N/A	0.5859	N/A	0.4082	0.5818	N/A

Table 7: Data from Haubold and Wiehe comparing the I_r scores for 5 species in the *Mycobacterium tuberculosis* cluster, with CS1 added. An IR score of 0 shows completely random sequence with no repetition.

Accession	Species and strain name	Genome length	I_r score
NC_002677	*M. leprae* TN	3,268,203	0.833
NC_002944	*M. avium* subsp. paratuberculosis K-10	4,829,781	0.6248
NC_000962	*M. tuberculosis* H37Rv	4,411,532	0.5647
NC_002755	*M. tuberculosis* CDC1551	4,403,837	0.453
NC_002945	*M. bovis* AF2122/97	4,345,492	0.4082
NZ_CP044345	*M. tuberculosis* CS1	4,416,671	0.607

Haubold and Weihe's I_r tool was used to investigate the organisation of repeat regions inside the CS1 draft. Three window sizes were used (1,000 bases, 5,000 bases, and 10,000 bases), and repeat levels were measured inside each window over the whole genome. It is notable that the lowest coverage regions around the opening of the genome have repeat peaks. As seen in the WinMasker data, most repeats are small and close together, causing peaks of high repetition levels which decrease as the window size increase.

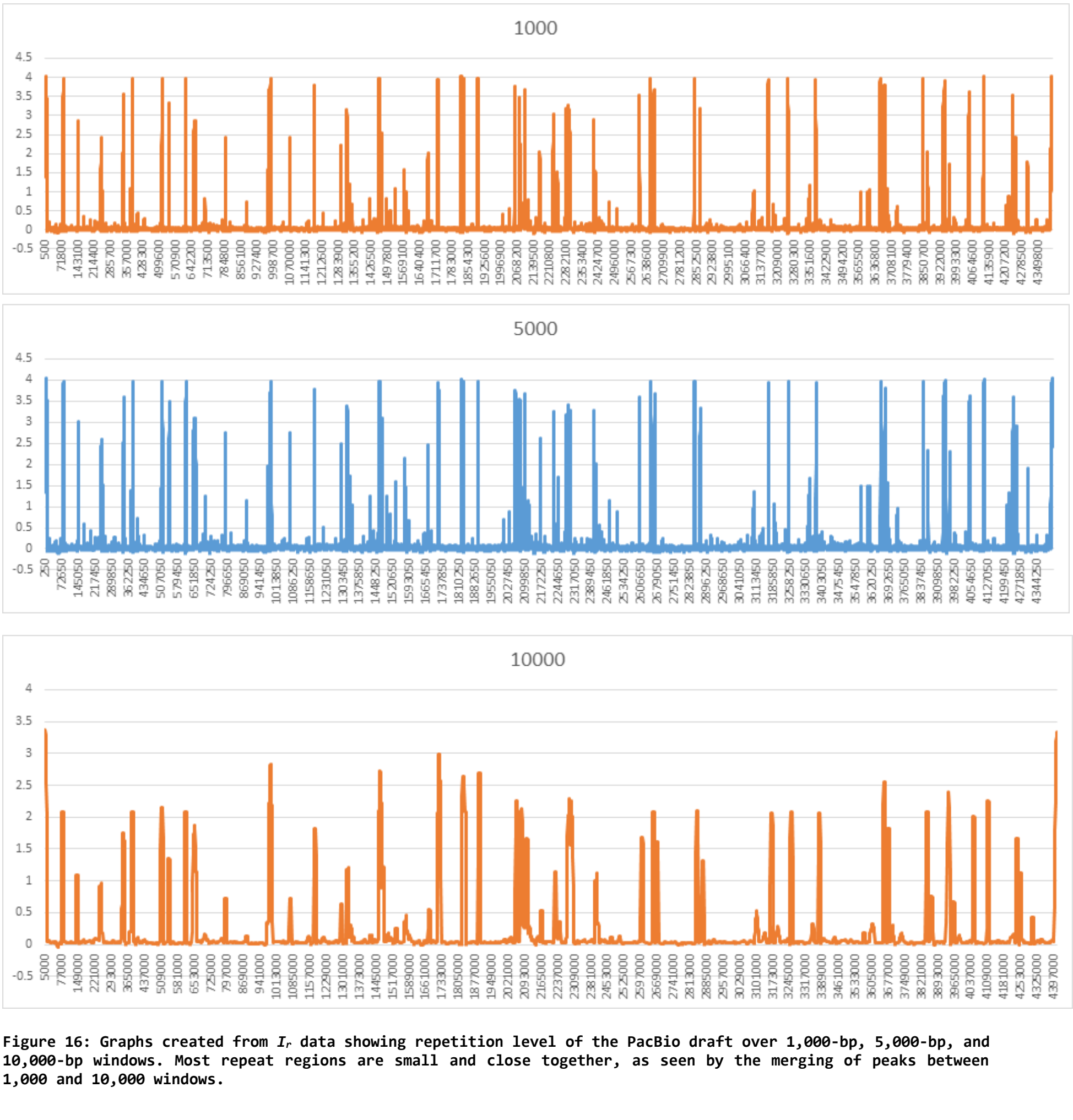

Figure 16: Graphs created from I_r data showing repetition level of the PacBio draft over 1,000-bp, 5,000-bp, and 10,000-bp windows. Most repeat regions are small and close together, as seen by the merging of peaks between 1,000 and 10,000 windows.

3.3 Genome Analysis

3.3.1 BRIG

As the quality of the data and assembly have been confirmed, the individual discrepancies were investigated using BRIG. Figure 17 shows the image created by BRIG with relevant genetic features, and the VRIPs shown there are listed in Table 7. VRIPs which overlap with low quality and low coverage regions or with large repeats can be disregarded as unverifiable by this project. Of the 32 identified VRIPs, 19 were in this category. This does not mean they are not true features of CS1, but they will need to be resequenced so better data quantity and quality can be obtained to resolve them. It is of note that because of the annotation system changes, Table 7 is not a complete list of VRIPs, nor of variable regions outside of proteins. The CS1 draft would benefit from searching for more of these discrepancy regions, and targeted resequencing and polishing for unverified VRIPs.

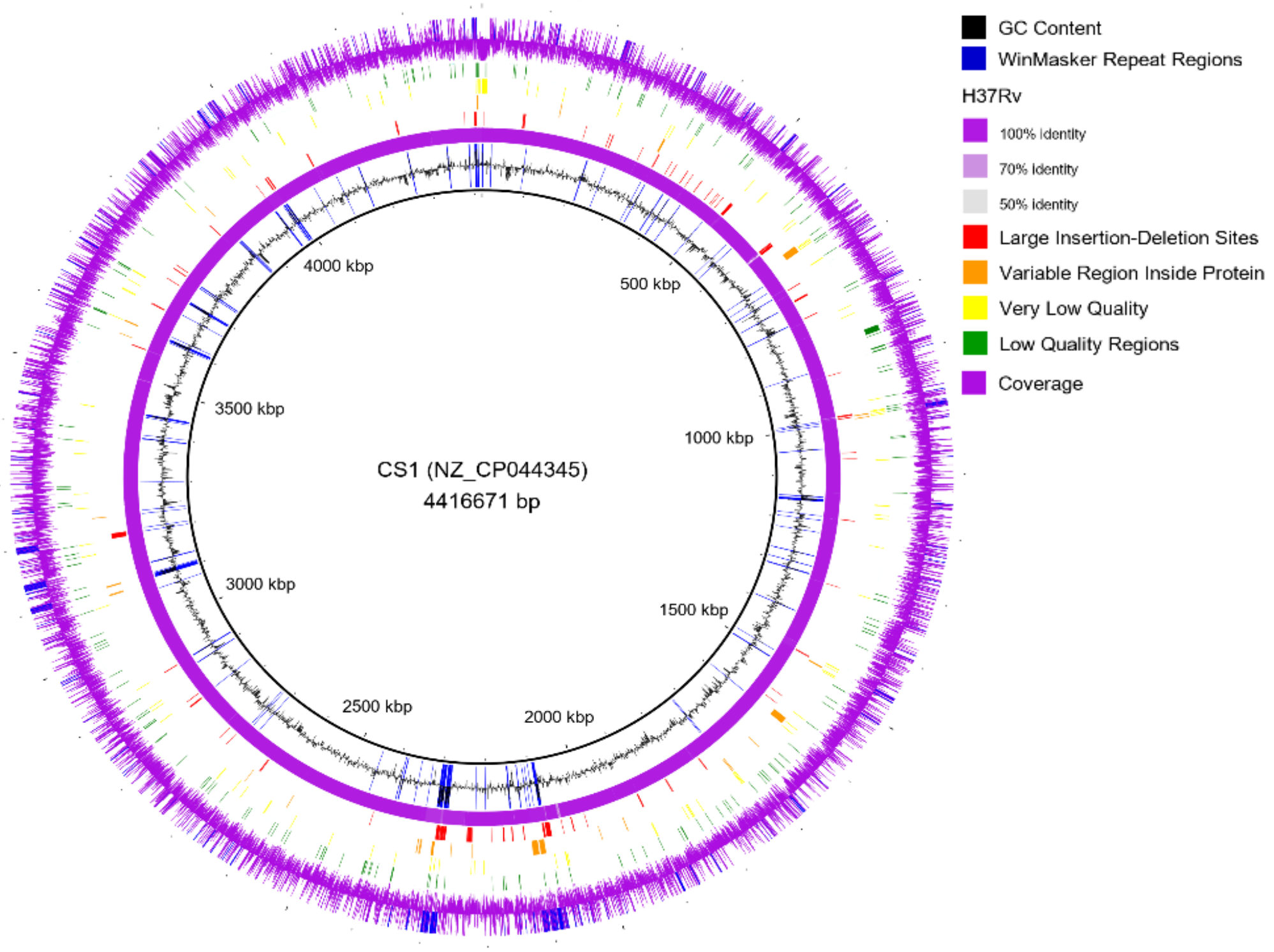

Figure 17: The circular CS1 genome in BRIG, showing (from the centre outwards) GC content and repeat regions (blue), H37RV comparison, sites of large insertion-deletions and Variable Regions Inside Proteins (VRIPs), regions where the cumulative quality score of all reads is under 300 (yellow) or 600 (green), and coverage (purple). Large insertion-deletions tended to be either suspected annotation errors or transposons and similar elements, and have been included out of interest. The figure shows only a small amount of clustering for VRIPs and large indels, with most repeat regions overlapping with low quality and low coverage regions but not with all VRIPs or large indels. Coverage values spike at various points over the genome. Note that the blue markings over the coverage graph appear to be a software artefact. Large insertion-deletions are notable and should be reviewed in another project, as many of these are transposons and similar elements, while others are genes and parts of genes which have been added or deleted.

The 32 VRIPs shown in Figure 17 are listed in Table 7 with base position, the name and locus tag of the gene it appears in, whether the gene coincides with a Low or Very Low quality region, and the number and largest size of repeats in the gene. Genes which had low quality regions were discarded. Not all genes had annotated names, and some annotation positions had changed between being found in December 2019 and the current analysis. Of the 13 VRIPs not explained by repeat regions or poor-quality data, 5 were chosen for further analysis on the basis of potential importance to virulence or cell survival.

Table 8: A list of the 32 CS1 VRIPs shown above. Each gene contains one VRI. The start and end base positions of the gene are shown, alongside the locus tag assigned in the latest-accessed version of the PacBio draft. Where genes have been split into 2 or more annotations since the VRIs were discovered, relevant accessions are noted. Quality shows if there is an overlap between the gene and a low-quality region as seen above. For VRIPs that did not overlap with a low-quality region, the number and largest size of repeats were measured. Of the 32 VRIPs, 5 had annotations that no longer exist, while another 8 overlapped with very-low-quality regions. Of the 19 remaining VRIPs, 6 had repeats over 100bp long, which could be the true cause of the discrepancies. The other 13 VRIPs cannot be explained by low quality data or repeat-related errors and are therefore of significant interest: GlyS, Cas10/csm1, QEX90791, QEX91071, manB, lprN, PPE61, kstD, QEX91665-QEX91666, pheA, hypothetical F6W99_02917, fadD2, and rplR-rpsE. It would be ideal to search again for VRIs to find all inside and outside current annotations. It would also be ideal to further investigate the repeats over 100, and the exact placement of low-quality regions in relation to the VRIs.

Gene Start	Gene End	Quality	Locus Tag	Gene name	Number of repeats	Largest repeat
6271	7056	Very Low	F6W99_00008	Antibiotic ABC transporter	Not measured	Not measured
144343	144864	Very Low	F6W99_00131	None	Not measured	Not measured
144912	145304	Very Low	F6W99_00132	None	Not measured	Not measured
338887	342277	Very Low	F6W99_00320	PPE24	Not measured	Not measured
648127	660577	Very Low	None	None	Not measured	Not measured
986012	987835	Good	F6W99_00962	Putative PPE40	11	117
991754	993145	Good	F6W99_00968	GlyS_2 – tRNA glycine ligase	4	17
1470411	1472849	Good	F6W99_01431-33	CRISPR-associated protein Cas10/Csm1	9	33
1584992	1596982	Very Low	None	None	Not measured	Not measured
1656532	1657752	Good	F6W99_01579	Unknown integral membrane protein, QEX90791.1	10	37
1718589	1719896	Very Low	F6W99_01640	PPE46_1	Not measured	Not measured
1721029	1722336	Very Low	F6W99_01645	PPE46_2	Not measured	Not measured
1938993	1940495	Good	F6W99_01864	Unknown two-component sensor kinase, QEX91071.1	10	24
1987859	1988938	Good	F6W99_01911	manB	8	31

Gene Start	Gene End	Quality	Locus Tag	Gene name	Number of repeats	Largest repeat
2092499	2101965	Good	F6W99_02005-06	PPE55	81	209
2104733	2115883	Good	None	None	Not measured	Not measured
2226435	2226731	Very Low	F6W99_02126	None	Not measured	Not measured
2264178	2265332	Good	F6W99_02168-69	lprN/mce4E (starts F6W99_02168)	9	33
2296601	2301024	Very Low	None	None	Not measured	Not measured
2320106	2321363	Good	F6W99_02208	PPE61	15	44
2321505	2323218	Good	F6W99_02209-10	PPE62	14	153
2326134	2327825	Good	F6W99_02214-16	kstD, with interrupting hypothetical	14	25
2574668	2575864	Good	F6W99_02468-69	Hypothetical aminotransferase. QEX91665.1 and QEX91666.1.	11	46
2664033	2664998	Good	F6W99_02535	pheA (prephenate dehydratase)	8	40
3088785	3089966	Good	F6W99_02917	None	13	53
3090068	3091696	Good	F6W99_02918	FadD2	13	34
3104942	3107455	Good	F6W99_02930	PE-PGRS4	16	513
3234963	3236294	Good	F6W99_03039	PPE32_2	16	129
3584427	3585477	Good	F6W99_03395, F6W99_03396	rplR and rpsE, originally in one annotation	12	78
3608936	3611257	Very Low	F6W99_03424	PE-PGRS10	Not measured	Not measured
3865113	3866120	Good	F6W99_03677	PE-PGRS17	1	274
4407495	4409706	Very Low	None	None	Not measured	Not measured

3.4 VRIP Analysis

The 5 chosen verifiable VRIPs were compared with H37Rv in Clustal Omega to show how the VRIPs affect the amino acid sequences (shown in Figures 18, 21, 25, 28, and 33). They were also compared in the same tool with other strains to explore if any similar features appear in other strains, and where any differences are in relation to the VRIP. If other strains have similar features, this makes their appearance in CS1 more plausible. Alternatively, if these features are unique to CS1 or found in strains known for their virulence, then this may provide some clues about its enhanced transmissibility. To do this comparison, the sequences were run through BLAST to find similar sequences in other strains, and non-redundant sequences were removed from the list as those stand in for many strains and are hard to trace. The exception to this were matches from candidates which were part of the MTC but non-*M. Tuberculosis,* such as *M. bovis*, as exact location and strain are less important in those cases. For the remaining strains, phylogenetic trees were created in Clustal Omega to find the closest matches. The closest matches and any drug resistant, unusual, or notable (Beijing, CDC1551, Haarlem, F1) strains were used to produce the final phylogenetic trees (shown in Figures 19, 22-23, 26, 29-31, and 34) and comparison images (shown in Figures 20, 24, 27, 32, and 35).

Protein accessions are used here, instead of the gene accessions. The symbols under the alignments refer to residues with similar properties as determined by the Gonnet PAM 250 matrix. Positions with strong conservation (a matrix score greater than 0.5) are shown with a “:”. These positions have amino acids with similar chemical properties, for example serine substituting for threonine. Positions with weak conservation (a matrix score greater than 0 but less than or equal to 0.5, such as cysteine and alanine) are denoted by “.”. Positions without any symbols show where either

there is no conservation or where one or more sequences have a deletion.

3.4.1 Cas10/csm1

The Cas10/csm1 VRIP causes the CS1 variant to split in two, as shown in Figure 18. Most of the VRIP is in the non-coding region between the two coding regions, however 4 amino acids are affected. The changes to those amino acids are non-conserved, which is consistent with their position at the end of the N-terminal region. Truncations are seen in other strains, but not in the same place as CS1 (Figure 20). Notable matches to the N-terminal region, as seen in Figure 19 include CDC1551 (also known as OshKosh), *M. africanum*, and *M. bovis*. The C-terminal region was not given a phylogenetic tree as this region was identical to H37Rv

Figure 18: A comparison of Cas10/csm1 in H37Rv (NP_217339.1) and CS1 (QEX90644.1 and QEX90645.1, split by the VRI). A frame-shift causes the early stop in QEX90645.1, and not seen in the amino acid sequence is a series of insertion-deletions after the end of QEX906445.1. Earlier in QEX90645.1 there is a 9bp insertion, which is part of the VRIP while distant from the main cluster. Strains with this sort of shortening are of interest, as are those with the added bases.

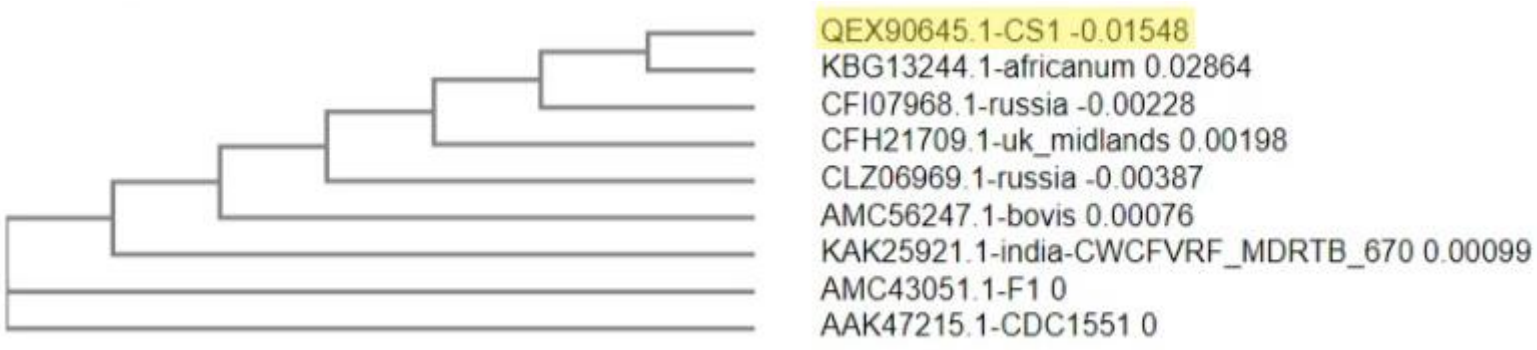

Figure 19: A phylogenetic tree from Clustal Omega showing the relatedness of matches for csm1.1. Two Russian strains are represented, as well as a strain from the UK. A multi-drug-resistant strain from India, F1, and CDC1551 (the latter two notable for outbreaks) also appeared as matches. *M. africanum* and *M. bovis* are represented, with *M. africanum* being the closest match to CS1.

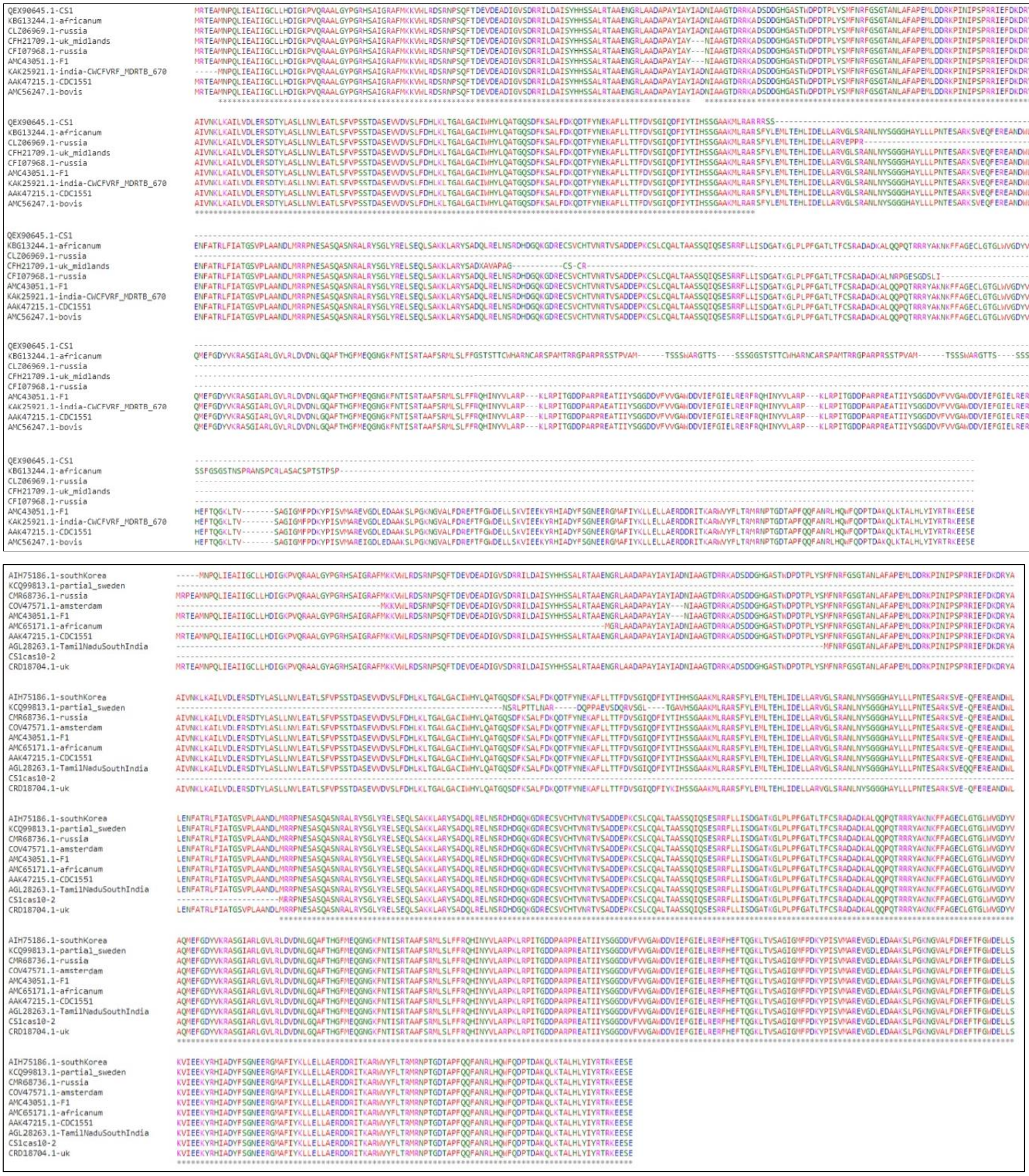

Figure 20: Filtered BLAST comparison results for csm1.1 (top) and csm1.2 (bottom). No strain shares the last four amino acids not truncation in csm1.1, although other strains are truncated slightly later. Csm1.2 is similar, with no strain starting as late, however the match is identical and overall the matches contain less variation than with the full-length proteins matching with csm1.1. This is of interest as the VRIP occurs at the end of csm1.1, while other strains appear to have variation nearby, such as the 3-residue deletion in F1 and the early truncation of both Russian strains.

3.4.2 lprN/mce4E

The lprN/mce4E VRIP causes the CS1 variant to split in two, as shown in Figure 21. While most of the VRIP is in the non-coding region between the two coding regions, 12 amino acids at the end of the N-terminal section are affected, with 2 more deleted. The changes to those amino acids are mainly non-conserved. Truncations are mainly seen in strains with only a partial sequence and so may not be a true reflection of length, and the amino acids affected by the VRIP do not match any other strains (Figure 24). The strains closest to CS1 changed between the N and C terminal sections, but many were unchanged, including several drug-resistant strains (Figures 22-23).

Figure 21: A comparison of lprN in H37Rv (NP_218012.1) and CS1 (QEX91374.1 and QEX91375.1, split by the VRI). The VRI is "bookended" by first a 10bp frameshifting deletion in CS1 which causes the early stop in QEX91375.1, then a 10bp insertion. QEX91375.1 starts with the next V. LprN2 is where the VRIP occurs. Strains with this sort of shortening are of interest.

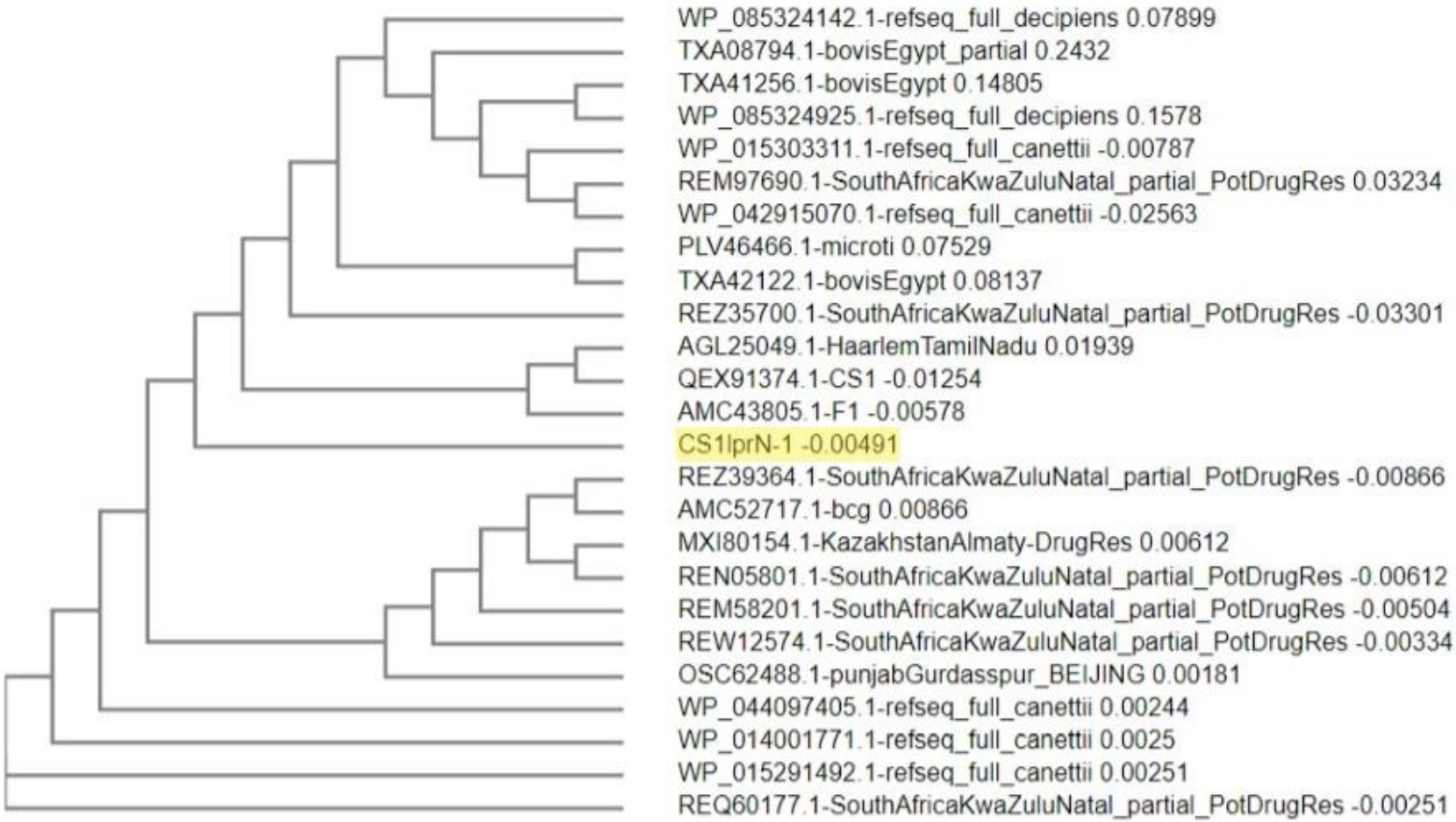

Figure 22: A phylogenetic tree from Clustal Omega showing the relatedness of matches for lprn1. Several partial proteins from potentially drug resistant outbreaks in South Africa appear, alongside Beijing-family strains, the vaccine BCG strain, a drug-resistant strain from Kazakhstan, contagious strains like Haarlem and F1, and several non-*M. tuberculosis* strains (*M. bovis*, *M. microti*, *M. canettii*, *M. decipiens*). CS1 has no close matches.

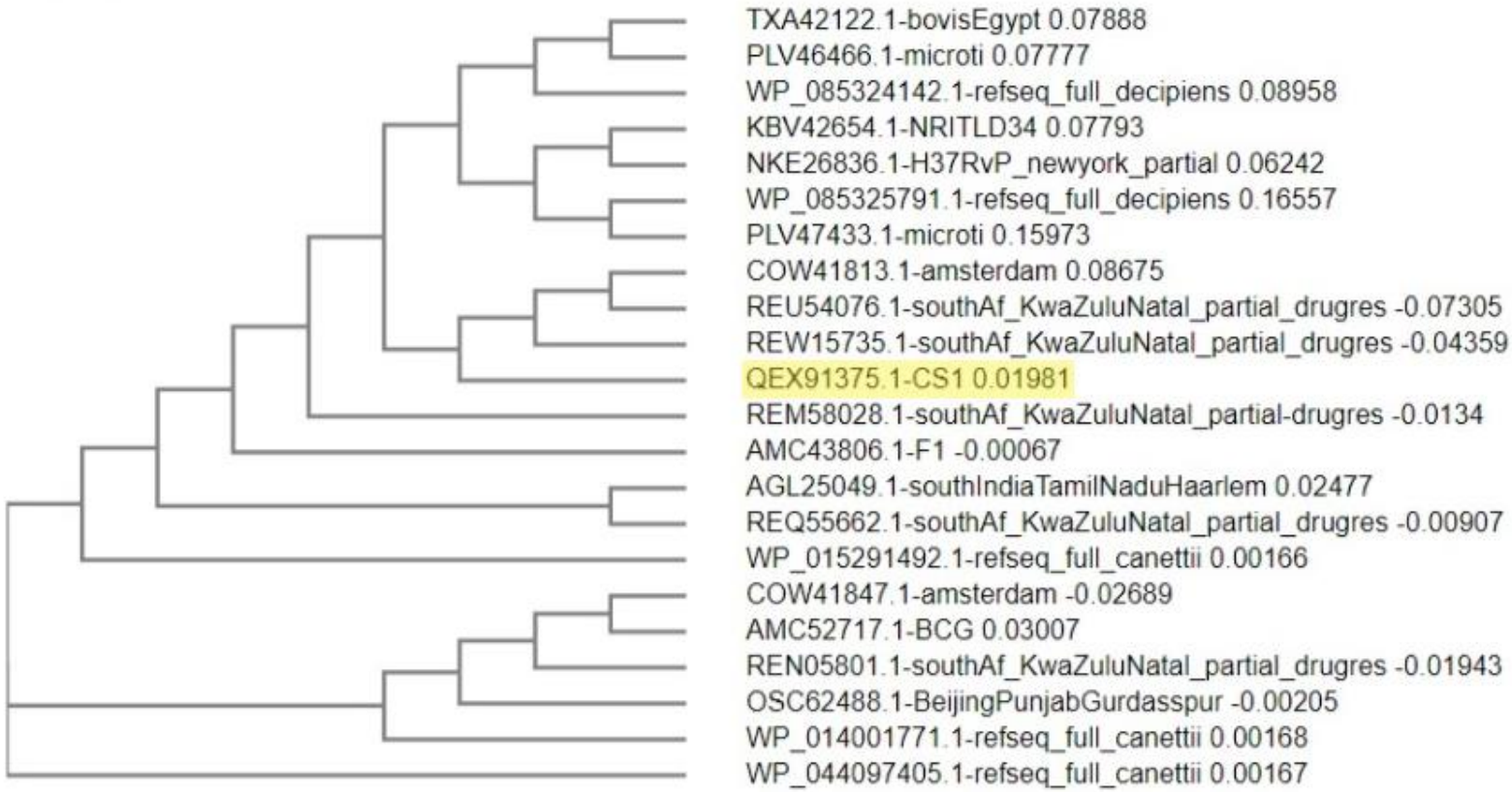

Figure 23: A phylogenetic tree from Clustal Omega showing the relatedness of matches for lprn2. Several partial proteins from potentially drug resistant outbreaks in South Africa appear, alongside Beijing-family strains, the vaccine BCG strain, two representatives from Amsterdam, contagious strains like Haarlem and F1, and several non-*M. tuberculosis* strains (*M. bovis*, *M. microti*, *M. canettii*, *M. decipiens*). CS1 has no close matches. NRITLD34 and a partial from an early H37Rv patient appear here. CS1 is in a cluster, with one Amsterdam strain and two from South Africa being the closest matches.

Figure 24: Filtered BLAST comparison results for lprN1 & 2 (in order, 2 being N-terminal). The VRIP occurs in lprN2 and there appears to be a greater level of variation in matched strains for lprN2, however the other strains matching lprN1 suggest a similar separation of the two protein segments. LprN2 has matched with strains where the protein appears to be full size, however the last 12 residues for CS1, which match the VRI location, do not match any other strain.

3.4.3 PE_PGRS17

The PE_PGRS17 VRIP includes a single 4 amino acid insertion and numerous point mutations, as shown in Figure 25. This makes it unusual, as most VRIPs contain numerous insertions and deletions even if this isn't seen at the amino acid level. The changes to amino acids in the VRIP are mainly conserved, however a surprising number are not conserved. Most of the matches are partial records missing the N terminus, which includes a lot of the VRIP region (Figure 27). Of those strains which are not missing the VRIP due to partial sequencing, all are identical except for *M. bovis*. The matches include 2 non-human strains and two drug-resistant strains (one of which is part of the Beijing strain family) (Figure 26).

Figure 25: A comparison of PE_PGRS17 in H37Rv (YP_177774.1) and CS1 (QEX92859.1). It is notable that the only insertion-deletion in this VRIP is the insertion of EQAL in CS1, the rest of the changes are due to point mutations. Many of the mutations in this VRIP are silent or complimentary.

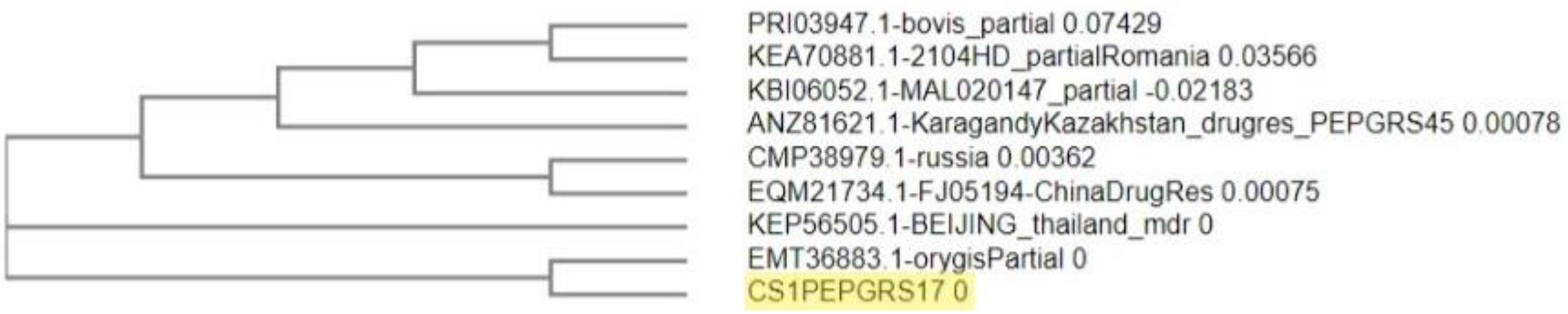

Figure 26: A phylogenetic tree from Clustal Omega showing the relatedness of matches for PE_PGRS17. Drug-resistant strains from China, Kazakhstan, and Thailand appear, with the latter being from the Beijing strain family. Partial records from *M. bovis*, *M. orygis*, Romania, and strain MAL020147 have matched as well. *M. orygis* is the closest match to CS1.

Figure 27: Filtered BLAST comparison results for PE_PGRS17. CS1 is longer which may be due to four of the results being partial records, otherwise it is no different from *M. orygis*. This suggests the large repeat found may not be causing disruption, especially as all matches with bases inside the VRI region have the VRI, except for *M. bovis*. The partial Romanian record does not have bases in the VRI, however it does have its own shortly after starting. Of note is that CS1 has a missing glycine inside the VRI region.

3.4.4 kstD

The kstD VRIP causes the CS1 variant to split in three parts, as shown in Figure 28. The middle section is reversed in direction to the others and overlaps with the N-terminal section. 14 amino acids at the end of the N-terminal section are affected with mainly non-conservative changes, however the C-terminus of the middle section is identical to H37Rv despite the N-terminus being entirely dissimilar. A similar truncation is seen in other strains (a UK strain, NITR204, and GM1503), but the C-terminus of the N-terminal section is not similar to any other strain, and the middle section shows multiple deletions in respect to other strains (Figure 32). The strains closest to CS1 change with a significant number of *M. canettii* strains and a Beijing-family strain from Fujian (Figures 29-31).

Figure 28: A comparison of kstD in H37Rv (NP_218054.1) and CS1 (QEX91420.1, QEX91421.1, and QEX91422.1). QEX91421.1 is

the protein accession of a hypothetical protein, and this gene is in the reverse direction (the other three are forward). This VRIP is marked by repeated base changes and short 1-8bp insertion-deletions.

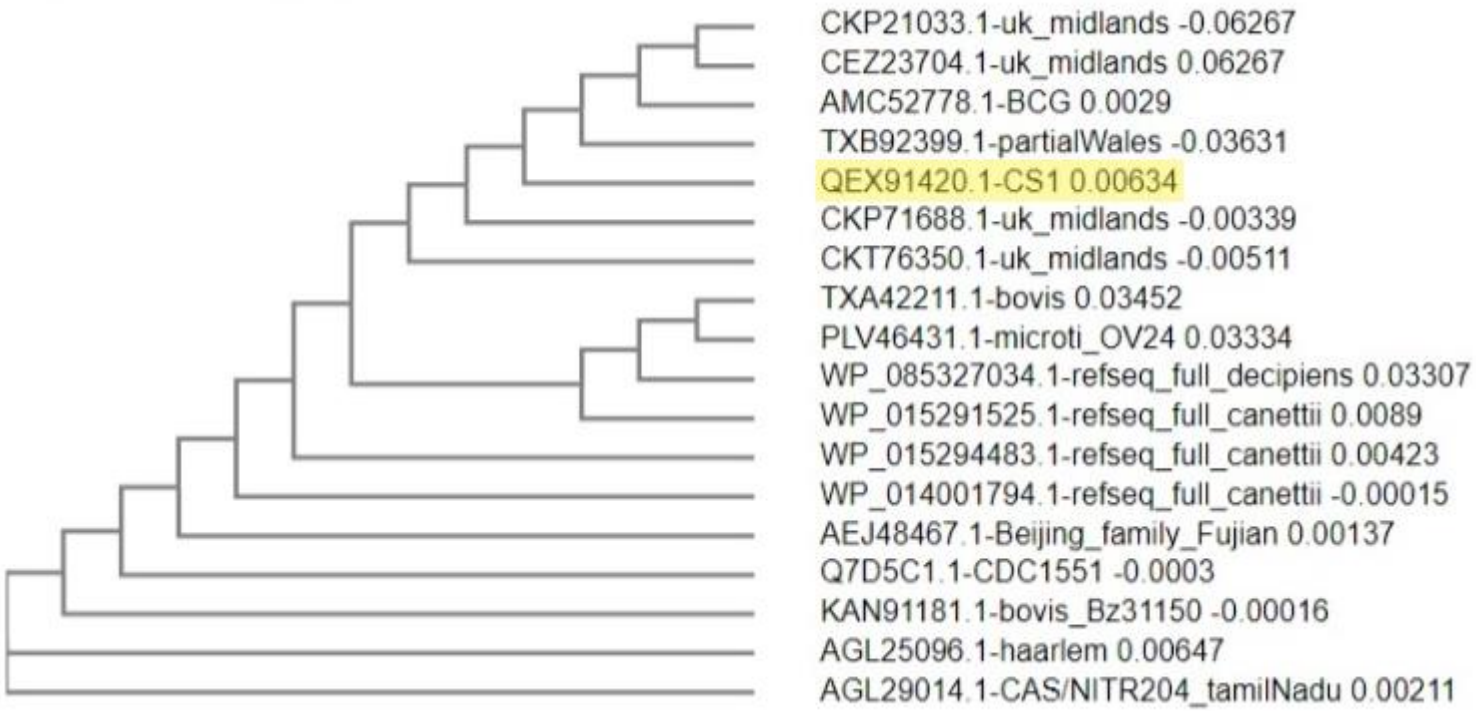

Figure 29: A phylogenetic tree from Clustal Omega showing the relatedness of matches for kstD1.0. Strains from the UK appear alongside a partial record from Wales. BCG, CDC1551, a Beijing-family strain from Fujian, NITR204 (of interest due to having a similar accession to Haarlem, appearing alongside it frequently, and appearing to be significant in its own right), and Haarlem are notable matches. *M. canettii, M. microti, M. decipiens,* and *M. bovis* are represented as well. CS1 does not have a close match but appears in a cluster with 4 UK strains, the partial record from Wales, and the vaccine strain BCG.

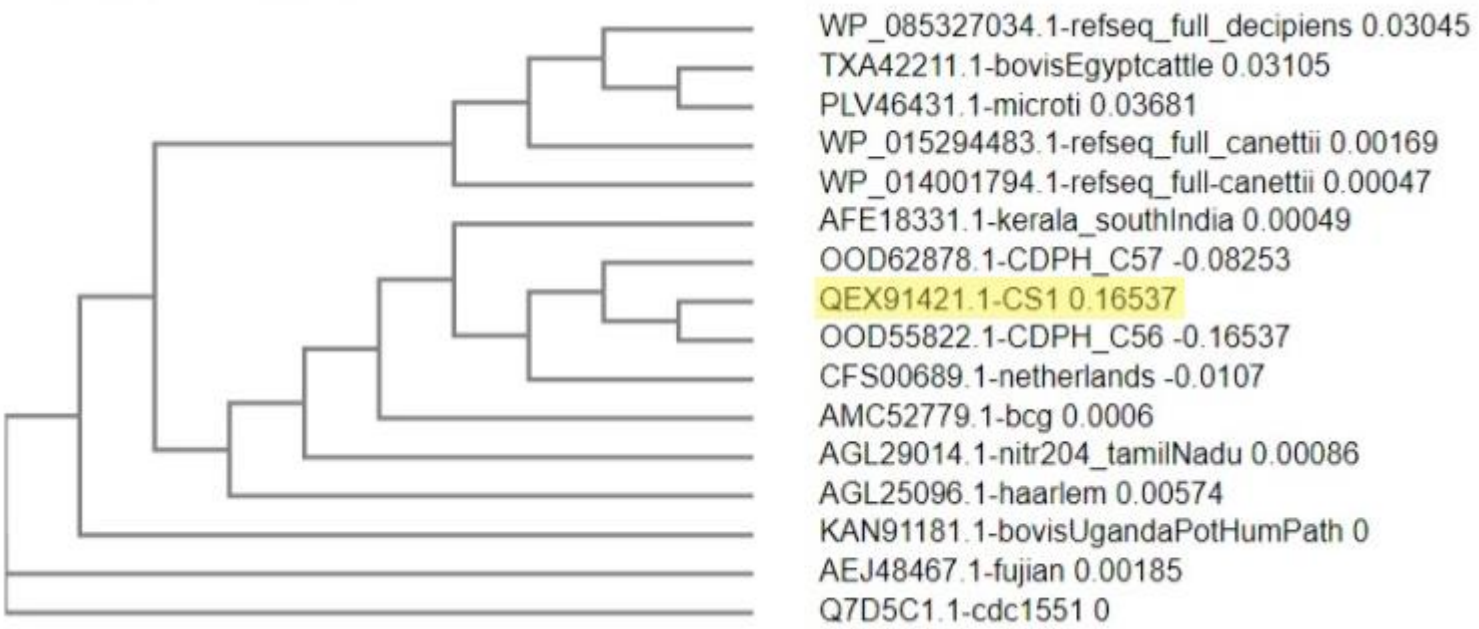

Figure 30: A phylogenetic tree from Clustal Omega showing the relatedness of matches for the hypothetical protein dubbed kstD1.5. BCG, CDC1551, a Beijing-family strain from Fujian, NITR204, and Haarlem are notable matches. *M. canettii, M. microti, M. decipiens,* and *M. bovis* (including a potentially human-infecting strain) are also represented. The closest match to CS1 is CDPH_C56.

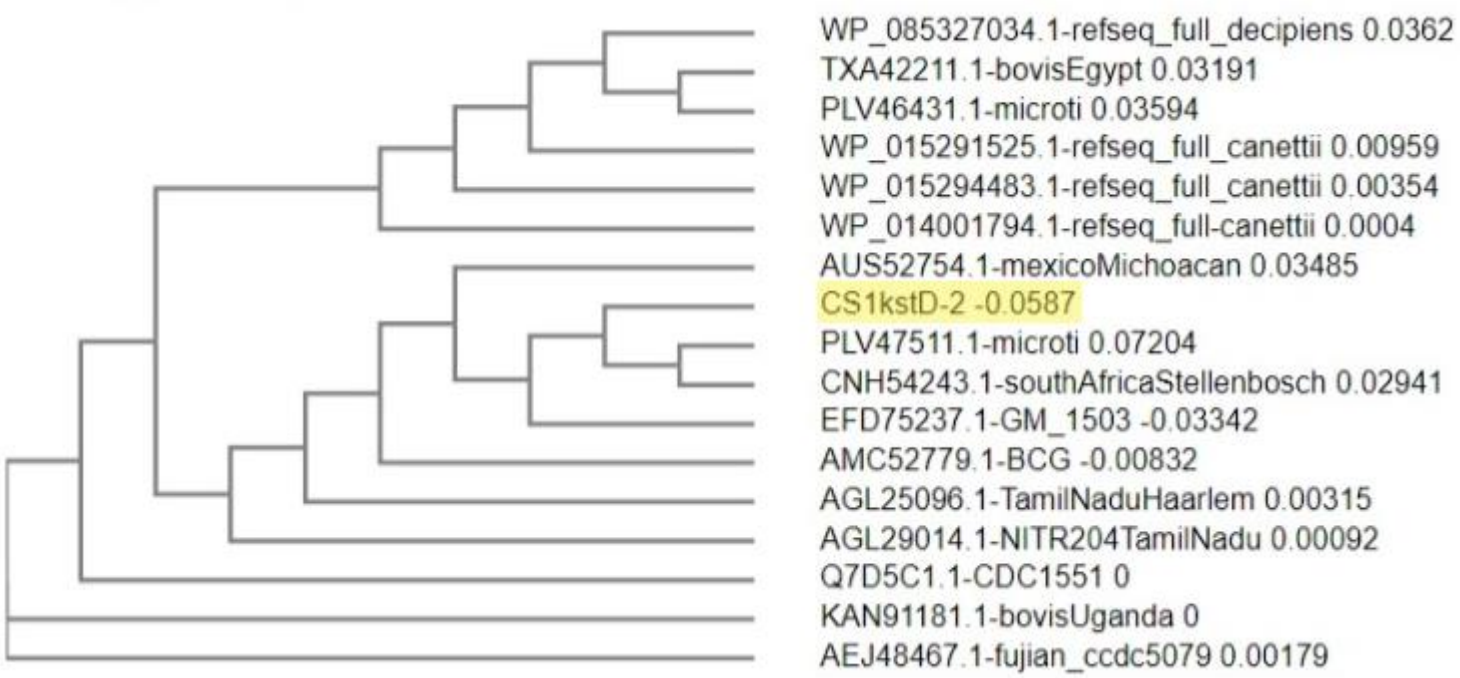

Figure 31: A phylogenetic tree from Clustal Omega showing the relatedness of matches for kstD2.0. BCG, CDC1551, a Beijing-family strain from Fujian, NITR204, and Haarlem are notable matches. *M. canettii, M. microti, M. decipiens,* and *M. bovis* are represented as matches. CS1 does not have a close match but appears in a cluster *M. microti*, the strain from Stellenbosch, GM_1503, and the Mexican strain.

```
CKP21033.1-uk_midlands          ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPGGRRWSRNRSTRASLVPTWPGWSPRMARFRSMWL----------
TXA42211.1-bovis                ---MSAQEYDVVVVGSGGAGMVAALTAAHRGLSTIVIEKAPHFGGSTARSGGGVWIPNNEVLKRDGVKDTPEAARTYLHGIIGDWVEPERIDTYLERGPEMLSFVLKHTPLKMCWVPRYSDVYPESPGGRA----------
PLV46431.1-microti_OV24         ---MTAQEYDVVVVGSGGAGMVAALTAAHRGLSTVVLEKAAHFGGSTARSGGGVWIPNNEVLKRDGVRDTPEAARTYLHGIIGDWVEPERIDTYLDRGPEMLSFVLKHTPLKMCWVPRYSDVYPEAPGGRA----------
WP_015291525.1-refseq_full_canettii ---MTAQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
AMC52778.1-BCG                  ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
CKP71688.1-uk_midlands          ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRP----------
KAN91181.1-bovis_Bz31150        ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPGGRSIEPKPFNARKLGADMAGLEPAYGKVPLNMVHRGPEMLSFVLKHTP
WP_015294483.1-refseq_full_canettii ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
CEZ23704.1-uk_midlands          ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRP----------
QEX91420.1-CS1                  ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
TXB92399.1-partialWales         ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
WP_085327034.1-refseq_full_decipiens ---MTAQEFDVVVVGSGGAGMVAALTTAHRGLSTLVVEKAPHYGGSTARSGGGVWIPNNEVLKRYGVRDTPEAARAYLHGIVGDIVEPERIDTYLDRAPEMLSFVTKHTPLKMCWVPGYSDVYPEAPGGRPG----------
AGL25096.1-haarlem              ------------------------------------------------------------------------------------------MLSFVLKHTPLKMCWVPGYSDVYPEAPGGRXG----------
AGL29014.1-CAS/NITR204_tamilNadu ------------------------------------------------------------------------------------------MLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
WP_014001794.1-refseq_full_canettii ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
Q7D5C1.1-CDC1551                MFYMTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
AEJ48467.1-Beijing_family_Fujian MFYMTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
CKT76350.1-uk_midlands          ---MTVQEFDVVVVGSGGAGMVAALVAAHRGLSTVVVEKAPHYGGSTARSGGGVWIPNNEVLKRGVRDTPEAARTYLHGIVGEIVEPERIDAYLDRGPEMLSFVLKHTPLKMCWVPGYSDVYPEAPGGRPG----------
                                                                                                                                        ***** ********** *******;****

CKP21033.1-uk_midlands          ----------------------------------------------------------------------------------------------------------------------------------
TXA42211.1-bovis                ----------EGRSIEPKPFNARKLGPDEAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSLKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVVLNTALTDLYLEDGVVRGVYVRDSQAAESAEPRLIRARRGVILASGGFEHNEQMRVKYQRAPIT
PLV46431.1-microti_OV24         ----------EGRSIEPKPFNARKLGPDEAGLEPAYGKVPLNVVVMQQDYVRLNLLKRHPRGVLRSLKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVRLNTALTDLYVADGVVRGVYVRDTGAPESSEPQLIRARRGVILACGGFEHNEQMRVKYQRAPIT
WP_015291525.1-refseq_full_canettii ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSNEAEAAEPQLIRARCGVILACGGFENNEQMRIKYQRAPIT
AMC52778.1-BCG                  ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGYTSAIPTRRNPLSRS----------
CKP71688.1-uk_midlands          ----------GGRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDMWFRA----------
KAN91181.1-bovis_Bz31150        LKMCWVPGYSDVYPEAPGGRPGGRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
WP_015294483.1-refseq_full_canettii ----------GGRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTALTDLFVESGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
CEZ23704.1-uk_midlands          ----------GGRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
QEX91420.1-CS1                  ----------GGRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
TXB92399.1-partialWales         ----------GGRSIEPKPFNARKLGADMAGLEPAYGKV----------------------------------------------------------------------------------
WP_085327034.1-refseq_full_decipiens ----------GRSIEPKPFNARKLGADMPGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVATRTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTALTDLYVENGVVSGVYVRDTNASES-AEPTLIRARRGVILACGGFEHNEQMRVKYQRAPIT
AGL25096.1-haarlem              ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTALTDLFVENGVVSGVYVXDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
AGL29014.1-CAS/NITR204_tamilNadu ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEASAKPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
WP_014001794.1-refseq_full_canettii ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
Q7D5C1.1-CDC1551                ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
AEJ48467.1-Beijing_family_Fujian ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT
CKT76350.1-uk_midlands          ----------GRSIEPKPFNARKLGADMAGLEPAYGKVPLNVVVMQQDYVRLNQLKRHPRGVLRSMKVGARTMWAKATGKNLVGMGRALIGPLRIGLQRAGVPVELNTAFTDLFVENGVVSGVYVRDSHEAESAEPQLIRARRGVILACGGFEHNEQMRIKYQRAPIT

CKP21033.1-uk_midlands          ----------
TXA42211.1-bovis                TEWTVGAKANTGDGIIAAEKLGAALELMEDAWMGPTVPLVGAPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEYGQGPGPGENIPAWLVFDQQYRDRYIFAGLQPGQRIPRKWLESGVIIQAETLEELASKAGL--PVDEFLATVQRFNGFARTGIDEDYHRGESAYDRY
PLV46431.1-microti_OV24         TEWTVGAKANTGDGIIAGEKLGAALDVMEDAWMGPTVPLVGAPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEFGQGPGPGENIPAWLVFDQQYRDRYIFAGLQPGQRIPNKWMESGVIVRADTLEELAEKAGL--PVAEFTATVARFNGFARQGVDEDFHRGESAYDRY
WP_015291525.1-refseq_full_canettii TEWTVGACANTGDGIIAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPRRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHRGESAYDRY
AMC52778.1-BCG                  ----------
CKP71688.1-uk_midlands          ----------
KAN91181.1-bovis_Bz31150        TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHRGESAYDRY
WP_015294483.1-refseq_full_canettii TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPRRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHRGESAYDRY
CEZ23704.1-uk_midlands          TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPASSSRPIPLR----------SWPARPVYPRTNSLPPSSVSTHSPGPVSTRTTTAGKVPTIATTA-----TRATSPIRT-
QEX91420.1-CS1                  TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGECVETLDGGSEFVR----------G----------
TXB92399.1-partialWales         ----------
WP_085327034.1-refseq_full_decipiens TEWTVGAKANTGDGIIAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRVPRRWLDSGVIVAADTLGELAGKAGL--PADELTATIERFNAFARSGVDEDYQRGESAYDRY
AGL25096.1-haarlem              TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLXFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHR-----GES
AGL29014.1-CAS/NITR204_tamilNadu TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHRGESAYDRY
WP_014001794.1-refseq_full_canettii TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHR-----GES
Q7D5C1.1-CDC1551                TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHR-----GES
AGL29014.1-nitr204_tamilNadu    TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHR-----GES
AMC52779.1-bcg                  TEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFMNESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHR-----GES
OOD62878.1-CDPH_C57             ----------GPGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSG----------
QEX91421.1-CS1                  ----------MADLAEVRIGLVARVAVVAIVGTFPAVVVLVDTGPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGPLAVFAAVHMMTGFD
OOD55822.1-CDPH_C56             ----------GPGENIPAWLVFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSG----------
AGL25096.1-haarlem              TTEWTVGASANTGDGILAAEKLGAALDLMDDAWMGPTVPLVGKPWFALSERNSPGSIIVNMSGKRFM-NESMPYVEACHMMYGGEHGQGPGPGENIPAWLXFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGL--PADELTATVQRFNAFARSGVDEDYHR-----GES
CFS00689.1-netherlands          ----------MFDQRYRDRYIFAGLQPGQRIPSRWLDSGVIVQADTLAELAGKAGLPADELTATVQRFNAFARSGVDEDYHR-----GES
TXA42211.1-bovisEgyptcattle      TTEWTVGAKANTGDGILAAEKLGAALELMEDAWMGPTVPLVGAPWFALSERNSPGSIIVNMSGKRFM-NESMPYVEACHMMYGGEHGQGPGPGENIPAWLVFDQQYRDRYIFAGLQPGQRIPRKWLESGVIIQAETLEELASKAGLPVDEFLATVQRFNGFARTGIDEDYHR-----GES
PLV46431.1-microti              TTEWTVGAKANTGDGIIAGEKLGAALDVMEDAWMGPTVPLVGAPWFALSERNSPGSIIVNMSGKRFM-NESMPYVEACHMMYGGEFGQGPGPGENIPAWLVFDQQYRDRYIFAGLQPGQRIPNKWMESGVIVRADTLEELAEKAGLPVAEFTATVARFNGFARQGVDEDFHR-----GES
                                                                                                                                
```

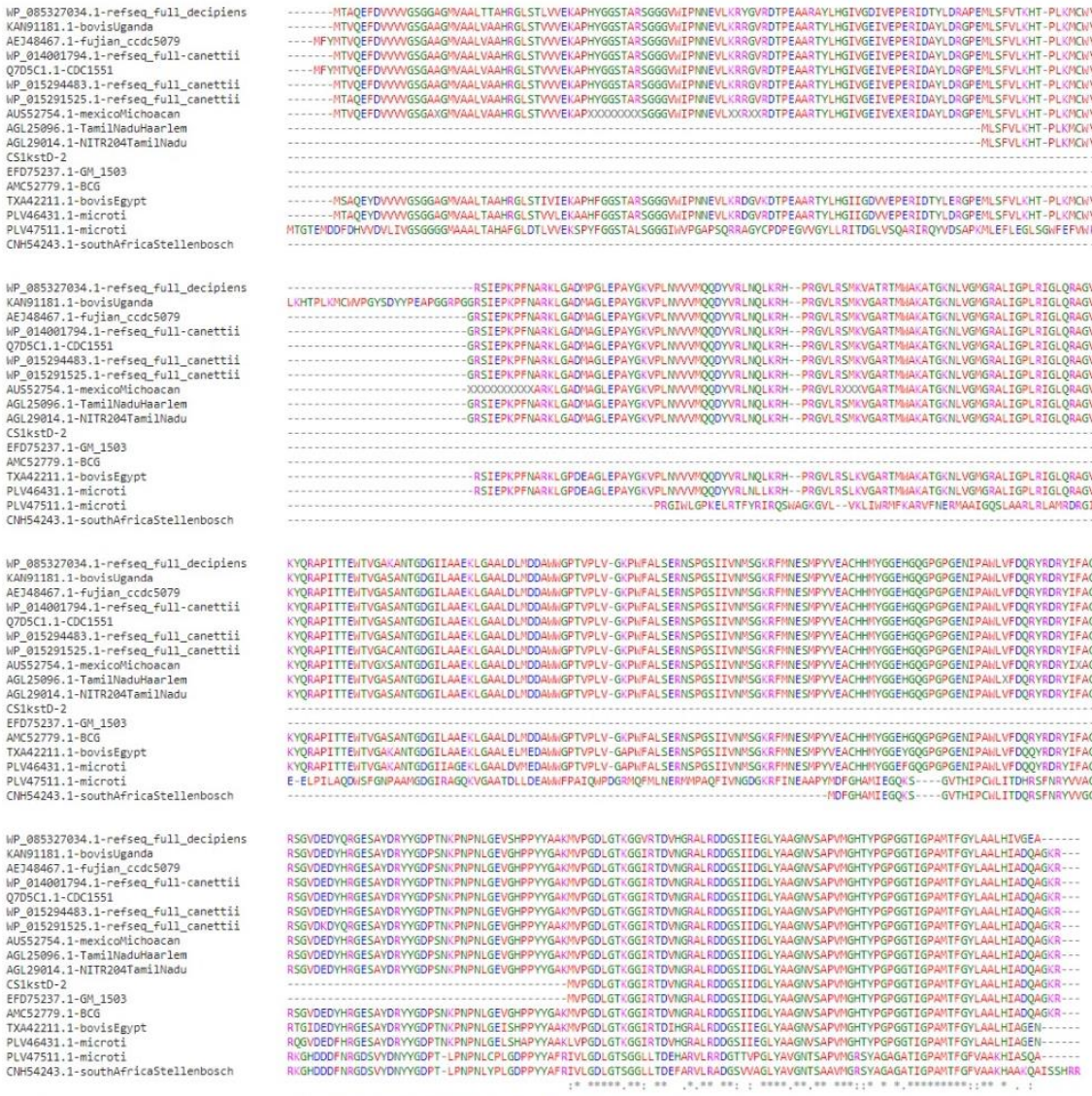

Figure 32: Filtered BLAST comparison results with for kstD1.0, the hypothetical protein kstD1.5, and kstD2.0. KstD1.0 and 1.5 contain the VRIP, however shortened kstD2 is of interest as well as kstD1.5. KstD2 mainly matched with far larger sequences, although lengths varied and one strain (GM_1503) was similarly truncated. The hypothetical appears to match non-truncated proteins as well, however it was notable that BCG and Haarlem were also shorter than expected, while CDC1551 was far longer. For kstD1, neither CDC1551 nor Haarlem share the early truncation of CS1, while BCG and a UK strain were much shorter. No matches had the exact VRIP as CS1. The level of variation found is surprising as malfunction in this gene is associated with failure to grow *in vivo*.

3.4.5 manB

The manB VRIP is 36 amino acids long with a “bookended” 2-residue deletion and 2-residue insertion (Figure 33). This VRIP has only 3 residues identical to H37Rv, and another 9 of some level of complementarity. The number of non-conserved residues is unusual for their position in the middle of the protein. manB is notable for being the only alignment with significant conservation over every strain for most places, however just under half of the VRIP does not match any other strain, and there is more variation in this area among other strains (Figure 35). There is a variety of matching strains, including 4 non-human strains and *M. africanum* (Figure 34).

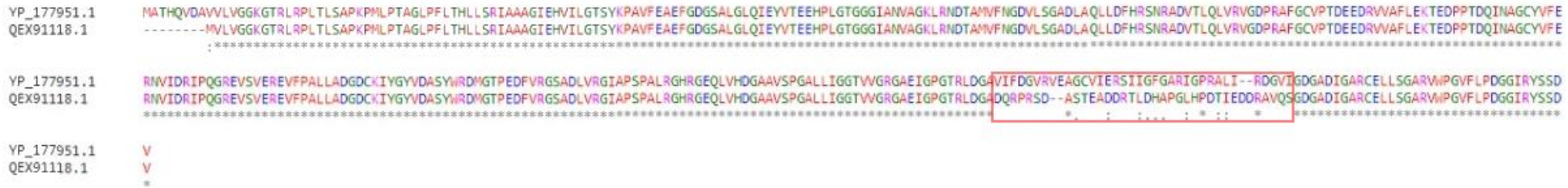

Figure 33: A comparison of manB in H37Rv (YP_177951.1) and CS1 (QEX91118.1). It is unclear why QEX91118.1 starts later than YP_177951.1 when they are identical at the nucleotide level, this could be an annotation error. Some of the amino acids in the VRI are complimentary. None of the insertion-deletions cause a stop-changing frameshift – in fact the two proteins are the same size other than the first 8 amino acids.

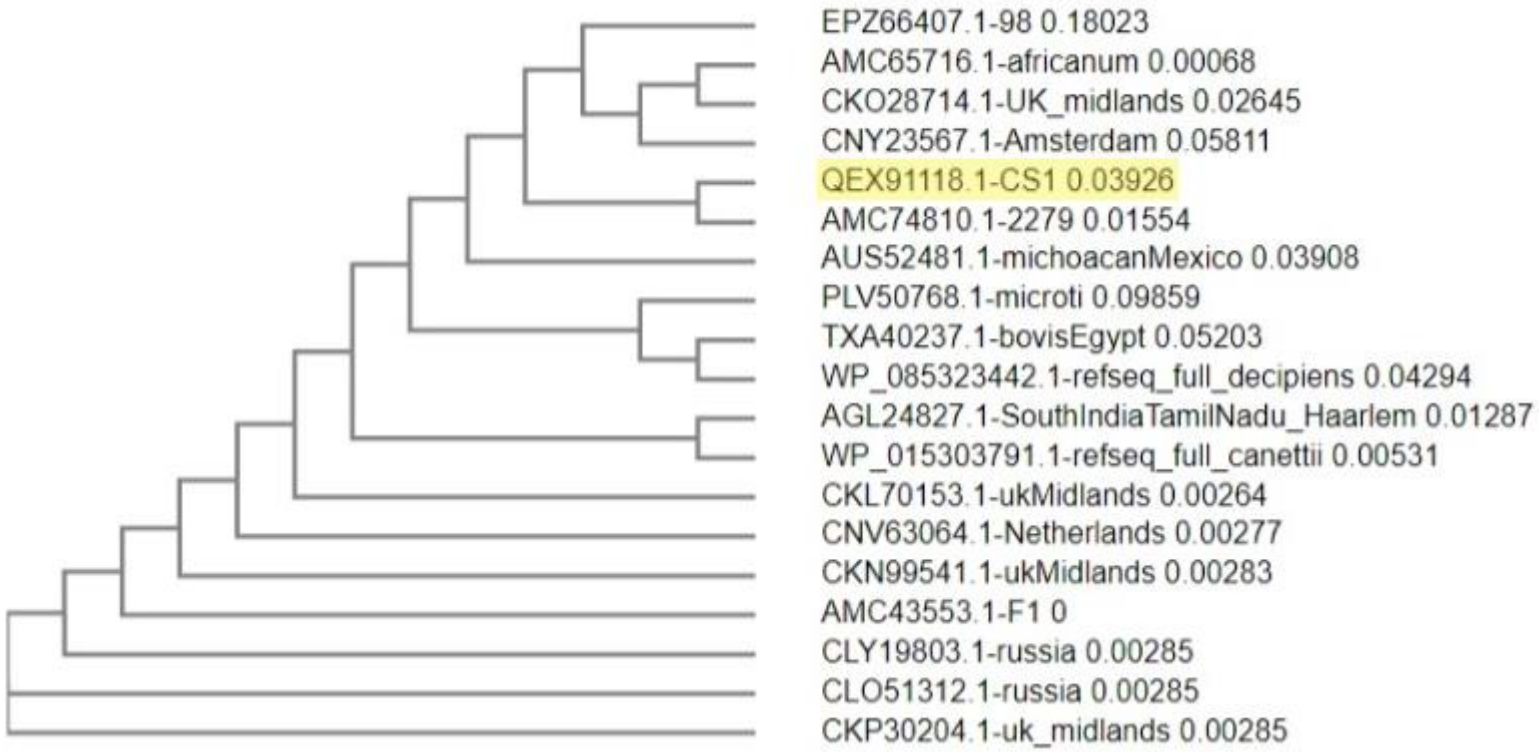

Figure 34: A phylogenetic tree from Clustal Omega showing the relatedness of matches for manB. Drug-resistant strains from China, Kazakhstan, and Thailand appear, with the latter being from the Beijing strain family. Notable matches include *M. bovis*, *M. canettii*, *M. microti*, *M. africanum*, F1, and Haarlem. Strain 2279 is the closest match to CS1.

Figure 35: Filtered BLAST comparison results for manB. *M. africanum*, strain and 2279, and Amsterdam have notable variation around CS1's VRI, but there are no exact matches for 16 bases nor for both the positioning and size of the deletion. Also of interest is that strain 98 is truncated before CS1's VRI, meaning the hexapeptide repeats forming the beta helix are missing.

3.4.6 VRIP Summary

The five VRIPs shown above are all characterised by amino acid and nucleotide changes when compared to H37Rv. VRIPs resulted in annotation splits in three cases, which was seen in a few other strains with some variations. For those which were not split, PE_PGRS17 had strains which appeared to have identical sequences, however this is uncertain as most of the matches were partial records. No identical matches were seen for manB, although variation was observed in the same region as the CS1 VRIP. For all five VRIPs, a variety of human and non-human strains were observed, with several candidates from the Beijing family as well as strains known for drug resistance and virulence. There is a possibility that the latter two candidates appear only due to sample bias, as these strains are the most likely to have been sequenced than slow-growing and less virulent strains.

Chapter 4

Discussion

4.1 Closing the CS1 Genome

While VRIPs could be signs of poor-quality data, all genomic data used to create the CS1 draft is of good quality overall, as is the quality of the draft itself. However, there are patches of poor quality and poor Illumina coverage which would benefit from further polishing. Nevertheless, the genome is high enough quality to be considered closed. This was determined by analysis of several factors (number of contiguous segments or contigs, coverage, error rate, core gene presence, and parameter sensibility) [46-48]. Contigs are pieces of an assembled genome made up of many reads, often creating gaps between contigs, meaning more complete genome assembly will have less contigs and fewer gaps [46; 48]. As PacBio was used for CS1, one single contig was produced, so the number of contigs is not an issue, and the use of Illumina to confirm the Pacbio sequence helps to improve certainty in the closed genome [46]. While 15% of the genome has coverage less than 7 reads, the majority of those regions are identical to H37Rv and so pose little concern. The error rate is low at 0.0072, and while this could be improved, the use of H37Rv as a comparison while noting local error values for SNPs increases reliability. All core genes are present, and parameters like repeat size, repeat number and percentage, GC percentage, and genome size are all within range for *Mtb* strains. It is difficult to achieve a true “closed” genome, with finished and near-finished genomes varying greatly in quality and number of coverage gaps [47; 48]. The CS1 genome is well within the acceptable range.

A related issue is the case of genome annotation. Between the 2020 Steele project [24] and this current work, the prokaryote annotation software in NCBI was updated, changing the annotations

on the CS1 genome. This changed the VRIPs, as by definition these were dense mutation clusters inside protein annotations. Some became defunct and did not have a locus tag, and presumably some variable regions ignored in the Steele project now have annotations around them. Annotation errors are common and can be widespread [49], and have been found in *Mtb* [50]. While automatic annotation is constantly improving with better algorithms and larger databases, automated with manual annotation ("curated annotation") remains the gold standard [25]. For CS1 VRIPs, the best solution would be to find all variable regions regardless of whether they appear inside an annotation.

4.2 Genetic Features of CS1

4.2.1 Repeat regions

The repeat analysis in WinMasker and I_r showed the CS1 repeats were within the expected range of other *M. tuberculosis* strains. Unlike many repeat-masking tools, neither Winmasker nor I_r rely on databases of known repeat regions [25; 36; 37], making the identification of unknown repeat regions easier and more accurate [25; 51]. CS1 has a larger genome than H37Rv, and most of the extra bases are repeats, which is of interest as many virulence genes contain repeats. Repeat regions are important components of the genome, involved in everything from mobile elements impacting gene regulation and CRISPRs, to protein structure and location signalling, to physical and informational organisation [51; 52]. The importance of mobile elements and CRISPRs in particular is discussed below, but both play integral roles in cellular survival and fitness. The importance of repeats is true in any genome, for example it is estimated that two-thirds of the human genome is made up of repeats, some of these being mobile elements, physical elements such as telomeres, and regulatory regions [51]. Similar roles are seen in prokaryotes, with many repeat regions part of regulatory regions and mobile elements [53]. As repeats play important genomic roles, defining their sequences and describing

their biological function is valuable information. Accurate representation of repeat sequences is best obtained by high-accuracy Illumina reads, while lower-base-accuracy PacBio sequencing can provide repeat size and location [54].

The size of the repeats were the same between CS1 and H37Rv, with most repeats being under 100 bases long. However, the “scatter pattern” of outlier repeats above 100 bases varied between H37Rv and CS1 (Figure 15). CS1 seems to have smaller outliers with one notable exception (discussed shortly), and a more even transition between dense clustering closer to 100 bases and more spacing as repeat size increases.

Outlier repeats are here defined as larger than 100 bases, for two reasons. Firstly, most repeats were 99 bases or smaller. Secondly, repeats larger than 100 bases would take up half the size of the average Illumina read in this sample, potentially causing slippage, issues with duplicated reads (as seen in the Qualimap output), and problems with assembly. Low quality regions and gaps could be expected inside large repeats due to ambiguity [55], and both low quality and gaps were seen to some extent, particularly around the opening point of the circular genome. Repeats often have subtle differences across their length, causing apparent base mismatches due to being mapped to the wrong position [55], and this was also seen in the CS1 Illumina reads. A sign of repeat-related duplication could be read-stacking, which was seen around many repeat regions, with sometimes hundreds of reads stacked precisely on each other and disagreements between reads. The disagreements may not be of any significance, as it is known that more read coverage often results in more disagreement simply due to the number of reads [48], and it could also be due to repeat region subtleties [55], however it could also be an indicator of low-quality. Low quality regions appeared throughout the genome, including in repeat regions, however many of the large repeats had few disagreements between high-quality Illumina reads.

Most outlier repeats were between 100 and 200 bases long, and this range appeared to be of the best quality more consistently. It would be valuable to investigate quality for large repeats which occur inside VRIP genes such as PPE32. PE/PPE proteins are a repeat-rich protein family characterised by repetitive Proline-Glutamine (PE) or Proline-Proline-Glutamine (PPE) motifs at the N terminus which have roles in virulence and disease progression. PPE32, for example, plays a role in apoptosis of host macrophages as well as in the survival of intracellular *Mtb* through cytokine production [56]. If the repeats in genes like PPE32 have not caused VRIPs and there are no low-quality regions, those VRIPs could be considered verified. As such, an investigation into the quality of repeats in and around VRIPs would provide valuable insight into the genetic features of CS1.

While repeat regions between 100 and 200 bases were common, larger repeats (such as one in PE_PGRS4, at 513 bases long) also were found around VRIPs, although not exclusively. These could prove more difficult to resequence where low-quality is an issue, however H37Rv had larger repeats than CS1. The only exception to this was CS1's 1519-base repeat, which interrupted PE_PGRS57. This repeat was unable to be explored, and due to mis-annotations, quality and coverage issues, and genes affected by the site, it would be worthy of further investigation. The site of the repeat appears as a truncated hypothetical protein (F6W99_02187, which was found to match the genome position PE_PGRS57 in H37Rv, as well as its protein sequence with a BLAST error value of 0.1) followed by a 22-repeat CRISPR mis-annotation, followed by an annotation (F6W99_02188) which runs into fadD19 and bears similarity to fadD19 in *M. africanum*. This repeat has read-stacking as previously described, as well as sections of low-coverage and poor-quality reads. The repeat region contains multiple repeats backing onto each other, however none were found to be CRISPRs.

4.2.2 CRISPR regions

While the 1519-base repeat was not found to contain a CRISPR region, CRISPRs are themselves of interest. Firstly, the 36-base Direct Repeat (DR) region in Mtb is a CRISPR region known to have mobile elements inside it, which is of note for virulence and strain typing [23] [57]. Secondly, CRISPR-Cas systems can be linked to virulence [57; 58]. Due to the link with virulence through mobile elements, the locations of CS1's CRISPR regions are of significance. In the future, comparisons between clinically-significant and less-virulent strains could lead to more insight into the role of CRISPRs in tuberculosis.

4.2.3 Mobile Elements: IS6110 regions

Mobile elements are found throughout *Mtb* genomes, including in CRISPR regions. The movement of mobile elements is likely the cause behind some of the apparent large insertion-deletions between CS1 and H37Rv shown in Figure 17. IS6110 elements in particular have been shown to influence gene regulation, by interrupting genes [59], facilitating deletions [60], and increasing gene expression [61]. IS6110 insertions have been linked to drug resistance and extrapulmonary tuberculosis [61]. Insertion sequences like IS6110 are particularly relevant in CS1, as IS6110 is found in H37Rv flanking the region deleted in both the "Beijing" RD152 and DS6^{Quebec} deletions (Figure 2) [1]. IS6110 create "fingerprints" which can be used to differentiate strain families, for example Beijing strains have similar IS6110 patterns, which is even more interesting when considering the regulatory effect on genes flanked by IS6110 [23]. Beggs, Eisenach, and Cave found IS6110 regions flanking a variety of genes in two clinically-significant Beijing strains, including 9 PE/PPE genes, and these regions showed differences to H37Rv [23]. Further work into IS6110, transposons, and other mobile elements is recommended to further enhance understanding of CS1 virulence. This would also

clear much of the list of large insertion-deletions seen in Figure 17.

4.2.4 VRIPs

VRIPs and similar regions are fascinating, as they could represent sequencing errors, non-coding regions accumulating mutations, or an adaptive advantage for CS1. The verified VRIPs appeared inside genes linked to survival and virulence. The five which were analysed in detail were chosen due to the potential of being adaptive advantages.

Cas10/Csm1 is the nuclease of the *Mtb* CRISPR-Cas system. *Mtb* has a type III-A CRISPR system which protects it from viral attack, particularly against double-stranded DNA bacteriophages [62; 63]. Cas10 has two domains, one which binds and cleaves DNA and another for synthesising cyclic oligoadenylate [64]. Cyclic oligoadenylates act as a signaller for Csm6, which indiscriminately degrades RNA [62; 63]. Interestingly, the CS1 cas10 VRIP causes a split which would be consistent with the two domains being separated. Furthermore, cas10 is potentially implicated in drug resistance, while others in the CRISPR-cas system have been linked to virulence [58; 65]. The abundance of CRISPR-cas proteins seems inversely linked to virulence and growth speed in Lineage 7 (known to be less virulent and slower growing) [65], while mutations and deletions appear in Lineage 2 [58], however much of the research focuses on Cas1 so the specific role of cas10 in virulence and growth speed is unknown. However, the CS1 cas10 VRIP has the potential to impact drug resistance, growth speed, and virulence.

LprN, also known as mce4E, is an immunomodulating lipoprotein involved in the invasion of host cells by *Mtb* [66]. It is highly immunogenic but was found to remove immune protection against *Mtb* when trialled as a vaccine candidate, with mice suffering increased tuberculosis-related tissue damage [66; 67]. LprN has a high rate of polymorphisms, particularly linked to drug resistant

strains but also to survival in general [68]. Differences have also been noted in lprN and other mce4 proteins between the non-pathogenic *Mycobacterium smegmatis* and H37Rv [69]. The CS1 LprN VRIP causes a split which also appears to split domains, but the significance is unknown as the mechanism of action of lprN is not clear.

PE_PGRS17 is part of the PE/PPE gene family, which are poorly characterised but involved in virulence. As one of two analysed VRIPs to not cause a split, the VRIP could cause an adaptive advantage. Certainly, this gene has great variation, with SNPs and insertion-deletions similar to the CS1 VRIP [70]. PE_PGRS17 activates and matures host dendritic cells, and other PE_PGRS proteins are known to be involved in apoptosis of this immune cell type [70; 71]. It also is antigenic and triggers cytokines from T- and dendritic cells during infection and is upregulated eight-fold in brain microvascular epithelium invasions [72]. PE_PGRS17 promotes T-cell proliferation [71], and is involved in *Mtb* persistence *in vivo* [73]. It increases survival in *M. smegmatis* while promoting host cell death [74]. Alphafold [75] structures (Figure 36) show a change in conformation between H37Rv and CS1 due to the VRIP, and these structures could assist in discovering the mechanism of action for PE_PGRS17.

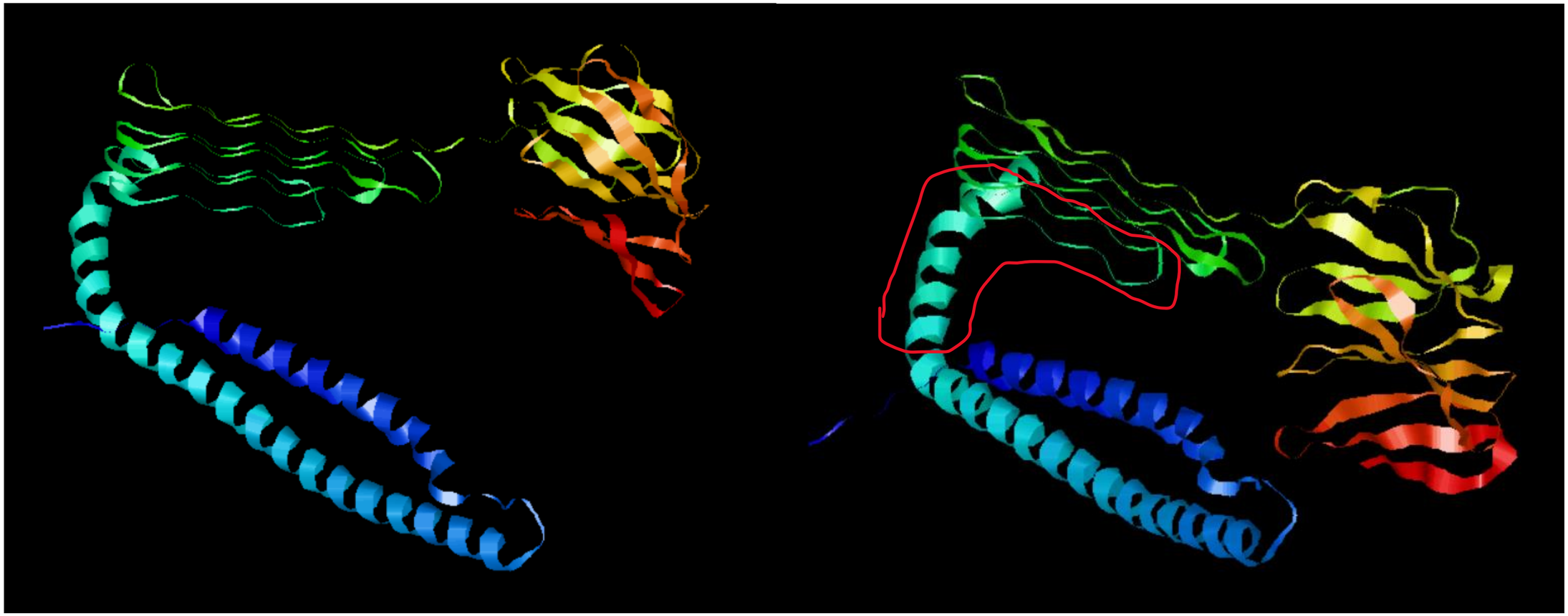

Figure 36: Alphafold predictions viewed in RasMol for PE_PGRS17, with H37Rv on the left and CS1 on the right. The VRIP is indicated in red, affecting residues 99-140. The gap between the N terminal alpha helix and the C terminal beta-sheet cluster in smaller in CS1 than H37Rv. The certainty for both structural predictions was 92%. Thanks to Daniel Schipper for implementing AlphaFold for these images.

KstD is crucial for utilising stored cholesterol as a carbon source [76], and *Mtb* cannot multiply inside macrophages without it [77]. Stored cholesterol affects cell wall permeability, and was found to reduce rifampin uptake [76]. Fitness and survival are reduced when KstD is inactivated [76], particularly for survival inside dendritic cells (which can act as an alternative to macrophages as a host) [78]. While the analysis showed variation between strains, this is not a protein to mutate lightly due to the catastrophic effect if it is deactivated. The way CS1 kstD was split into three pieces, with the middle piece in the reverse direction, was not seen in other strains, and there were unique regions caused by the VRIP. As CS1 infects and grows successfully, the changes in kstD could potentially provide a fitness advantage, and if not, this could provide valuable insight into the diversity and action of kstD in tuberculosis.

The final gene analysed was manB, a mannophosphomutase on the lipoarabinomannan pathway. Lipoarabinomannan binds to host macrophages [79], and is important in host macrophage survival and *Mtb* virulence [80]. For example, overexpression of manB in *M. smegmatis* causes increased lipoarabinomannan levels and increased association with human macrophages [81]. The CS1 manB VRIP, like for PE_PGRS17, does not cause any splits, however it does produce a unique sequence inside the hexapeptide repeat, which is predicted to change the structure of that region and therefore could impact function [24].

These five proteins provide some clues into CS1 virulence for further investigation, in particular PE_PGRS17 and manB have predicted structural changes caused by the VRIP. This points towards the VRIPs in these cases providing a fitness advantage, as while the effect could be neutral, the VRIPs occur in the middle of domains and impact overall structure rather than in less functionally-important regions (such as the ends, which had similar non-conserved residues in other VRIPs, for example kstD

and cas10). However, there is the possibility that apparently-unique regions and features, such as in PE_PGRS17 and the reversed domain in kstD, are shared by slower-growing and less virulent strains which have not been sequenced yet. Clinically-significant strains, such as those displaying drug resistance, are more likely to be prioritised for sequencing, so the appearance of many of these strains as matches to VRIP-affected genes is not necessarily an indicator of a virulence-enhancing effect. Of more interest were unusual results, such as *M. africanum* (an ancient member of the MTC causing disease in North Africa [82]), *M. canettii* (a rare smooth variant found in humans around the Horn of Africa [83]), *M. bovis* (found in cows, with occasional human infection [84]), *M. microti* (found in small rodents like voles as well as cats [85]), *M. orygis* (found mainly in oryx with occasional human infection [86]), and the vaccine strain BCG. These results were not often a close match for CS1 with the exception of *M. africanum*, but their appearance could not be attributed to an abundance of sequenced strains, and many are slower-growing.

Another note on the VRIP analyses is that not all strains had completely-sequenced proteins (indicated in those figures with "partial"), which particularly impacted PE_PGRS17 results. Partial protein records do not have the full amino acid sequence and may contain errors. It is also possible that annotation error affected some of the proteins, for example the middle annotation of CS1 kstD is largely dissimilar to any other strain, and while the other two segments are forward-direction genes, this one is not. This calls into question whether this segment of CS1 kstD is shorter than the annotation, and what protein it produces in the context of a split-domain kstD.

The domain-splitting seen in Cas10, lprN, and kstD was shared by very few strains, however independent domains is the first step in domain shuffling. Domain shuffling involves copying a gene encoding a protein domain and insertion into a different locus,

leading to hybrid proteins where one domain is used by different proteins, a process which has led to new, sometimes crucial, proteins and is key to evolution [87]. Domain shuffling events have previously been observed in MTC genomes [88], which gives credibility to this observation.

Further investigation into VRIPs would be valuable for both CS1 and for *Mtb* research in general. Finding, verifying or resolving, and analysing VRIP genes (and perhaps also non-coding VRIP-like regions) not only gives clues into the cause of increased transmissibility in CS1, but also could help clarify the roles and mechanism of action for many *Mtb* genes and proteins, which currently are poorly understood. This in turn could contribute to the discovery of new drug targets and new drugs, particularly in the case of antibiotic persistent and resistant *Mtb*.

Chapter 5

Conclusions and Future Work

The aim of this project was to close the CS1 genome and investigate its high transmission rate through investigating VRIPs. CS1 makes up a large portion of New Zealand-origin tuberculosis cases, and a closed genome will itself help discover the cause of its high transmission rate. This research could also elucidate the advantageous traits of international strains of clinical significance and new drug targets. A variety of computational tools were used to investigate the draft genome quality and investigate potential non-genetic causes of VRIPs, as well as the effect of a selection of verified VRIPs.

Many VRIPs were not verifiable by this work, however the genome was found to be of good enough quality to be declared closed. A closed genome gives better certainty for strain-typing and research into *Mtb* genomes, not just CS1. Poor quality, low coverage, and high repeat regions have been found, which can be used to re-sequence these regions as needed, particularly where VRIPs and SNPs appear. The region around the genome opening would benefit from resequencing in particular. Knowledge about high repeat regions can also help with research on repeat elements related to virulence, such as mobile elements.

This project verified 13 VRIPs, which provides valuable information into the diversity of proteins in *Mtb*. Several were found to have potential virulence and survival effects, which could be individually investigated in the future. Further investigation into the diversity of *Mtb* proteins, especially in slow-growing and less virulent strains, would be useful in discovering protein mechanisms of action and effect on patient outcomes. Part of this would be to investigate the unverified VRIPs, ensuring quality and coverage is adequate in order to resolve or verify them. There are also likely to be VRIPs

unaccounted for due to annotation error, and variable regions outside coding proteins were out of scope for this project; finding and resolving both would provide valuable information.

Annotation errors were found, and in the future manual annotation curation would be valuable. Correcting the CRISPR mis-annotation in the 1519 base repeat would be one immediate step, and adding CRISPR annotations found in this project would be another. This will help researchers of CRISPRs and protein diversity to gather accurate results of the feature they are investigating.

Outside of CS1 and similar to the need to investigate protein diversity in less clinically-significant strains, mapping IS6110 regions in those strains can facilitate understanding of the role of IS6110 in virulence, protein expression, and gene deletion, as well as its history. A comparison of IS6110 in clinically-significant strains, animal strains, Lineage 7 strains, and MTC relatives such as *M. africanum* would assist in this purpose.

This work confirmed the genome of New Zealand *Mtb* strain CS1 is closed and gave insight into its genetic features and characteristics. Future work could focus on CS1 (investigating the verified CS1 VRIPS, the unverified and yet-to-be-found CS1 VRIPS, or resolving the identified poor-quality regions and annotation errors) or on the wider MTC (analysing virulent and less-virulent strains for VRIP-like elements and IS6110, investigating virulence and survival mechanisms of action for proteins with verified CS1 VRIPs), but will regardless have an impact on the understanding of virulence in CS1 and *Mtb*.

References

[1] Mulholland, C. V. (2019). *Mycobacterium tuberculosis strains in New Zealand: phylogeny and structural biology* thesis, University of Waikato, Hamilton, New Zealand.

[2] WHO. (2020). *Global tuberculosis report 2020.* World Health Organization, Geneva.

[3] Lawn, S. D., & Zumla, A. (2011). Tuberculosis. *The Lancet, 378*(9785), 57-72.

[4] Houben, E. N. G., Nguyen, L., & Pieters, J. (2020). Interaction of pathogenic mycobacteria with the host immune system. *Current Opinion in Microbiology, 9*(6), 76-85.

[5] Miller, L. G., Asch, S. M., Yu, E. I., Knowles, L., Gelberg, L., & Davidson, P. (2000). A Population-Based Survey of Tuberculosis Symptoms: How Atypical Are Atypical Presentations? *Clinical Infectious Diseases, 30*(2), 293-299.

[6] Houben, R. M. G. J., & Dodd, P. J. (2016). The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLOS Medicine, 13*(10), e1002152.

[7] WHO. (2019). *New Zealand Tuberculosis Profile 2019 (accessed 10th August 2021)*. Retrieved 10 August, 2021, from https://www.who.int/tb/country/data/download/en.

[8] ESR. (2019). *Tuberculosis in New Zealand Annual Report 2016.* The Institute of Environmental Science and Research Ltd (ESR). Porirua.

[9] Aung, H. L., & Devine, T. J. (2019). Reducing the burden of tuberculosis in the Māori, the Indigenous people of New Zealand. *The Lancet Global Health, 7*(7).

[10] WHO. (2018). BCG vaccines: WHO position paper – February 2018. *Weekly Epidemiological Record, 93*(8), 73-96.

[11] Roy, A., Eisenhut, M., Harris, R. J., Rodrigues, L. C., Sridhar, S., Habermann, S., Snell, L., Mangtani, P., Adetifa, I., Lalvani, A., & Abubakar, I. (2014). Effect of BCG vaccination against *Mycobacterium tuberculosis* infection in children: systematic review and meta-analysis. *BMJ : British Medical Journal, 349*, g4643.

[12] WHO. (2010). *Treatment of Tuberculosis: Guidelines. 4th edition.* World Health Organization, Geneva.

[13] WHO. (2017). *Guidelines for treatment of drug-susceptible tuberculosis and patient care, 2017 update.* World Health Organization, Geneva.

[14] Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., Trauner, A., Fenner, L., Rutaihwa, L., Borrell, S., Luo, T., Gao, Q., Kato-Maeda, M., Ballif, M., Egger, M., Macedo, R., Mardassi, H., Moreno, M., Vilanova, G. T., Fyfe, J., Globan, M., Thomas, J., Jamieson, F., Guthrie, J. L., Asante-Poku, A., Yeboah-Manu, D., Wampande, E., Ssengooba, W., Joloba, M., Boom, W. H., Basu, I., Bower, J., Saraiva, M.,

Vasconcellos, S. E. G., Suffys, P., Koch, A., Wilkinson, R., Gail-Bekker, L., Malla, B., Ley, S. D., Beck, H.-P., de Jong, B. C., Toit, K., Sanchez-Padilla, E., Bonnet, M., Gil-Brusola, A., Frank, M., Penlap Beng, V. N., Eisenach, K., Alani, I., Ndung'u, P. W., Revathi, G., Gehre, F., Akter, S., Ntoumi, F., Stewart-Isherwood, L., Ntinginya, N. E., Rachow, A., Hoelscher, M., Cirillo, D. M., Skenders, G., Hoffner, S., Bakonyte, D., Stakenas, P., Diel, R., Crudu, V., Moldovan, O., Al-Hajoj, S., Otero, L., Barletta, F., Carter, E. J., Diero, L., Supply, P., Comas, I., Niemann, S., & Gagneux, S. (2016). Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nature Genetics, 48*(12), 1535-1543.

[15] Littleton, J., Park, J., Thornley, C., Anderson, A., & Lawrence, J. (2008). Migrants and tuberculosis: analysing epidemiological data with ethnography. *Australian and New Zealand Journal of Public Health, 32*(2), 142-149.

[16] Park, J., Littleton, J., Chambers, A., & Chambers, K. (2011). Whakapapa in anthropological research on tuberculosis in the Pacific. *SITES: New Series 8*(2), 6-31.

[17] Mulholland, C. V., Shockey, A. C., Aung, H. L., Cursons, R. T., O'Toole, R. F., Gautam, S. S., Brites, D., Gagneux, S., Roberts, S. A., Karalus, N., Cook, G. M., Pepperell, C. S., & Arcus, V. L. (2019). Dispersal of Mycobacterium tuberculosis Driven by Historical European Trade in the South Pacific. *Frontiers in Microbiology, 10*(2778).

[18] Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., Blum, M. G. B., Rüsch-Gerdes, S., Mokrousov, I., Aleksic, E., Allix-Béguec, C., Antierens, A., Augustynowicz-Kopeć, E., Ballif, M., Barletta, F., Beck, H. P., Barry, C. E., Bonnet, M., Borroni, E., Campos-Herrero, I., Cirillo, D., Cox, H., Crowe, S., Crudu, V., Diel, R., Drobniewski, F., Fauville-Dufaux, M., Gagneux, S., Ghebremichael, S., Hanekom, M., Hoffner, S., Jiao, W.-w., Kalon, S., Kohl, T. A., Kontsevaya, I., Lillebæk, T., Maeda, S., Nikolayevskyy, V., Rasmussen, M., Rastogi, N., Samper, S., Sanchez-Padilla, E., Savic, B., Shamputa, I. C., Shen, A., Sng, L.-H., Stakenas, P., Toit, K., Varaine, F., Vukovic, D., Wahl, C., Warren, R., Supply, P., Niemann, S., & Wirth, T. (2015). Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. *Nature Genetics, 47*(3), 242-249.

[19] De Zoysa, R., Shoemack, P., Vaughan, R., & Vaughan, A. (2001). A prolonged outbreak of tuberculosis in the North Island. *New Zealand Public Health Report, 8*(1).

[20] Gautam, S. S., Mac Aogáin, M., Bower, J. E., Basu, I., & O'Toole, R. F. (2017). Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of Mycobacterium tuberculosis. *Infectious Diseases, 49*(9), 680-688.

[21] Stucki, D., & Gagneux, S. (2013). Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. *Tuberculosis (Edinburgh, Scotland), 93*(1), 30-39.
[22] Coscolla, M., & Gagneux, S. (2014). Consequences of genomic diversity in Mycobacterium tuberculosis. *Seminars in Immunology, 26*(6), 431-444.
[23] Beggs, M. L., Eisenach, K. D., & Cave, M. D. (2000). Mapping of IS6110 Insertion Sites in Two Epidemic Strains of *Mycobacterium tuberculosis*. *Journal of Clinical Microbiology, 38*(8), 2923-2928.
[24] Steele, M. (2020). Final Research Report for SCIEN313-19C on Bioinformatics Analysis from M. Tuberculosis Genomes; Submitted to Graham Saunders on 14th February 2020 by email. University of Waikato.
[25] Ejigu, G. F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology, 9*(9).
[26] Aung, H. L. (Compiler) (2019). *Complete genome sequence of a New Zealand Mycobacterium tuberculosis strain responsible for the ongoing transmission over the last 30 years*
[27] Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLOS ONE, 5*(6), e11147.
[28] Geneious Prime 2020.0.3. (https://www.geneious.com).
[29] Bushnell, B. (2020). BBMap.
[30] Andrews, S. (2019). FastQC: a quality control tool for high throughput sequence data.
[31] Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics, 32*(2), 292-294.
[32] Schipper, D. A. (2020). Matcher: Python Algorithm for Computing and Filtering Quality Scores of Assembled Reads by Base Position.
[33] Liao, P., Satten, G. A., & Hu, Y. J. (2017). PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genet Epidemiol, 41*(5), 375-387.
[34] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.
[35] Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics, 14*(2), 178-192.
[36] Morgulis, A., Gertz, E. M., Schäffer, A. A., & Agarwala, R. (2006). WindowMasker: window-based masker for sequenced genomes. *Bioinformatics, 22*(2), 134-141.

[37] Haubold, B., & Wiehe, T. (2006). How repetitive are genomes? *BMC bioinformatics, 7*, 541-541.
[38] Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., & Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics, 20*(4), 1085-1093.
[39] The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research, 45*(D1), D158-D169.
[40] Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research, 46*(W1), W246-W251.
[41] Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics, 12*(1), 402.
[42] Information, N. C. f. B. (accessed December 2020). The NCBI C++ Toolkit.
[43] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-10.
[44] NCBI. (2021). BLASTx [Online software]. https://blast.ncbi.nlm.nih.gov/Blast.cgi.
[45] Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., & Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Research, 38*(suppl_2), W695-W699.
[46] A reference standard for genome biology. (2018). *Nature Biotechnology, 36*(12), 1121-1121.
[47] Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T., & Salzberg, S. L. (2002). The value of complete microbial genome sequencing (you get what you pay for). *Journal of bacteriology, 184*(23), 6403-6405.
[48] Mardis, E., McPherson, J., Martienssen, R., Wilson, R. K., & McCombie, W. R. (2002). What is finished, and why does it matter. *Genome Res, 12*(5), 669-71.
[49] Lockwood, S., Brayton, K. A., Daily, J. A., & Broschat, S. L. (2019). Whole Proteome Clustering of 2,307 Proteobacterial Genomes Reveals Conserved Proteins and Significant Annotation Issues. *Frontiers in Microbiology, 10*(383).
[50] Johnston, J. M., Arcus, V. L., Morton, C. J., Parker, M. W., & Baker, E. N. (2003). Crystal structure of a putative methyltransferase from Mycobacterium tuberculosis: misannotation of a genome clarified by protein structural analysis. *Journal of bacteriology, 185*(14), 4057-4065.

[51] de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics, 7*(12), e1002384.
[52] Shapiro, J. A., & von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biol Rev Camb Philos Soc, 80*(2), 227-50.
[53] Delihas, N. (2011). Impact of small repeat sequences on bacterial genome evolution. *Genome biology and evolution, 3*, 959-973.
[54] Mangin, A., de Pontual, L., Tsai, Y.-C., Monteil, L., Nizon, M., Boisseau, P., Mercier, S., Ziegle, J., Harting, J., Heiner, C., Gourdon, G., & Tomé, S. (2021). Robust Detection of Somatic Mosaicism and Repeat Interruptions by Long-Read Targeted Sequencing in Myotonic Dystrophy Type 1. *International Journal of Molecular Sciences, 22*(5).
[55] Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics, 13*(1), 36-46.
[56] Deng, W., Yang, W., Zeng, J., Abdalla, A. E., & Xie, J. (2016). Mycobacterium tuberculosis PPE32 promotes cytokines production and host cell apoptosis through caspase cascade accompanying with enhanced ER stress response. *Oncotarget, 7*(41), 67347-67359.
[57] Singh, A., Gaur, M., Sharma, V., Khanna, P., Bothra, A., Bhaduri, A., Mondal, A. K., Dash, D., Singh, Y., & Misra, R. (2021). Comparative Genomic Analysis of Mycobacteriaceae Reveals Horizontal Gene Transfer-Mediated Evolution of the CRISPR-Cas System in the Mycobacterium tuberculosis Complex. *mSystems, 6*(1).
[58] Wei, W., Zhang, S., Fleming, J., Chen, Y., Li, Z., Fan, S., Liu, Y., Wang, W., Wang, T., Liu, Y., Ren, B., Wang, M., Jiao, J., Chen, Y., Zhou, Y., Zhou, Y., Gu, S., Zhang, X., Wan, L., Chen, T., Zhou, L., Chen, Y., Zhang, X.-E., Li, C., Zhang, H., & Bi, L. (2019). Mycobacterium tuberculosis type III-A CRISPR/Cas system crRNA and its maturation have atypical features. *The FASEB Journal, 33*(1), 1496-1509.
[59] Sampson, S. L., Warren, R. M., Richardson, M., van der Spuy, G. D., & van Helden, P. D. (1999). Disruption of coding regions by IS6110 insertion in Mycobacterium tuberculosis. *Tuber Lung Dis, 79*(6), 349-59.
[60] Sampson, S. L., Warren, R. M., Richardson, M., Victor, T. C., Jordaan, A. M., van der Spuy, G. D., & van Helden, P. D. (2003). IS6110-mediated deletion polymorphism in the direct repeat region of clinical isolates of Mycobacterium tuberculosis. *Journal of bacteriology, 185*(9), 2856-2866.
[61] Roychowdhury, T., Mandal, S., & Bhattacharya, A. (2015). Analysis of IS6110 insertion sites provide a glimpse into genome evolution of Mycobacterium tuberculosis. *Scientific Reports, 5*(1), 12567.

[62] Grüschow, S., Athukoralage, J. S., Graham, S., Hoogeboom, T., & White, M. F. (2019). Cyclic oligoadenylate signalling mediates Mycobacterium tuberculosis CRISPR defence. *Nucleic acids research, 47*(17), 9259-9270.
[63] Jiang, W., Samai, P., & Marraffini, Luciano A. (2016). Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. *Cell, 164*(4), 710-721.
[64] Athukoralage, J. S., Graham, S., Rouillon, C., Grüschow, S., Czekster, C. M., & White, M. F. (2020). The dynamic interplay of host and viral enzymes in type III CRISPR-mediated cyclic nucleotide signalling. *eLife, 9*, e55852.
[65] Yimer, S. A., Kalayou, S., Homberset, H., Birhanu, A. G., Riaz, T., Zegeye, E. D., Lutter, T., Abebe, M., Holm-Hansen, C., Aseffa, A., & Tønjum, T. (2020). Lineage-Specific Proteomic Signatures in the Mycobacterium tuberculosis Complex Reveal Differential Abundance of Proteins Involved in Virulence, DNA Repair, CRISPR-Cas, Bioenergetics and Lipid Metabolism. *Frontiers in microbiology, 11*, 550760-550760.
[66] Becker, K., & Sander, P. (2016). Mycobacterium tuberculosis lipoproteins in virulence and immunity – fighting with a double-edged sword. *FEBS Letters, 590*(21), 3800-3819.
[67] Pasricha, R., Saini, N. K., Rathor, N., Pathak, R., Sinha, R., Varma-Basil, M., Mishra, K., Brahmachari, V., & Bose, M. (2014). The Mycobacterium tuberculosis recombinant LprN protein of mce4 operon induces Th-1 type response deleterious to protection in mice. *Pathogens and Disease, 72*(3), 188-196.
[68] Pasricha, R., Chandolia, A., Ponnan, P., Saini, N. K., Sharma, S., Chopra, M., Basil, M. V., Brahmachari, V., & Bose, M. (2011). Single nucleotide polymorphism in the genes of mce1 and mce4 operons of Mycobacterium tuberculosis: analysis of clinical isolates and standard reference strains. *BMC Microbiology, 11*(1), 41.
[69] He, L., Zhou, X., Yin, X., Tian, L., Yang, L., Fan, K., & Zhao, D. (2014). Comparative Study of the Growth and Survival of Recombinant Mycobacterium smegmatis Expressing Mce4A and Mce4E from Mycobacterium bovis. *DNA and Cell Biology, 34*(2), 125-132.
[70] Meena, L. S. (2015). An overview to understand the role of PE_PGRS family proteins in Mycobacterium tuberculosis H37Rv and their potential as new drug targets. *Biotechnology and Applied Biochemistry, 62*(2), 145-153.
[71] Bansal, K., Elluru, S. R., Yeddula, N., Chaturvedi, R., Patil, S., Kaveri, S., Bayry, J., & Balaji, K. (2010). PE_PGRS Antigens of Mycobacterium tuberculosis Induce Maturation and Activation of Human Dendritic Cells. *Journal of immunology (Baltimore, Md. : 1950), 184*, 3495-504.
[72] Mukhopadhyay, S., & Balaji, K. N. (2011). The PE and PPE proteins of Mycobacterium tuberculosis. *Tuberculosis, 91*(5), 441-447.

[73] Li, W., Deng, W., & Xie, J. (2018). Expression and regulatory networks of Mycobacterium tuberculosis PE/PPE family antigens. *Journal of Cellular Physiology*.
[74] Chen, T., Zhao, Q., Li, W., & Xie, J. (2013). Mycobacterium tuberculosis PE_PGRS17 promotes the death of host cell and cytokines secretion via Erk kinase accompanying with enhanced survival of recombinant Mycobacterium smegmatis. *Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research, 33*(8), 452-458.
[75] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583-589.
[76] Brzostek, A., Pawelczyk, J., Rumijowska-Galewicz, A., Dziadek, B., & Dziadek, J. (2009). Mycobacterium tuberculosis is able to accumulate and utilize cholesterol. *Journal of bacteriology, 191*(21), 6584-6591.
[77] Brzezinska, M., Szulc, I., Brzostek, A., Klink, M., Kielbik, M., Sulowska, Z., Pawelczyk, J., & Dziadek, J. (2013). The role of 3-ketosteroid 1(2)-dehydrogenase in the pathogenicity of Mycobacterium tuberculosis. *BMC Microbiology, 13*(1), 43.
[78] Mendum, T. A., Wu, H., Kierzek, A. M., & Stewart, G. R. (2015). Lipid metabolism and Type VII secretion systems dominate the genome scale virulence profile of Mycobacterium tuberculosis in human dendritic cells. *BMC Genomics, 16*(1), 372.
[79] Schlesinger, L. S., Hull, S. R., & Kaufman, T. M. (1994). Binding of the terminal mannosyl units of lipoarabinomannan from a virulent strain of Mycobacterium tuberculosis to human macrophages. *J Immunol, 152*(8), 4070-9.
[80] Maiti, D., Bhattacharyya, A., & Basu, J. (2001). Lipoarabinomannan from Mycobacterium tuberculosis Promotes Macrophage Survival by Phosphorylating Bad through a Phosphatidylinositol 3-Kinase/Akt Pathway*. *Journal of Biological Chemistry, 276*(1), 329-333.
[81] McCarthy, T. R., Torrelles, J. B., MacFarlane, A. S., Katawczik, M., Kutzbach, B., DesJardin, L. E., Clegg, S., Goldberg, J. B., & Schlesinger, L. S. (2005). Overexpression of Mycobacterium tuberculosis manB, a phosphomannomutase that increases phosphatidylinositol mannoside biosynthesis in Mycobacterium smegmatis and mycobacterial association with human macrophages. *Molecular Microbiology, 58*(3), 774-790.

[82] Gehre, F. A., Martin, Otu, J. K., Sallah, N., Secka, O., Faal, T., Owiafe, P., Sutherland, J. S. A., Ifedayo M., Ota, Martin O., Kampmann, Beate, Corrah, T., & de Jong, B. C. (2013). Immunogenic Mycobacterium africanum Strains Associated with Ongoing Transmission in The Gambia. *Emerging Infectious Diseases, 19*(10), 1599-1605.

[83] Supply, P., & Brosch, R. (2017). The Biology and Epidemiology of Mycobacterium canettii. In S. Gagneux (Ed.), *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control* (pp. 27-41). Cham: Springer International Publishing.

[84] van Soolingen, D., de Haas, P. E., Haagsma, J., Eger, T., Hermans, P. W., Ritacco, V., Alito, A., & van Embden, J. D. (1994). Use of various genetic markers in differentiation of Mycobacterium bovis strains from animals and humans and for studying epidemiology of bovine tuberculosis. *J Clin Microbiol, 32*(10), 2425-33.

[85] Smith, N. H., Crawshaw, T., Parry, J., & Birtles, R. J. (2009). Mycobacterium microti: More diverse than previously thought. *J Clin Microbiol, 47*(8), 2551-9.

[86] van Ingen, J. R., Z Mulder, A.,Boeree, M. J.,Simeone, R., Brosch, R.,van Soolingen, D. (2012). Characterization of Mycobacterium orygis as M. tuberculosis Complex Subspecies. *Emerging Infectious Diseases, 18*(4), 653-655.

[87] de Souza, S. J. (2012). Domain shuffling and the increasing complexity of biological networks. *BioEssays, 34*(8), 655-657.

[88] Płociński, P., Macios, M., Houghton, J., Niemiec, E., Płocińska, R., Brzostek, A., Słomka, M., Dziadek, J., Young, D., & Dziembowski, A. (2019). Proteomic and transcriptomic experiments reveal an essential role of RNA degradosome complexes in shaping the transcriptome of Mycobacterium tuberculosis. *Nucleic Acids Research, 47*(11), 5892-5905.