Mycobacterium tuberculosis strains in New Zealand: phylogeny and structural biology

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Biological Sciences

at

The University of Waikato

by

Claire Vignette Mulholland

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2019

Abstract

Mycobacterium tuberculosis is an obligate human pathogen and is the primary causative agent of tuberculosis. New Zealand has a relatively low incidence of tuberculosis disease, however, Māori (the indigenous people of New Zealand) and Pacific People are disproportionally affected. Molecular typing shows that approximately two-thirds of *M. tuberculosis* isolates from New Zealand-born patients can be assigned to clusters of related strains. The largest *M. tuberculosis* cluster in New Zealand is known as the 'Rangipo' cluster and is predominantly found in Māori. This strain has been the source of several tuberculosis outbreaks over the last 30 years and anecdotal evidence suggests it may be particularly virulent. Two other large clusters, known as the 'Southern Cross' and 'Otara' clusters, most commonly occur in Pacific People.

Here, whole genome sequencing, phylogenetics and structural biology were used to investigate evolutionary origins and functional consequences of genomic diversity in New Zealand *M. tuberculosis* clusters, with a particular focus on the Rangipo strain. Analysis of Rangipo strain non-synonymous single nucleotide polymorphisms (nsSNPs) identified bacterial genetic factors that may contribute to the high transmissibility of this strain. The F_{420}-dependent oxidoreductase Rv2893 harbours a Rangipo-specific G72S mutation encoded by a nsSNP. H37Rv and Rangipo Rv2893 structures were solved and show the effect of this G72S mutation. Binding of the F_{420} cofactor was confirmed and characterised. SNP analyses also guided the optimisation of a diagnostic assay for rapid Rangipo strain classification at low cost and with high discriminatory power. Phylogenetic analyses revealed the Rangipo and Otara clusters belong to a larger *M. tuberculosis* clade of French/European origin that is also prevalent in indigenous populations in Canada. Molecular dating indicates dispersal of this clade to the South Pacific was driven by expanding European trade networks in the early 19th century and identifies host factors that have contributed to the dispersal and expansion of the Rangipo and Otara strains.

Overall, these results show that relatively recent changes in host ecology have likely played a crucial role in driving the success of the Rangipo strain in New Zealand and point to bacterial genetic factors that may influence its virulence and thereby also contribute to its prevalence.

Acknowledgements

First and foremost, I would like to thank my primary supervisor Prof Vic Arcus. Your ongoing support, expertise and guidance have been invaluable throughout this work. I admire your intellect and enthusiasm and I am most grateful to have been a student of yours. Thank you to my secondary supervisor Dr Ray Cursons, your encouragement and advice along the way has been greatly appreciated.

To the other members of the Arcus lab; Jo, Erica, Emma S, Emma A, Heng, Vikas, Chelsea, Tifany, Kirsty, Annmaree, Liz, Mitchell, Brooke and Daniel, thank you. I couldn't imagine getting through this without you and I am grateful for your support and friendship, it has made the days enjoyable and the challenging times more manageable. A special thank you to Judith for being a wonderful 'lab Mum'. Your kindness and encouragement are deeply appreciated, and of course your excellent baking too. Jo, Erica and Emma S, thank you for reading chapters of this thesis, you are all researchers I admire and am grateful to have worked with.

To Prof Greg Cook, Dr Htin Lin Aung, Dr Sally Roberts, Dr Noel Karalus and Duncan Thorpe, thank you for collaboration, support and enthusiasm. I have been fortunate to collaborate with such a great range of people; your varying expertise and help has been invaluable. Htin, I have greatly appreciated your encouragement and mentoring throughout my PhD and am grateful to have worked with you. To Caitlin Pepperell at the University of Madison-Wisconsin, thank you for giving me the opportunity to learn from you and your group and for your support. Your involvement and expertise have been instrumental to part of this work.

Thank you to the Maurice Wilkins Centre for my doctoral scholarship and other opportunities provided along the way, and also the Waikato Graduate Women Educational Trust for funding to support me through this research.

Last but not least, to my friends and family, thank you for your continued love and support throughout my studies. Grandma and Grandad, thank you for backing me along this journey and for keeping my freezer well stocked. And especially to my parents, the opportunities in life you've provided have enabled me to get here and I cannot express my gratitude enough. Dad, thank you for believing in me and for your unwavering support. Mum, thanks to you I'm always reminded to "just try", I know you'd be proud.

Table of Contents

Abstract i

Acknowledgements ii

Table of Contents iii

List of Figures x

List of Tables xiii

Abbreviations xv

1 Introduction 1

1.1 Tuberculosis: The disease 1

1.2 *Mycobacterium tuberculosis*: The pathogen 2

1.3 Evolution and genetic diversity of the *M. tuberculosis* complex 4

1.3.1 The *M. tuberculosis* complex 4

1.3.2 Emergence of the *M. tuberculosis* complex 6

1.3.3 Phylogeography of the *M. tuberculosis* complex 7

1.3.4 Genetic variation within the *M. tuberculosis* complex 8

1.4 Consequences of genetic variation in *M. tuberculosis* 9

1.4.1 Impact on infection and disease outcome 9

1.4.2 Molecular determinants of virulence 10

1.5 Genotyping *M. tuberculosis* 12

1.5.1 Traditional molecular typing methods 12

1.5.2 Large sequence polymorphisms and single nucleotide polymorphisms 15

1.6 Whole genome sequencing of *M. tuberculosis* 17

1.7 Tuberculosis in New Zealand 18

1.7.1 The New Zealand population 18

1.1.1 Tuberculosis incidence in New Zealand 19

1.1.2 Molecular typing in New Zealand 21

1.1.3 The Rangipo cluster 23
1.8 Objectives 25

2 Rangipo SNP-based analyses 26
2.1 Introduction 26
2.1.1 Objectives 28
2.1 Methods 28
2.1.1 Previous Rangipo strain outbreaks 28
2.1.2 *M. tuberculosis* genomic DNA 28
2.1.3 Identification of putative Rangipo specific SNPs 29
2.1.3.1 *In silico* SNP functional assessment 30
2.1.4 Validation of SNPs by Sanger sequencing 30
2.1.4.1 Primers 30
2.1.4.2 Polymerase Chain Reaction 30
2.1.4.3 Agarose Gel Electrophoresis 31
2.1.4.4 ExoISAP treatment and Sanger sequencing of PCR products 31
2.1.5 Sublineage classification 32
2.1.5.1 MIRU based sublineage classification 32
2.1.5.2 S lineage PCR-RFLP assay 32
2.1.6 $DS6^{Quebec}$ deletion PCR assays 33
2.1.7 Contig mapping and analysis 35
2.1.8 Rangipo diagnostic assay 35
2.1.8.1 Rangipo PCR-RFLP assay 36
2.1.8.2 Sputum samples 37
2.1.9 Ethical Considerations 37
2.2 Results 37
2.2.1 Previous Rangipo strain outbreaks 37
2.2.2 Rangipo SNPs 38

2.2.2.1 Gene functional categories 41
2.2.2.2 SNP functional effect prediction 43
2.2.2.3 Rv1631/*coaE* Y363D SNP 45
2.2.2.4 Rv2893 G72S SNP 47
2.2.3 Lineage classification 49
2.2.3.1 S lineage classification PCR-RFLP assay 49
2.2.4 DS6^{Quebec} deletion 50
2.2.5 Rangipo diagnostic 54
2.2.5.1 Rangipo PCR-RFLP assay 55
2.2.5.2 Evaluation of the Rangipo diagnostic in a clinical setting 57
2.3 Discussion 58
2.3.1 Rangipo SNPs 59
2.3.2 The S lineage 62
2.3.3 DS6^{Quebec} deletion 64
2.3.4 Rangipo diagnostic 65
2.3.5 Conclusions 65

3 Structure of the F_{420}-dependent oxidoreductase Rv2893 67
3.1 Introduction 67
3.1.1 Cofactor F_{420} 68
3.1.2 F_{420}-dependent enzymes 69
3.1.2.1 Rv2893 70
3.1.3 Objectives 71
3.2 Methods 72
3.2.1 Molecular cloning of Rv2893 72
3.2.2 Protein expression in *M. smegmatis* 73
3.2.2.1 Small scale protein expression trails 73
3.2.3 Protein purification 74

3.2.3.1 Immobilised metal affinity chromatography purification 74
3.2.3.2 Anion exchange chromatography 74
3.2.3.3 Buffer exchange 75
3.2.3.4 Size exclusion chromatography 75
3.2.3.5 Determination of protein concentration by A280 76
3.2.4 Protein characterisation 76
3.2.4.2 Molecular weight determination by size exclusion chromatography 76
3.2.4.2 Melt temperature determination by differential scanning fluorimetry 76
3.2.5 Protein Crystallography 77
3.2.5.1 Crystallisation robot screens 77
3.2.5.2 Crystallisation fine screens 77
3.2.5.3 Heavy atom derivatisation 77
3.2.5.4 Ligand soaks 78
3.2.6 X-Ray diffraction data collection and structure determination 78
3.2.6.1 Data collection 78
3.2.6.2 Structure solution using single-wavelength anomalous dispersion 79
3.2.6.3 Structure solution by molecular replacement 79
3.2.6.4 Model Refinement 80
3.2.6.5 Structure Analysis 80
3.3 Results 81
3.3.1 Heterologous expression and purification of Rv2893 81
3.3.1.1 Small scale expression trails 81
3.3.2 Large scale expression and protein purification 83
3.3.3 Melt temperature of $Rv2893^{G72S}$ and $Rv2893^{H37Rv}$ 86
3.3.4 Crystallisation of Rv2893 86
3.3.5 Experimental Phasing 88

3.3.6 Structure determination by molecular replacement 91
3.3.6.1 Rv2893 structural models 94
3.4 Analysis of Rv2893 structures 96
3.4.1 β_6–α_6 loop 102
3.4.2 F_{420} binding 103
3.4.3 Substrate binding region and active site 105
3.4.4 The Rangipo specific SNP encoding a G72S mutation 108
3.5 Discussion 110
3.5.1 Conclusions 114

4 Evolutionary analysis of the Rangipo and Otara tuberculosis clusters 115

4.1 Introduction 115
4.1.1 Historical and demographic context 116
4.1.1.1 Early European presence in Polynesia 117
4.1.1.2 European discovery and early settlement of New Zealand 118
4.1.1.3 New Zealand and the Pacific Islands 119
4.1.1.4 A brief history of tuberculosis in New Zealand 120
4.1.2 Phylodynamics and molecular dating 121
4.1.3 Objectives 122
4.2 Methods 122
4.2.1 Whole genome sequencing of New Zealand *M. tuberculosis* isolates 122
4.2.1.1 Clinical *M. tuberculosis* isolates 122
4.2.1.2 Whole genome sequencing 123
4.2.2 Lineage assignment 123
4.2.3 Global *M. tuberculosis* L4.4 dataset 124
4.2.3.1 Additional Rangipo *M. tuberculosis* genomes 124
4.2.3.2 Canadian $DS6^{Quebec}$ lineage *M. tuberculosis* genomes 124
4.2.3.3 Publicly available L4.4 genomes 124

4.2.3.4 Datasets for phylogenetic analyses 125
4.2.4 Reference guided assembly and variant calling 125
4.2.5 Phylogenetic inference 126
4.2.5.1 Nucleotide Alignments 126
4.2.5.2 Selection of nucleotide substitution model 126
4.2.5.3 Maximum likelihood phylogenetic analysis 127
4.2.6 Diversity and clustering analyses 127
4.2.7 Rangipo SNPs 127
4.2.8 Bayesian evolutionary analysis 128
4.2.8.1 Assessment of temporal signal in data 128
4.2.8.2 BEAST analyses 128
4.2.8.3 Model selection 130
4.3 Results 130
4.3.1 New Zealand *M. tuberculosis* cluster isolates 130
4.3.1.1 Illumina whole genome sequencing 130
4.3.1.2 Lineage assignment 131
4.3.1.3 Phylogeny, clustering and genetic diversity of the New Zealand isolates 131
4.3.1.4 Phylogenetic analysis of additional Rangipo strain genomes 136
4.3.2 Rangipo strain SNPs 137
4.3.2.1 Common Rangipo SNPs 137
4.3.2.2 Rangipo specific SNPs 139
4.3.3 Global phylogeny of L4.4 142
4.3.3.1 L4.4 dataset 142
4.3.3.2 Maximum likelihood phylogeny of L4.4 143
4.3.3.3 $DS6^{Quebec}$ deletion 145
4.3.4 Molecular dating 146
4.3.4.1 L4.4.1.1/S dataset 146

4.3.4.2 Assessment of temporal signal for tip-dating 147
4.3.4.3 Molecular dating in BEAST2 149
4.3.4.4 Model selection 152
4.3.4.5 Validity of results 154
4.4 Discussion 155
4.4.1 The New Zealand *M. tuberculosis* clusters 155
4.4.2 Origins and dispersal of the DS6Q clade in Polynesia 157
4.4.2.1 History of the Otara cluster 159
4.4.2.2 History of the Rangipo cluster 160
4.4.3 Tuberculosis stigma 161
4.4.4 The L4.4 sublineage 162
4.4.4.1 L4.4.1.1/S sublineage substitution rate 163
4.4.5 Study Limitations 165
4.4.6 Conclusions 165

5 Conclusions and future directions 166

6 References 169

7 Appendices 187

Appendix A: Appendices relating to Chapter Two 187
Appendix B: Appendices relating to Chapter Three 189
Appendix C: Appendices relating to Chapter Four 195
Appendix D: Publications 202

List of Figures

Figure 1.1. Risk factors for tuberculosis infection and progression to active disease 2

Figure 1.2. Phylogeny of the MTBC 5

Figure 1.3. Schematic showing the principles of traditional molecular typing methods 14

Figure 1.4. Phylogeny of the MTBC based on LSPs and SNPs 16

Figure 1.5. Map of New Zealand showing the main islands and geographical regions 19

Figure 1.6. New Zealand tuberculosis notification rates by ethnicity (2011–2015) 20

Figure 1.7. Proportion of clustered tuberculosis cases in New Zealand (2011–2015) 22

Figure 2.1. *M. tuberculosis* Rangipo strain isolates sequenced by Colangeli *et al.* (2014) 27

Figure 2.2. Domain structure of *M. tuberculosis* CoaE and position of the Y363D RS-nsSNP 45

Figure 2.3. The conserved protein GrpB from *E. faecalis* (2NRK) 46

Figure 2.4. Multiple sequence alignment of F_{420}-dependent of bacterial LLHT family proteins 48

Figure 2.5. Corresponding position of the Rangipo G72S SNP site in *M. tuberculosis* FGD (3B47) 48

Figure 2.6. S lineage PCR-RFLP assay 50

Figure 2.7. Schematic of the $DS6^{Quebec}$ and RD152 deletions 51

Figure 2.8. $DS6^{Quebec}$ deletion PCR assays 52

Figure 2.9. $DS6^{Quebec}$ deletion PCR assay typing of New Zealand *M. tuberculosis* cluster isolates 54

Figure 2.10. Rangipo strain specific PCR-RFLP diagnostic assay 56

Figure 2.11. Rangipo PCR-RFLP diagnostic screening of clinical sputum samples 57

Figure 3.1. Molecular structure of cofactor F_{420} 68

Figure 3.2. Genomic context of Rv2893 in the *M. tuberculosis* H37Rv genome 71

Figure 3.3. Small scale protein expression trails and his-tag binding of Rv2893 82

Figure 3.4. Rv2893 protein purification by immobilised metal affinity chromatography (IMAC) and anion exchange chromatography 84

Figure 3.5. Melting temperature (T_m) comparison of $Rv2893^{H37Rv}$ and $Rv2893^{G72S}$ 86

Figure 3.6. Typical morphology of Rv2893 crystals 87

Figure 3.7. Heavy metal native PAGE gel shift assay to identify heavy atom derivatives for experimental phasing of Rv2893 88

Figure 3.8. Diffraction of Rv2893 heavy metal derivatives 89

Figure 3.9. COOT electron density maps for ligand binding in Rv2893 95

Figure 3.10. Structure of Rv2893 97

Figure 3.11. Dimeric arrangement of Rv2893 98

Figure 3.12. Overall topology of different F_{420}-dependent LLHTs 100

Figure 3.13. Multiple sequence alignment of LLHT proteins 101

Figure 3.14. The β_6-α_6 loop 102

Figure 3.15. F_{420} binding to Rv2893 and other LLHTs. 104

Figure 3.16. Rv2893 substrate binding cleft 106

Figure 3.17. The unidentified ligand bound in the Rv2893 active site 108

Figure 3.18. Rangipo specific G72S SNP 108

Figure 3.19. Structural location of the Rangipo specific G72S SNP 109

Figure 4.1. Map of Oceania highlighting the Polynesian sub-region 116

Figure 4.2. Māori and European population in New Zealand, 1840–1901 119

Figure 4.3. Pacific Island Polynesian population in New Zealand, 1916–1976 119

Figure 4.4. New Zealand tuberculosis notification rates for European and Māori, 1945–1978 120

Figure 4.5. Phylogeny of New Zealand *M. tuberculosis* cluster isolates 132

Figure 4.6. WGS-based clustering of New Zealand *M. tuberculosis* isolates 133

Figure 4.7. Phylogeny of 21 *M. tuberculosis* Rangipo strain genomes 136

Figure 4.8. STRING protein-protein associations among the largest cluster of proteins harbouring common Rangipo SNPs..... 138

Figure 4.9. STRING protein-protein associations among proteins harbouring Rangipo specific nsSNPs 141

Figure 4.10. Global distribution of 236 L4.4 sublineage genomes included in this study (dataset 1, n = 236) 142

Figure 4.11. Global phylogeny of the L4.4 sublineage..... 143

Figure 4.12. Global phylogeny of the L4.4 sublineage and regional distribution of sublineages 144

Figure 4.13. Phylogenetic placement of the DS6^{Quebec} deletion in the L4.4 sublineage 146

Figure 4.14. Global distribution of L4.4.1.1/S sublineage genomes used for molecular dating analyses (dataset 2, n = 117) 147

Figure 4.15. Assessment of temporal signal for tip-dating 148

Figure 4.16. Bayesian phylogeny of the L4.4.1.1/S sublineage 149

Figure 4.17. Bayesian phylogeny and population dynamics of the L4.4.1.1/S sublineage 150

Figure 4.18. Marginal likelihood estimates for replicate path sampling runs in BEAST2..... 153

Figure 4.19. Comparison of parameter estimates using differing population size prior specifications 154

Figure 4.20. Dated Bayesian phylogeny of the DS6Q clade and historical timeline 158

Figure 4.21. Comparison of substitution rate estimated in this study with previously published studies 164

List of Tables

Table 2.1. Phylogenetically informative S lineage marker SNPs..... 32

Table 2.2. S lineage PCR-RFLP assay..... 33

Table 2.3. DS6^{Quebec} deletion assays..... 34

Table 2.4. Rangipo specific PCR-RFLP diagnostic assay..... 36

Table 2.5. Contact tracing of previous Rangipo strain outbreaks..... 38

Table 2.6. Classification of 247 common Rangipo SNPs..... 39

Table 2.7. Non-synonymous Rangipo specific SNPs..... 40

Table 2.8. Functional categories of genes harbouring Rangipo SNPs..... 42

Table 2.9. SNAP2 functional effect predictions for RS-nsSNPs..... 44

Table 3.1. Conditions used to grow Rv2893 crystals used for data collection 87

Table 3.2. Data collection statistics for Rv2893 phasing and statistics for the experimentally phased model..... 90

Table 3.3. Data collection statistics for apo and F_{420} bound Rv2893^{G72S} and Rv2893^{H37Rv}..... 92

Table 3.4. Refinement and model statistics for native and F_{420} bound Rv2893^{G72S} and Rv2893^{H37Rv}..... 93

Table 3.5. RMSD in the Cα atomic positions between Rv2893 structures..... 94

Table 3.6. Proteins in the PDB with structural similarity to Rv2893..... 99

Table 4.1. Summary of Illumina WGS data for New Zealand *M. tuberculosis* cluster isolates..... 131

Table 4.2. Genetic variation in New Zealand *M. tuberculosis* cluster isolates.. 133

Table 4.3. Pairwise SNP distances..... 135

Table 4.4. STRING network analysis of proteins harbouring common Rangipo strain nsSNPs..... 137

Table 4.5. Functional enrichment analysis for clusters of ≥10 proteins harbouring common Rangipo strain nsSNPs..... 138

Table 4.6. Rangipo specific nsSNPs identified from Illumina WGS data..... 140

Table 4.7. Gene functional categories of Rangipo specific nsSNPs..... 141

Table 4.8. Substitution rate and TMRCA estimates for the L4.4.1.1/S sublineage 151

Table 4.9. Model evaluation using path sampling 153

Abbreviations

SI (Systeme Internationale d'Unités) abbreviations are used throughout this thesis for units and elements. Further abbreviations are listed below.

3D	three-dimensional
aa	amino acid
Adf	F_{420}-dependent alcohol dehydrogenase
AJHR	Appendices to the Journals of the House of Representatives
ASU	asymmetric unit
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BLAST	Basic Local Alignment Search Tool
bMer	*Methanosarcina barkeri* Mer
bp	base pair
BSP	Bayesian skyline plot
C-terminal	carboxy-terminus
CoA	Coenzyme A
DHB	district health board
DNA	deoxyribonucleic acid
DR	direct repeat
DSF	differential scanning fluorometry
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid (disodium salt)
ENA	European Nucleotide Archive
ESR	The Institute of Environmental Science and Research
FAD	flavin adenine dinucleotide
FDOR	flavin/deazaflavin oxidoreductase
FGD	F_{420}-dependent glucose-6-phosphate dehydrogenase
FMN	flavin mononucleotide
FPLC	fast protein liquid chromatography
F_{ST}	pairwise fixation index
G6P	glucose-6-phosphate
gDNA	genomic DNA
GTR	general time reversible

HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
His-tag	poly-histidine tag
HMM	Hidden Markov model
HPD	highest posterior density
IMAC	immobilized metal affinity chromatography
indel	small insertions and deletions
IS*6110*	Insertion sequence *6110*
kb	kilobase
kDa	kilodalton
KeV	kiloelectron volt
kMer	*Methanopyrus kandleri* Mer
LAM	Latin-American Mediterranean
LB	Luria-Bertani
LLHT	luciferase-like hydride transferase
LSP	large sequence polymorphism
M	molar
MCMC	Markov chain Monte Carlo
MELAA	Middle Eastern/Latin American/African
Mer	methylenetetrahydromethanopterin reductase
MIRU	mycobacterial interspersed repetitive units
MRCA	most recent common ancestor
MTBC	*Mycobacterium tuberculosis* complex
mtbFGD	*Mycobacterium tuberculosis* FGD
MW	molecular weight
N-terminal	amino-terminus
NAD	nicotinamide adenine dinucleotide
NCBI	National Center for Biotechnology Information
nsSNP	non-synonymous single nucleotide polymorphism
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PDB	protein data bank
PDIM	phthiocerol dimycocerosate
PEG	polyethylene glycol
pI	isoelectric point
RD	region of difference

RE	restriction enzyme
RFLP	restriction fragment length polymorphism
rhFGD	*Rhodococcus jostii* FGD
RMSD	root mean squared deviation
RNA	ribonucleic acid
rpm	revolutions per minute
RS-SNP	Rangipo-specific single nucleotide polymorphism
RS-nsSNP	Rangipo-specific non-synonymous single nucleotide polymorphism
s/s/y	substitutions per site per year
SAD	single-wavelength anomalous diffraction
SDS	sodium dodecyl sulfate
SNP	single nucleotide polymorphism
SOC	super optimal broth with catabolite repression
SRA	short read archive
sSNP	synonymous single nucleotide polymorphism
TE	tris ethylenediaminetetraacetic acid
T_m	melting temperature
tMer	*Methanobacterium thermoautotrophicum* Mer
TMRCA	time to most recent common ancestor
UV	ultra-violet
v/v	volume per volume
VCF	variant call format
VNTR	variable number tandem repeats
w/v	weight per volume
WGS	whole genome sequencing
x *g*	times the force of gravity

Chapter One

Introduction

1.1 Tuberculosis: The disease

Tuberculosis is primarily a pulmonary disease but can also affect almost every organ of the body. Tuberculosis disease in humans is most commonly caused by the bacterial species *Mycobacterium tuberculosis,* an obligate human pathogen. Infection occurs when a person inhales airborne bacteria aerosolised by an infected individual. Inhaled bacilli are engulfed by alveolar macrophages in the lung and a host immune response is triggered involving immune cell recruitment to the site of infection, and an immune structure called a granuloma is formed. The granuloma is a compact aggregation of immune cells and serves as a host protective response to wall off and eradicate the infection (reviewed in Ramakrishnan 2012). However, in some individuals, eradication and containment fails and infection progresses to active disease. Active tuberculosis occurs when the granuloma structure undergoes caseous necrosis and breaks down, causing localised tissue damage and lung cavitation. This allows bacilli to be released into the airways and infected individuals may transmit the disease to others by coughing or through saliva. Symptoms of the disease include a persistent cough, night sweats, fever and fatigue, and left untreated has around a 70% mortality rate (Tiemersma *et al.* 2011).

The result of *M. tuberculosis* infection is highly variable and comprises a spectrum of responses ranging from clearance, to subclinical latent infection, to the development of active disease (Young *et al.* 2009). The course of the disease is also variable and comprises of a range of clinical presentations, including pulmonary disease and extrapulmonary forms such as lymph node tuberculosis and miliary tuberculosis (widespread dissemination via the blood stream). The outcome of exposure and infection is determined by a complex interplay between the host and pathogen. The virulence of the infecting strain, size of the bacterial inoculum, host

clinical characteristics, and social/environmental factors all contribute to the outcome of infection (Lönnroth *et al.* 2009; Salgame *et al.* 2015) (Figure 1.1). For the majority of individuals infected with *M. tuberculosis*, clinical symptoms do not develop and they are not a risk for spreading the disease. Approximately 5–10% of infected individuals however will progress to active disease in their lifetime, with the majority doing so in the first two years of infection (Salgame *et al.* 2015). As active pulmonary tuberculosis is required for transmission to occur, progression to active disease and transmission are related in tuberculosis.

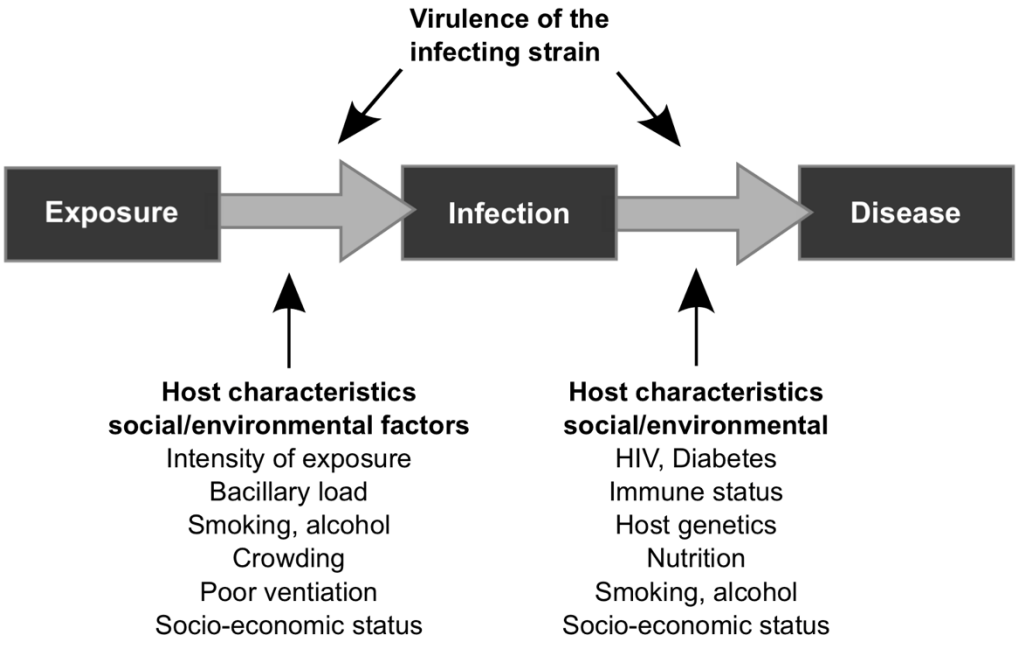


Figure 1.1. *Risk factors for tuberculosis infection and progression to active disease.*

1.2 *Mycobacterium tuberculosis*: The pathogen

M. tuberculosis is a slow growing mycobacterial species with a generation time of around 16–24 hrs. It is a gram-positive acid-fast aerobe and unlike other gram-positive bacteria, it has a thick and highly complex lipid-rich cell wall (reviewed in Brennan 2003; Cook *et al.* 2009). This provides a hydrophobic permeability barrier that plays an important role in intrinsic antimicrobial resistance (Smith *et al.* 2012)

and the array of lipids and glycolipids are also thought to be important for bacterial pathogenicity (Neyrolles and Guilhot 2011).

M. tuberculosis is a highly specialised pathogen and the mechanisms underlying its virulence are complex and multifaceted. *M. tuberculosis* virulence involves the ability of the bacilli to survive extracellularly until a new host is found and then to invade, survive and multiply within host macrophages upon infection. Two central features of *M. tuberculosis* virulence are; (1) its ability to cause active disease leading to cavitation and the expulsion of bacilli allowing transmission, and (2) its capability to persist in a subclinical latent state whilst maintaining the potential to reactivate even decades after infection. In order to achieve this, the pathogen has developed sophisticated strategies and acquired virulence factors to enable it to survive in challenging environments, counteract host antimicrobial responses, persist in a non-replicative state, and facilitate the progression to active disease.

M. tuberculosis lacks classical virulence factors, such as toxins, that are typical of other bacterial pathogens. In *M. tuberculosis*, a virulence factor can generally be considered as a gene whose loss leads to a decrease in pathogenicity but does not impair bacterial growth under standard *in vitro* conditions (Forrellad *et al.* 2013). Sassetti and Rubin (2003) have identified 194 genes that are essential for growth of *M. tuberculosis* in mice and represent essential virulence determinants, and dozens of additional virulence factors have been identified in experimental studies of individual genes and using comparative approaches. The manually curated virulence factor database PATRIC_VF (Wattam *et al.* 2014) currently lists over 450 virulence factors for the laboratory reference strain H37Rv. Identifying and understanding the role of virulence factors in *M. tuberculosis* is essential for the development of new drugs and vaccines to combat this important disease.

1.3 Evolution and genetic diversity of the *M. tuberculosis* complex

1.3.1 The *M. tuberculosis* complex

M. tuberculosis is a member of a group of closely related mycobacteria known as the *M. tuberculosis* complex (MTBC). The MTBC comprises seven human adapted lineages, the animal adapted strains, and the smooth tubercle bacilli *Mycobacterium canettii* (Figure 1.2). The human adapted lineages, lineages 1–7 (L1–L7), belong to the species *M. tuberculosis sensu stricto* (L1–L4 and L7) and *Mycobacterium africanum* (L5 and L6). Among these, L2, L3 and L4 are often referred to as the "modern" lineages and the remaining lineages as the "ancient" lineages (Brosch *et al.* 2002). The human adapted lineages show a strong phylogeographical structure, as further discussed further in Section 1.3.3. L4, the 'Euro-American' lineage, is the most widely globally dispersed of the human-adapted lineages. This lineage comprises ten separate sublineages (Figure 1.2), which have differing phylogeographic distributions (Stucki *et al.* 2016b). Within the animal adapted strains, different ecotypes are adapted to particular mammalian host species and have distinct species designations. For example, *Mycobacterium bovis* and *Mycobacterium pinnipedii* usually infect cattle and seals, respectively. The last member of the MTBC, *M. canettii*, is found only on the Horn of Africa and is assumed to be an opportunistic pathogen with an environmental reservoir (Koeck *et al.* 2011). In this thesis, the term MTBC is used to refer to all members of the MTBC excluding *M. canettii*.

Over a century after *M. tuberculosis* was originally isolated by Robert Koch in 1882, the first whole genome sequence of a *M. tuberculosis* strain (the laboratory reference strain, H37Rv) was determined, revealing a GC rich genome of 4.4 Mb, containing ~4,000 protein coding genes (Cole *et al.* 1998b). Since this time, the complete genome sequence of *M. africanum* (Zhu *et al.* 2015), *M. bovis* (Garnier *et al.* 2003), *M. canettii* (Supply *et al.* 2013), and numerous additional strains of *M. tuberculosis* have been determined.

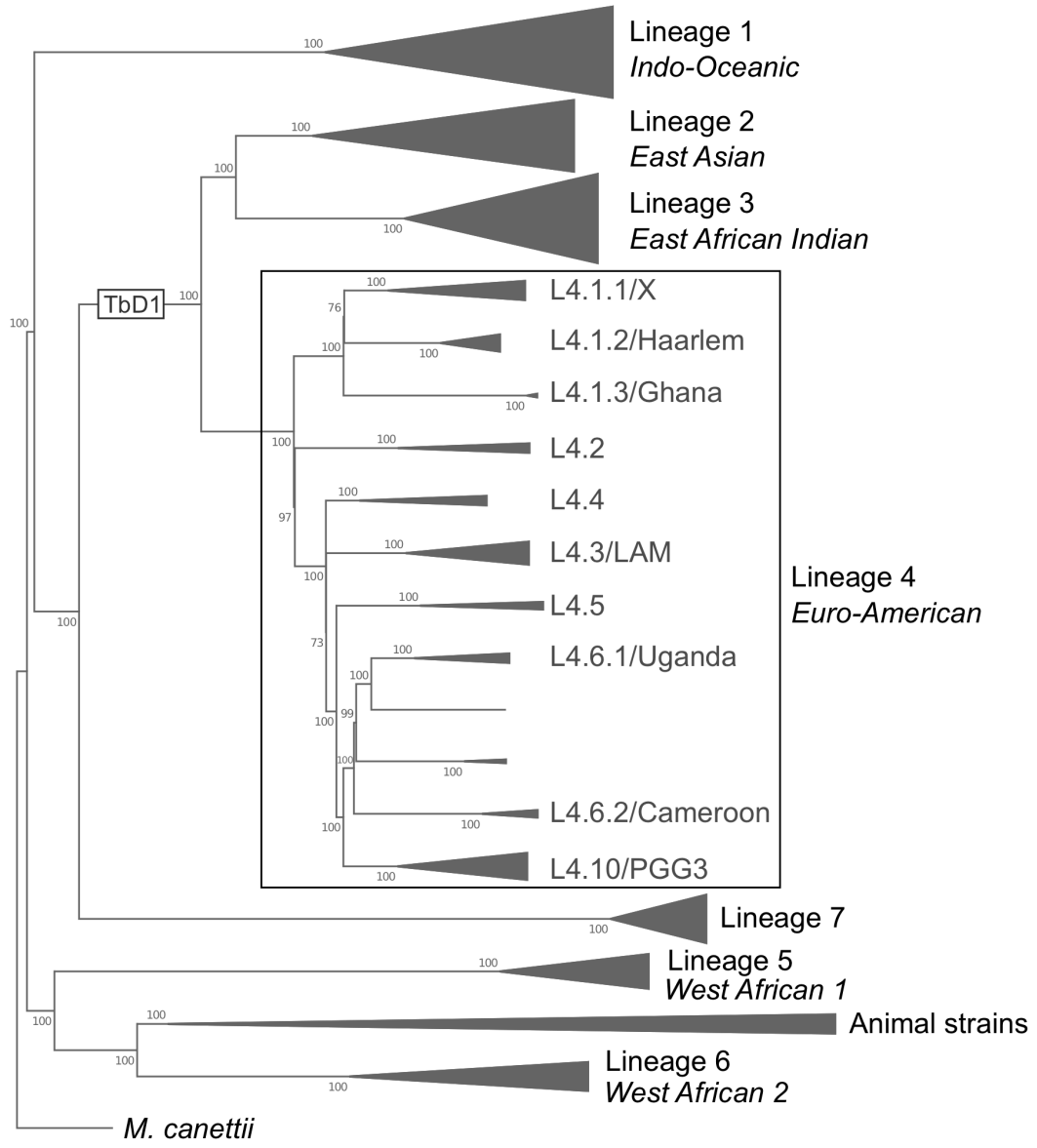


Figure 1.2. *Phylogeny of the MTBC.* The MTBC comprises seven human adapted lineages (L1-L7) and the animal adapted strains. L4 splits into ten major sublineages as shown. Phylogenetic placement of the TbD1 deletion that differentiates between the "modern" (L2, L3, and L4) and ancient lineages (all other lineages) is shown. (Adapted from Stucki *et al.* (2016b)).

1.3.2 Emergence of the *M. tuberculosis* complex

M. tuberculosis was originally thought to have arisen in humans as a result of a zoonosis linked to animal domestication ~10,000 years ago. Molecular data however, has revealed that the animal adapted strains are nested within the diversity of extant human adapted lineages and have a reduced genome compared to *M. tuberculosis*, refuting the hypothesis that *M. tuberculosis* evolved from *M. bovis* (Brosch *et al.* 2002; Hershberg *et al.* 2008). The ancestor of the MTBC is generally now hypothesised to have evolved from a pool of environmental smooth tubercle bacilli similar to *M. canettii* (Supply *et al.* 2013). The acquisition of new genes by horizontal gene transfer as well as genomic decay appear to have played a crucial role in the early evolution and transition to a pathogenic lifestyle in the ancestor of the MTBC (Veyrier *et al.* 2009; Veyrier *et al.* 2011). Despite its important role in the early evolution of the pathogen, members of MTBC do not undergo horizontal gene transfer, and have a strongly clonal population structure (Baker *et al.* 2004; Hirsh *et al.* 2004; Supply *et al.* 2003).

Africa is the only continent where all extant MTBC lineages occur and the earliest branching clades are restricted to Africa, suggesting an African origin of the MTBC and subsequent dispersal throughout the world via human migrations (Comas *et al.* 2013; Hershberg *et al.* 2008; Wirth *et al.* 2008). The date at which the MTBC emerged remains much debated. Archaeological evidence suggests that tuberculosis is an ancient disease. The oldest biomolecular evidence for the MTBC comes from a bison bone from Wyoming dated to 17,000 years before the present (Rothschild *et al.* 2001), and infection with an MTBC strain dating back 9,000 years has been reported in Israel (Hershkovitz *et al.* 2008), however, the authenticity of this finding has been questioned (Wilbur *et al.* 2009). The 'out-of-Africa' hypothesis suggests the most recent common ancestor (MRCA) of the MTBC existed ~70,000 years ago and that the MTBC accompanied migrations of modern humans out of Africa (Comas *et al.* 2013). Conversely, two separate ancient DNA studies both estimated the MRCA existed fewer than 6,000 years ago (Bos *et al.* 2014; Kay *et al.* 2015). If the MRCA of the extant members of the MTBC is indeed younger than previously thought, this does not necessarily represent the time the disease first emerged as wholescale clonal replacement could have displaced ancient strain diversity.

1.3.3 Phylogeography of the *M. tuberculosis* complex

The geographical distribution of the human adapted lineages of the MTBC differ markedly and show a strong phylogeographical structure (Gagneux *et al.* 2006a; Gagneux and Small 2007; Hershberg *et al.* 2008). The most widely globally dispersed lineages are L4 and L2 (which includes the "Beijing" family). Conversely, L7 is almost exclusively found in Ethiopia (Firdessa *et al.* 2013) and the *M. africanum* lineages (L5 and L6) are largely restricted to West Africa (De Jong *et al.* 2010). The phylogeographical distribution of the human adapted lineages is suggestive of co-evolution and adaptation of particular lineages to sympatric human host populations (Baker *et al.* 2004; Brudey *et al.* 2006; Comas *et al.* 2013; Gagneux *et al.* 2006a; Hershberg *et al.* 2008; Hirsh *et al.* 2004; Wirth *et al.* 2008). Adaptation of MTBC lineages to sympatric host human populations is supported by epidemiological studies in human populations whereby association of bacterial lineages with their sympatric host populations persist even in cosmopolitan environments where there is presumed mixing of human populations and bacterial strains (Gagneux *et al.* 2006a; Hirsh *et al.* 2004; Reed *et al.* 2009). Variation in host immunity is thought to play a role in this, a hypothesis supported by the finding that HIV infection, which causes immunodeficiency, disrupts the sympatric host-pathogen relationship (Fenner *et al.* 2013).

Comparison of the MTBC phylogeny with that of the main human mitochondrial haplogroups reveals congruence between the tree topologies, consistent with the MTBC diverging and evolving in parallel with its human host (Comas *et al.* 2013). In contrast, a similar analysis of the global phylogeographic structure of human Y-chromosome data did not identify strong signatures of co-divergence between *M. tuberculosis* and human populations (Pepperell *et al.* 2013). The more recent origin proposed for the MTBC (Bos *et al.* 2014; Kay *et al.* 2015) also precludes the possibly of co-evolution between humans and the MTBC and further, raises the question as to how the pathogen dispersed to achieve its current global distribution in this relatively recent timeframe. In a recent study, O'Neill *et al.* (2017) (preprint) sought to address this question using migration and demographic analyses. The authors connect major human events with global dispersal of the MTBC and present an historical context for a relatively recent evolutionary history of the MTBC based on mutation rates derived from ancient DNA analyses (O'Neill *et al.* 2017).

Regardless of the age of the MTBC, the association of different phylogenetic lineages of the MTBC with different host populations can still be hypothesised to reflect possible adaptation to specific human populations. Furthermore, the MTBC lineages appear to differ in their range of host specificity; some are suggested to behave as generalists able to persist in a range of human host populations, whilst others appear to be more specialist and infect only a small range of host populations (Brites and Gagneux 2015). This phenomenon is also observed amongst the ten major sublineages of the L4 lineage, which vary from being widely globally distributed to geographically restricted (Stucki *et al.* 2016b).

1.3.4 Genetic variation within the *M. tuberculosis* complex

Although the MTBC organisms are low in genomic diversity compared to other bacteria (Frothingham *et al.* 1994), they are genetically more diverse than traditionally assumed. There is increasing evidence that suggests this genomic variation can have phenotypic, clinical and epidemiological consequences (Section 1.4.1, reviewed in Coscolla (2017)). Members of the MTBC differ in their content of single nucleotide polymorphisms (SNPs), small insertions and deletions (indels) and large genomic deletions.

SNPs are one of the most common forms of genetic variation. The majority of these occur in coding regions, which comprise slightly over 90% of the *M. tuberculosis* genome (Cole *et al.* 1998a; Fleischmann *et al.* 2002). On average, *M. tuberculosis* clinical genomes harbour a SNP approximately every 3000 base pairs and two human adapted members of the MTBC differ by an average of around 1200 SNPs (Coscolla and Gagneux 2014). Analysis of the whole genome sequences of 1,601 MTBC isolates from around the world identified over 90,000 SNPs, of which almost 7,000 were strain specific (Coll *et al.* 2014). Strains belonging to L4 were found to contain 15–962 SNPs relative to the H37Rv reference genome, with an average of 746 per genome (Coll *et al.* 2014). A large proportion of SNPs occur as singletons (i.e. are only found in a single strain), a phenomenon suggested to be attributable to background selection (Pepperell *et al.* 2013). Due to their low frequency, such variants are better referred to as single nucleotide mutations rather than SNPs.

1.4 Consequences of genetic variation in *M. tuberculosis*

1.4.1 Impact on infection and disease outcome

It was previously thought that genetic variability between *M. tuberculosis* strains was not of clinical significance. However, increasing evidence now suggests that bacterial genetic background can have important clinical and epidemiological consequences. Numerous animal and *in vitro* studies of *M. tuberculosis* virulence have provided evidence of strain dependent differences in immunogenic and pathogenic properties of *M. tuberculosis* (Aguilar *et al.* 2010; Dormans *et al.* 2004; Lopez *et al.* 2003; Marquina-Castillo *et al.* 2009; Portevin *et al.* 2011; Theus *et al.* 2005). Evidence also suggests strain genotype can influence disease epidemiology and transmissibility (Anderson *et al.* 2013; De Jong *et al.* 2008; Kato-Maeda *et al.* 2010; Yang *et al.* 2012), the efficacy of BCG vaccination (Lopez *et al.* 2003; Tsenova *et al.* 2007), and the emergence of drug resistance (Drobniewski *et al.* 2005; Ford *et al.* 2013).

Strains belonging to the "modern" lineages typically have more virulent phenotypes in animal and *in vitro* models and are generally more transmissible than other strains (reviewed in Coscolla 2017; Coscolla and Gagneux 2014). For example, modern lineage strains are reported to replicate faster *in vitro* and in aerosol infected mice (Reiling *et al.* 2013), induce lower early inflammatory responses (Portevin *et al.* 2011), and are more likely to progress to active disease when compared to ancient lineage strains (De Jong *et al.* 2008). Whole genome sequencing (WGS) provides a powerful tool to study *M. tuberculosis* transmission (Section 1.6). Recent WGS studies have shown elevated transmission and a higher proportion of disease due to recent transmission for L2 compared with L1 (Guerra-Assunção *et al.* 2015; Holt *et al.* 2018). The L2/Beijing family is also commonly associated with drug resistance, and it is suggested that the genetic background of these strains is better suited to tolerate the fitness costs of drug resistance conferring mutations (Borrell and Gagneux 2009).

Strain dependent differences in virulence phenotypes and transmissibility have also been reported at the sublineage level (Aguilar *et al.* 2010; Anderson *et al.* 2013; Kato-Maeda *et al.* 2010; Kato-Maeda *et al.* 2012). For example, differences in the transmissibility of Beijing strains have been associated with virulence phenotypes

in a mouse model; mice infected with highly transmitted strains were shown to have higher mortality, more extensive tissue damage, and differential cytokine expression (Aguilar *et al.* 2010). The association of virulence phenotypes with transmission also supports an elementary role for increased virulence in enhanced transmission.

1.4.2 Molecular determinants of virulence

While it is becoming evident there are important differences in virulence phenotypes between different lineages of *M. tuberculosis,* the molecular mechanisms underlying strain dependent variation in virulence remain unclear. Thousands of single nucleotide changes have been identified in clinical *M. tuberculosis* isolates and have potential phenotypic consequences. Around two-thirds of SNPs in *M. tuberculosis* are non-synonymous (nsSNPs), meaning they change the amino acid sequence of encoded proteins. Non-synonymous changes can result in amino acid substitutions, frame shifts, or premature stop codon insertion, potentially altering protein structure and function, which in turn has the potential to affect clinical and epidemiological characteristics of the disease. Rose *et al.* (2013) have estimated that around 44% of SNPs may impact protein function, although it remains to be determined whether this variation has biological and/or clinical consequences. Determining the functional consequences of SNPs at the molecular, biochemical and biological level, is important to ultimately elucidate the role this variation plays in the outcome of infection and disease.

The most well characterised effects of single nucleotide changes on bacterial fitness (and amongst the most consequential) are those involved in the development of drug resistance. As the MTBC does not undergo horizontal gene transfer, drug resistance in *M. tuberculosis* is primarily caused by single nucleotide mutations. Other SNPs that mitigate the fitness costs associated with drug resistance, termed 'compensatory mutations', have also been well documented (Casali *et al.* 2014; Comas *et al.* 2012; De Vos *et al.* 2013). Furthermore, strain genetic background also influences the fitness costs associated with the acquisition of drug resistance mutations (Gagneux *et al.* 2006b).

However, the majority of single nucleotide variants in the MTBC are not associated with drug resistance. Hence, this presents the obvious question: what effects do the remainder of these have on bacterial phenotype? Elucidating the effect of SNPs on protein function is an important step towards understanding the practical consequences of genomic variation in the MTBC and also can shed new light on *M. tuberculosis* biology. Systematic biochemical *in vitro* and *in vivo* analysis of the catalogue of *M. tuberculosis* SNP diversity is simply not practical. Comparative approaches and *in silico* prediction of functional effects provide a means to identify SNPs that may affect protein function, presenting candidates for further study. Recent genomic analysis of outbreak strains in Denmark and New Zealand have identified SNPs in virulence factors that may contribute to the success of these strains (Folkvardsen *et al.* 2018; Gautam *et al.* 2017), however this is yet to be experimentally validated. Holt *et al.* (2018) have recently identified a threonine to alanine SNP at position two (T2A) in *EsxW* that is under positive selection and is suggested to be a possible contributor to enhanced transmission of Beijing family strains (Holt *et al.* 2018). This SNP was shown to lead to a slightly increased affinity and stability of the EsxW/EsxV heterodimer (Holt *et al.* 2018), which is thought to be involved in ESX-5 substrate selection – a crucial virulence determinant for *M. tuberculosis.* SNPs that alter transcriptional regulation also have the potential to alter bacterial phenotype. SNPs affecting promoter activity and transcription start sites and nsSNPs in transcriptional regulators, have been associated with lineage specific differences in transcription profiles (Rose *et al.* 2013).

Understanding the complex relationship between bacterial genotype, virulence and clinical outcomes, is important for the development of improved vaccines, treatment regimens and public health measures. Characterising the effects of strain specific variation at the molecular level is an essential step towards understanding how genomic variation in the MTBC affects bacterial phenotype and its clinical and epidemiological outcomes. Studying genotype-phenotype associations also require robust methods to genotype isolates and is discussed in the following section.

1.5 Genotyping *M. tuberculosis*

Molecular typing of *M. tuberculosis* is important both in public health and research settings. Genotypic information can be used to discriminate between reactivation and re-infection, link index cases and contacts, and to track the spread of strains throughout the population. Strain lineage can also influence important clinical and epidemiological aspects of the disease (Section 1.4.1). The ability to distinguish genotype for *M. tuberculosis* with high confidence is therefore important for tuberculosis control efforts and to study genotype-phenotype associations. Genotyping data can also be used to investigate phylogenetic relationships within the MTBC to understand the population structure and evolution of the pathogen and provide insights as to how it has achieved its current global distribution.

1.5.1 Traditional molecular typing methods

Traditional molecular typing methods for *M. tuberculosis* involve the analysis of repetitive genetic elements using restriction fragment length polymorphism (RFLP) and/or polymerase chain reaction (PCR) based approaches. The three main typing methods used in epidemiological investigations are IS*6110*-RFLP, mycobacterial interspersed repetitive unit variable number tandem repeat analysis (MIRU-VNTR) and spoligotyping (Figure 1.3).

IS*6110*-RFLP is based on differences in the copy number and genomic locations of the IS*6110* insertion element, which is exclusively found in the MTBC (van Embden *et al.* 1993) (Figure 1.3B). This method was the first "gold standard" for molecular typing of the MTBC and has been used in hundreds of studies around the globe. This approach however has limited resolution for strains with low numbers of IS*6110* repeats (Cowan *et al.* 2002), is laborious, and relies on having large amounts of high-quality DNA requiring long culturing times. Spoligotyping and MIRU-VNTR typing are PCR based genotyping techniques developed to overcome some of the limitations of IS*6110*-RFLP. Spoligotyping patterns are defined by the presence or absence of 43 unique spacer sequences located in the direct repeat (DR) region of the *M. tuberculosis* genome (Kamerbeek *et al.* 1997) (Figure 1.3C). This approach provides typing information at the sublineage level, however, its ability to determine epidemiological links between tuberculosis cases is limited.

MIRU-VNTR typing classifies strains based on the number of repeats at different VNTR loci in the *M. tuberculosis* genome (Figure 1.3D). MIRU-VNTR typing using 15 or 24 loci has a similar overall discriminatory power to IS*6110*-RFLP when used alone or in combination with spoligotyping (reviewed in Merker *et al.* 2017). Both spoligotyping and MIRU-VNTR typing patterns can be readily compared to hundreds of other patterns from around the globe using the SITVITWEB (Demay *et al.* 2012) and MIRU-VNTR*plus* (Weniger *et al.* 2010) databases. They also require only small amounts of DNA, and therefore have turnaround times significantly faster than IS*6110*-RFLP. As a result of these benefits, MIRU-VNTR alone, or combined with spoligotyping, replaced IS*6110*-RFLP as the gold standard for tuberculosis genotyping until quite recently. These methods are in the process of being superseded by WGS as advances in next generation sequencing technologies and reduced costs are making this more practical (Section 1.6).

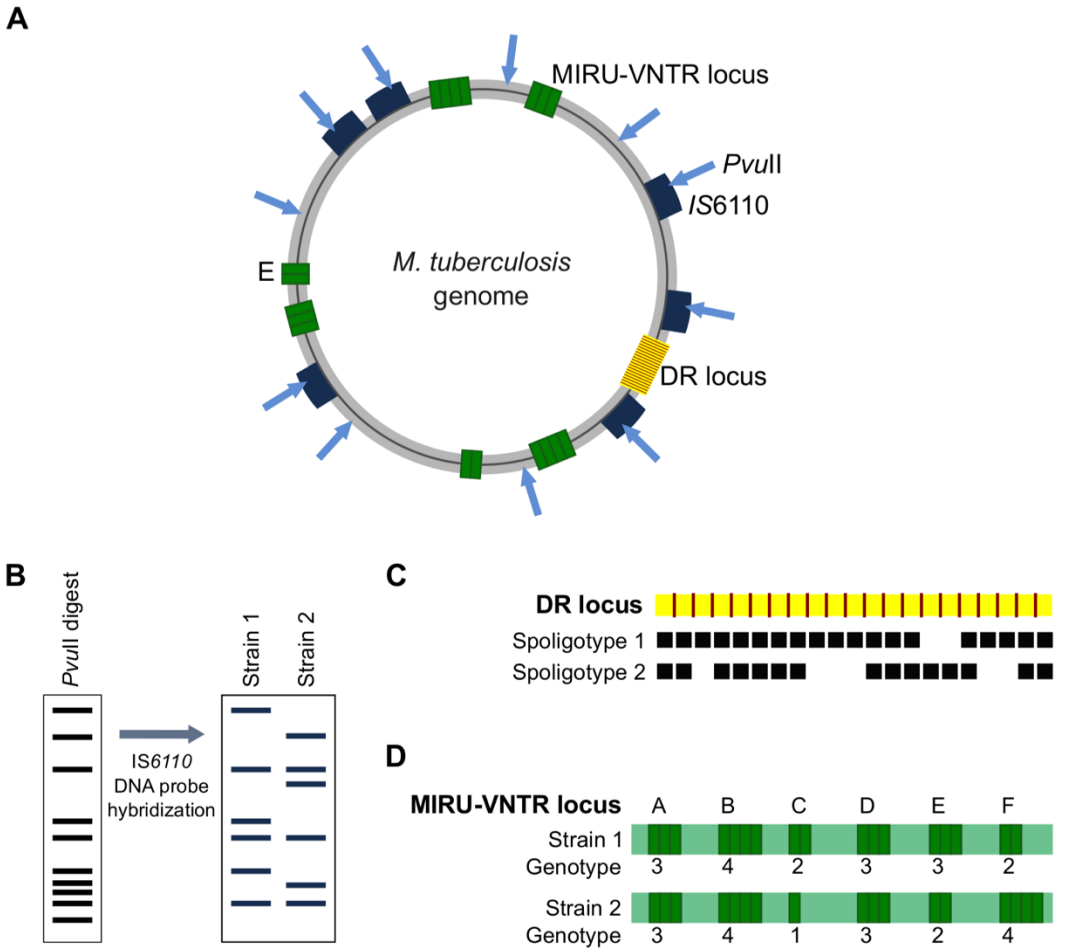


Figure 1.3. *Schematic showing the principles of traditional molecular typing methods.* (A) *M. tuberculosis* genome showing IS*6110* elements (dark blue) and *Pvu*II cleavage sites (light blue arrows); MIRU-VNTR loci with different numbers of repeats scattered around the genome (green boxes); and the direct repeat (DR) locus (yellow). (B) IS*6110*-RFLP typing. *M. tuberculosis* genomes differ in the locations and number of IS*6110* elements they possess. The whole genome is digested with *Pvu*II and then separated by gel electrophoresis. Hybridisation of fragments with an IS*6110* specific DNA probe produces the characteristic banding pattern for each isolate. (C) Spoligotyping. The DR region is amplified by PCR and products are hybridised to spacer-specific oligonucleotide probes. The presence (black square) or absence (white square/blank) of 43 unique spacers located between direct repeats at the DR locus determines the spoligotyping pattern. (D) MIRU-VNTR typing. The number of repeats present at each MIRU-VNTR locus determines the MIRU-VNTR genotype. Specific primers amplify each locus and amplicon sizes are determined by gel electrophoresis. Standardised sets of 12, 15 or 24 different loci are used for genotyping.

1.5.2 Large sequence polymorphisms and single nucleotide polymorphisms

The molecular markers used for MIRU-VNTR and spoligotyping are prone to convergent evolution and have limited discriminatory power, limiting their use in phylogenetic studies. Although not typically used clinically, genotyping methods based on the analysis of genomic deletions (commonly also referred to as large sequence polymorphisms (LSPs), or regions of difference (RDs)), have been used as more robust markers for phylogenetic analysis of the MTBC. LSPs can be used to assign strains to the major lineages and sublineages of the MTBC (Figure 1.4A) and have provided valuable insights into the global phylogeographic distribution and evolution of the pathogen (Comas and Gagneux 2009; Gagneux *et al.* 2006a; Hershberg *et al.* 2008).

SNPs provide useful epidemiological and phylogenetic information and can also have important functional consequences (Section 1.4.2). SNPs show very low levels of homoplasy (~1%) (i.e. do not frequently undergo convergent evolution) and thus represent robust markers for phylogenetic analysis (Comas *et al.* 2009; Homolka *et al.* 2012). Phylogenetic trees constructed using SNP data are congruent with the main lineages identified by LSPs and better resolve phylogenetic relationships, identifying previously un-discriminated lineages (Coll *et al.* 2014; Homolka *et al.* 2012) (Figure 1.4B). Accordingly, several SNP classification schemes describing sets of phylogenetically informative marker SNPs for the assignment of strains to the main lineages and sublineages have been reported (Coll *et al.* 2014; Filliol *et al.* 2006; Homolka *et al.* 2012). The most comprehensive of these to date describes a set of 62 marker SNPs, allowing assignment to the all the main lineages and sublineages with high confidence and classifies a greater number of lineages than previously published alternatives (Coll *et al.* 2014). SNPs also offer practical advantages as genetic markers due to their suitability for the development of rapid and relatively inexpensive high throughput assays to genotype clinical isolates. For example, Stucki *et al.* (2012) have developed two rapid SNP-typing methods to classify isolates into the main MTBC lineages with high confidence and comparably little cost. Analysis of genome-wide SNPs using WGS provides the ultimate resolution for strain typing, epidemiological and phylogenetic analysis, further discussed in the following section.

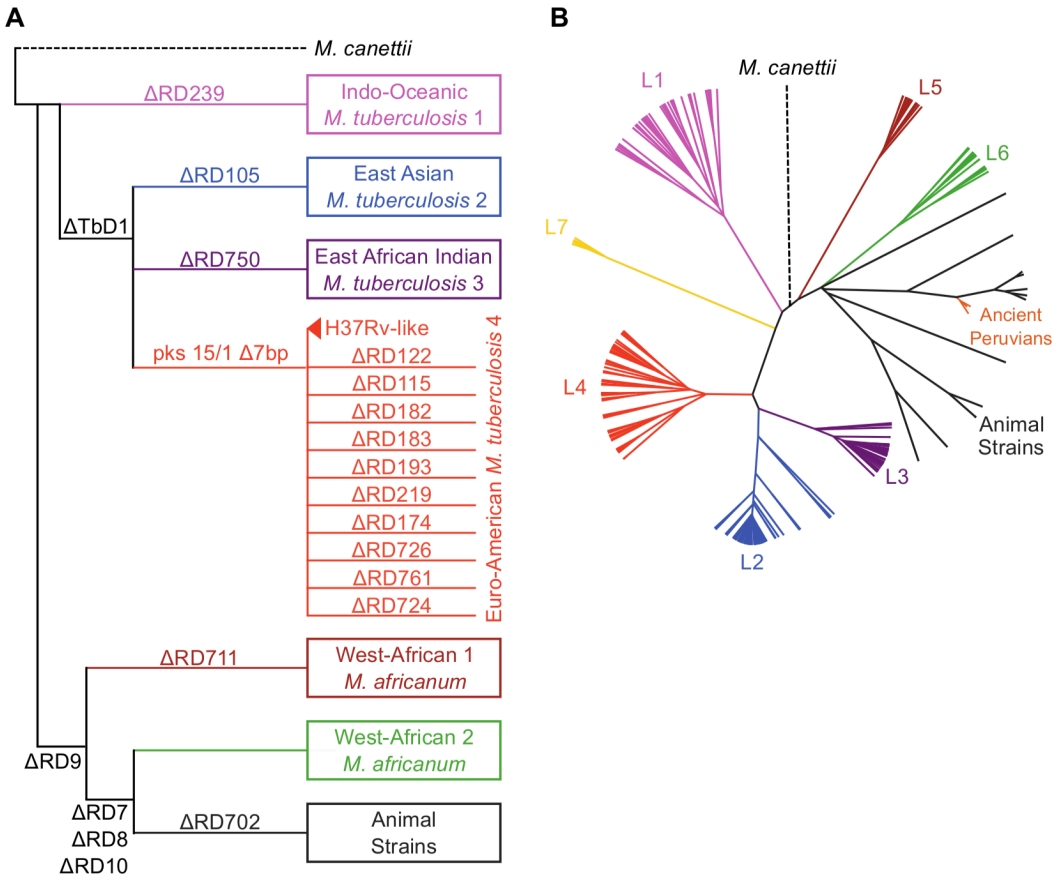


Figure 1.4. *Phylogeny of the MTBC based on LSPs and SNPs.* (A) Schematic phylogeny of the main MTBC lineages based on LSPs. Specific lineage defining regions of difference are indicated on branches, lineage colours correspond the colouring in (B). RDs as in Gagneux *et al.* (2006a) and Brosch *et al.* (2002), additional LSPs enabling further classification of the animal adapted strains and *M. tuberculosis* lineage 2 are not shown. (B) Bayesian phylogeny inferred from genome-wide SNPs of 261 genomes, including three *M. pinnipedii* strains from 1000-year-old Peruvian mummies (orange) (Adapted from Bos *et al.* (2014)).

1.6 Whole genome sequencing of *M. tuberculosis*

WGS is now considered the new gold standard for strain genotyping and phylogenetic analysis, providing the ultimate resolution to discriminate between closely related clones. Due to advances in next generation technology and reduced costs, WGS is likely to become routine in typing laboratories (at least in low incidence countries) and hundreds of genomes from different *M. tuberculosis* strains from around the globe have now been sequenced. This is providing valuable insights into the transmission, emergence of drug resistance, evolution and global dispersal of the pathogen.

WGS provides a powerful approach to study tuberculosis transmission and outbreaks in high resolution. Genomic network analysis provides detailed insights into *M. tuberculosis* transmission dynamics, enabling inference of transmission pathways, detection of sub-clusters and micro-epidemics, identification of "super spreaders", and has the potential to predict undiagnosed cases, potentially leading to earlier identification and treatment of infectious cases and their contacts (Gardy *et al.* 2011; Guerra-Assunção *et al.* 2015; Roetzer *et al.* 2013; Stucki *et al.* 2015; Walker *et al.* 2013b). Important to the determination of SNP thresholds for distinguishing linked and unlinked *M. tuberculosis* cases is the short-term mutation rate, which has been estimated at 0.3–0.5 SNPs per year (Bryant *et al.* 2013b; Roetzer *et al.* 2013; Walker *et al.* 2013a). Most isolates from epidemiologically linked patients differ by five of fewer SNPs (Perez-Lago *et al.* 2014; Walker *et al.* 2013a). Walker *et al.* (2013a) have proposed a threshold of five or fewer SNPs between isolates as suggestive of epidemiological linkage and more than 12 SNPs to rule out recent transmission.

Molecular dating provides another layer of information, enabling evolutionary timelines to be reconstructed characterising the dynamics of outbreaks and dating key events, such as the emergence of drug resistance and expansion of successful clones (Bjorn-Mortensen *et al.* 2016; Eldholm *et al.* 2015; Eldholm *et al.* 2016). Such high-resolution studies combining molecular-dating with spatial and/or epidemiological data have only recently been made possible due to increased accessibility and affordability of WGS, and offer a powerful way to study pathogen evolution in space and time.

In addition to studying inter-patient diversity, WGS can provide valuable insights into intra-patient variation and micro-evolutionary processes (Perez-Lago *et al.* 2014; Trauner *et al.* 2017) and differentiate relapse and re-infection cases with greater resolution than traditional typing methods (Bryant *et al.* 2013a; Guerra-Assunção *et al.* 2014). WGS also offers great promise as a tool to rapidly identify drug resistance mutations, potentially reducing the time taken to diagnose drug resistance from weeks to days (Koser *et al.* 2013; Pankhurst *et al.* 2016).

1.7 Tuberculosis in New Zealand

The research presented in this thesis is based on the genomic analysis of prevalent *M. tuberculosis* strains found in New Zealand. New Zealand is considered a low incidence country, having an annual tuberculosis notification rate of around 6.5 cases per 100,000 population, which translates to approximately 300 cases per year (ESR 2018). Tuberculosis is a notifiable disease in New Zealand under the Tuberculosis Act 1948 and is one of several communicable diseases (including meningococcal disease and acute rheumatic fever) that are a major source of socioeconomic and ethnic inequalities in New Zealand (Baker *et al.* 2012).

1.7.1 The New Zealand population

Aotearoa New Zealand has a population of approximately 4.9 million people and is the largest island nation in the Polynesian region of Oceania (national population estimate as at June 2018, StatsNZ (2018)). The majority of the population live on the two main islands, the North Island (*Te Ika a Maui*) and the South Island (*Te Wai Pounamu*) (Figure 1.5). Māori are the indigenous people of New Zealand and New Zealand is also home to the largest diaspora communities of indigenous Pacific People globally (Spickard *et al.* 2002). The European ethnic group is the dominant ethnic group and 74% of people identify with at least one European ethnicity (2013 National Census, StatsNZ (2013)). The other major ethnic groups are Māori (15%), Asian (12%), Pacific People (7%) and Middle Eastern/Latin American/African (MELAA) (1%) (it is important to note that some people identify with more than

one ethnic group). Nearly two-thirds of people who identify with at least one Pacific ethnicity were born in New Zealand.

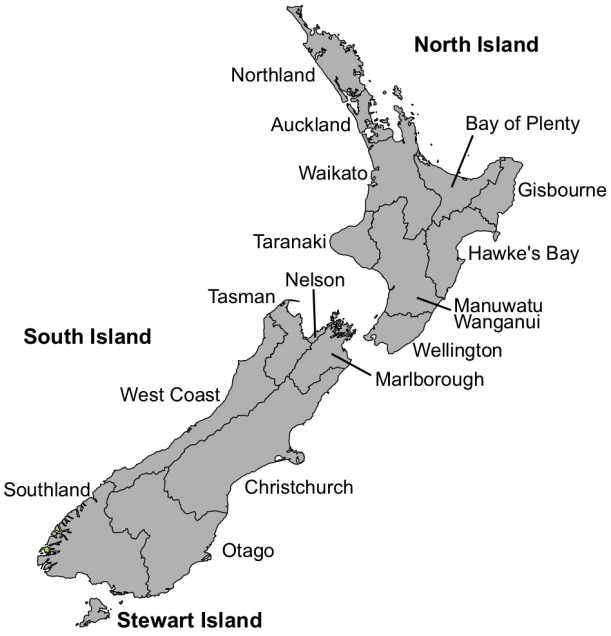


Figure 1.5. *Map of New Zealand showing the main islands and geographical regions.*

1.1.1 Tuberculosis incidence in New Zealand

Unless otherwise indicated, all of the following statistics are from the 2015 annual surveillance report of tuberculosis in New Zealand released by The Institute of Environmental Science and Research (ESR) (ESR 2018). Statistics pre-dating this are either provided in the 2015 report, or extracted from older reports (Bissielo *et al.* 2012; ESR 2015a; 2015b; Lim and Heffernan 2013).

From 1990 to 2003, New Zealand tuberculosis notification rates fluctuated between 8.5–11.6 per 100,000. Between 2003 and 2007, rates decreased to 6.7/100,000 and have remained relatively stable since. In 2015 there were 294 notified cases of tuberculosis, resulting in a rate of 6.4 per 100,000 population (6.4/100,000). The majority of notifications were for new cases (97.3%, 286/294) and most were laboratory confirmed (88.8%, 261/294). The New Zealand 2015 notification rate is lower than the corresponding rate reported for the United Kingdom (10.5/100,000) (England 2016), but higher than for the United States (3.0/100,000) (CDC 2017),

Canada (4.6/100,000) (Gallant *et al.* 2017) and Australia (5.2/100,000) (NNDSS 2015).

There are several demographic differences in tuberculosis notifications in New Zealand, including age, ethnicity and place of birth. The majority of new cases occur in patients over 14 years of age (96.9%, 277/286), and the highest rate is observed in the 15–39 age group (9.5/100,000). Higher rates of tuberculosis occur in socioeconomically deprived areas and this trend is most apparent among New Zealand-born cases. Regional differences in notification rates are also apparent. The Auckland and Counties Manukau district health boards (DHBs) have the highest rates at 12.7/100,00 and 12.3/100,000, respectively, while in other DHBs rates range from zero (i.e. no cases) to 7.0/100,000. Rates vary markedly with ethnicity and over the last five years the Asian ethnic group has consistently had the highest notification rate (Figure 1.6A). In 2015, the highest notification rates were observed for the Asian ethnic group (34.3/100,000), followed by Pacific Peoples (20.2/100,000), MELAA (15.7/100,00), Māori (3.2/100,000) and European or Other ethnic groups (0.6/100,000).

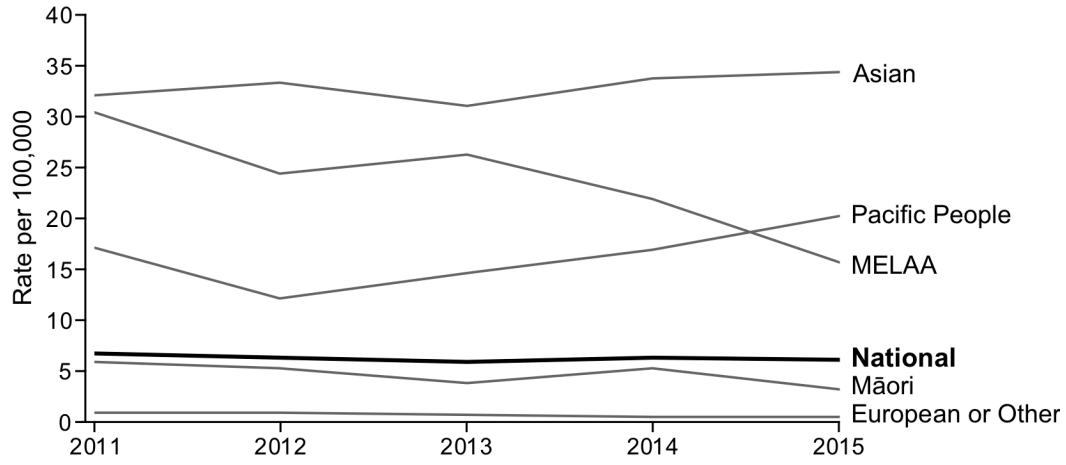


Figure 1.6. *New Zealand tuberculosis notification rates by ethnicity (2011–2015)*. Annual notification rates of new cases for different ethnic groups and the overall national rate are shown. It is important to note that although rates are relatively high for the MELAA (Middle Eastern/Latin American/African) group, the absolute number of cases in this group are low (8–16 cases annually). (Source: ESR (2018)).

For the five years from 2011–2015, slightly over three-quarters of all notified cases (75.5–82.9%) were born outside of New Zealand, and around half of these occurred within the first six years of arrival in New Zealand. The proportion of tuberculosis patients born in the Pacific Islands has recently increased from 13.2% in 2014, to 21.5% in 2015. Among New Zealand-born cases in 2015, 44.9% occurred in Māori, 24.5% in Pacific People, 22.4% in the European or Other group, and 8.2% in the Asian ethnic group.

New Zealand has a relatively low rate of drug resistant tuberculosis and only one case of extensively drug-resistant tuberculosis has been identified (this was in an immigrant from Myanmar in 2010 (Goh *et al.* 2011)). Drug susceptibility testing to isoniazid, rifampicin, ethambutol, streptomycin and pyrazinamide is routinely performed. In 2015, 89.7% (218/243) of tested isolates were fully drug susceptible and two cases of multi-drug-resistant tuberculosis were identified (0.8% of cases). Drug resistance is higher among non-New Zealand-born than New Zealand-born cases (11.6% compared to 2.8% of isolates, respectively) and only two multi-drug resistant tuberculosis cases have been identified in New Zealand-born patients in the ten years from 2004 to 2015.

1.1.2 Molecular typing in New Zealand

Universal typing of tuberculosis isolates was introduced in New Zealand 2002. Prior to 2009, IS*6110*-RFLP was the primary typing method, after which MIRU-VNTR became the primary method with secondary typing by IS*6110*-RFLP. Since 2011 typing has been undertaken by MIRU-VNTR analysis alone. At present, all molecular typing is carried out by the national tuberculosis reference laboratory, LabPLUS (Auckland, N.Z.). Primary typing is performed by MIRU-VNTR 12-loci analysis, and secondary typing by MIRU-VNTR 24-loci when an isolate shares its 12-loci pattern with previously typed isolate (Lim and Heffernan 2013).

Over one-third of MIRU-VNTR typed *M. tuberculosis* isolates in New Zealand have non-unique molecular typing results and are able to be assigned to ‘clusters’ (groups of strains with identical typing patterns) (ESR 2018). Genotypic clustering of *M. tuberculosis* isolates is typically interpreted to indicate recent transmission, although clustering can also represent reactivation of latent infection with endemic strains (Nguyen *et al.* 2004). A higher proportion of cases with non-unique

molecular types occur in the Māori (74.1%), and Pacific People (80.1%) ethnic groups, compared to all other groups (European or Other 32.5%, Asian 24.3%, MELAA 12.8%) (Figure 1.7). Cases born in the Pacific Islands have the highest proportion of non-unique molecular types for any region of birth (72.1%), although overall, overseas born cases have a lower proportion of non-unique molecular types (40.3%) compared to New Zealand-born patients (61.4%).

Although the majority of tuberculosis clusters in New Zealand are small (<5 cases), there are a small number of large clusters, some of which have been a source of tuberculosis infection for over two decades. The largest cluster in New Zealand based on identical or highly similar MIRU-VNTR typing patterns is most prevalent in Māori and is known as the 'Rangipo' cluster (Section 1.1.3). Two other major clusters, called the 'Southern Cross' and 'Otara' clusters, are most common in Pacific People. Since 2002, the Rangipo cluster has been responsible for at least 165 tuberculosis cases, the Southern Cross cluster 152 cases, and the Otara cluster 32 cases (V Playle, LabPLUS, personal communication).

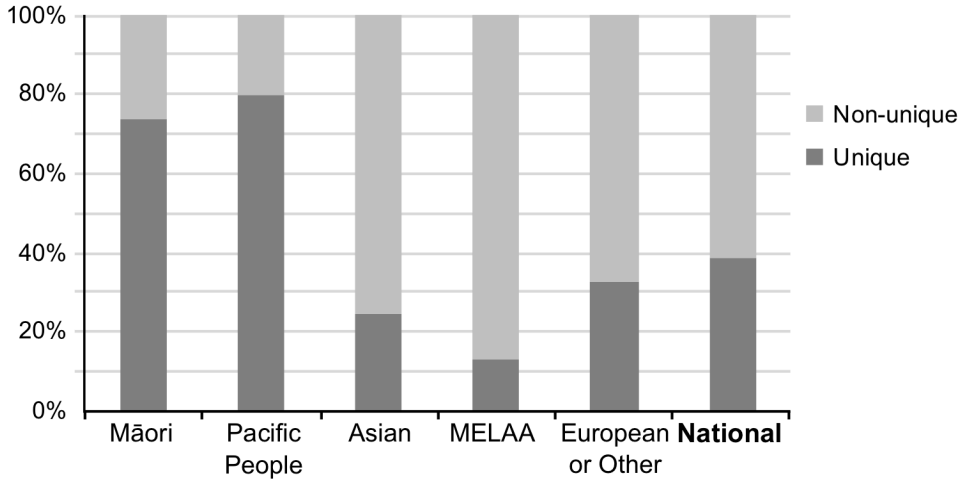


Figure 1.7. *Proportion of clustered tuberculosis cases in New Zealand (2011–2015).* Percentage of non-unique (i.e. clustered) and unique MIRU-VNTR typing profiles for different ethnic groups and the overall national proportion are shown. (Source: ESR (2018)).

1.1.3 The Rangipo cluster

The Rangipo cluster is the largest New Zealand *M. tuberculosis* cluster identified by related MIRU-VNTR typing patterns. This cluster has been a cause of ongoing outbreaks for around 30 years (Calder 2013; Colangeli *et al.* 2014; De Zoysa *et al.* 2001; McElnay *et al.* 2004). Approximately 12% of clustered cases in New Zealand belong to the Rangipo cluster and it is associated with more DHBs than any other cluster (Sexton *et al.* 2008). The majority of Rangipo strain cases occur in Māori (88%, 81/92 201), and it accounts for around one-quarter of tuberculosis cases in this population (J Sherwood, ESR, personal communication).

The Rangipo strain is most notably associated with a large outbreak in the late 1990s linked to the Rangipo prison (De Zoysa *et al.* 2001). During a three and half year period, this outbreak resulted in at least 61 cases of active tuberculosis and 119 cases of latent tuberculosis throughout the North Island (De Zoysa *et al.* 2001). It has since become known as the Rangipo strain, although cases caused by this strain have been recognised from the early 1990s, pre-dating the prison outbreak (N Karalus, Waikato Hospital, personal communication). In a subsequent 2002 Hawke's Bay outbreak, 19 new cases of active tuberculosis and 42 cases of latent disease were diagnosed, including contacts only casually connected to the index case (McElnay *et al.* 2004). Such large outbreaks present a serious challenge to tuberculosis control efforts in New Zealand and stretch local public health and clinical services. Recently, an outbreak of the Rangipo strain involving six cases occurred in a small Waikato town in 2017/2018, showing the strain continues to spread and cause disease in New Zealand (R Hoskins, Waikato Hospital, personal communication).

Although no virulence testing has yet been undertaken, anecdotal evidence suggests this strain may be highly transmissible and previous outbreaks demonstrate high infectivity of index cases, infection of casual contacts, and the generation of large numbers of secondary cases (Calder 2013; De Zoysa *et al.* 2001; McElnay *et al.* 2004). Host social/environmental factors are also likely important contributors to the success of the Rangipo strain. For example, prison incarceration and/or gang affiliations were a feature of several cases in previous Rangipo outbreaks (De Zoysa *et al.* 2001; McElnay *et al.* 2004).

A clear understanding of the bacterial and host factors that underlie the success of this prevalent strain are imperative to improve public health measures and develop effective strategies to halt its further spread. Prompt identification of Rangipo strain cases is also important for this. To address this, WGS-data has been used in our lab to guide the development of a rapid SNP-based Rangipo strain-specific diagnostic assay (Ruthe 2015). WGS offers a promising route to shed light on transmission and bacterial genetic factors that may contribute to the success of the Rangipo strain and other prominent clusters in New Zealand. WGS of Rangipo isolates has revealed very little SNP diversity among isolates (0–19 SNP differences between isolates), consistent with relatively recent transmission (Colangeli *et al.* 2014; Gautam *et al.* 2017). A number of nsSNPs in virulence associated genes in Rangipo strains have been identified (Colangeli *et al.* 2014; Gautam *et al.* 2017; Ruthe 2015), providing valuable insight into the genomic context of the Rangipo cluster, although questions still remain concerning the functional consequences of common Rangipo genetic variants. Important evolutionary and social questions surrounding the origins and expansion of the Rangipo strain are yet to be addressed and are necessary to understand how this strain has become established as an important cause of tuberculosis in New Zealand.

1.8 Objectives

In my doctoral research I have used WGS data, phylogenetics and structural biology to examine the evolutionary origins and functional consequences of genetic variation in the New Zealand *M. tuberculosis* Rangipo strain and other prominent local strains. The overall aim of this research was to enhance our understanding of factors contributing to the success of the Rangipo strain in New Zealand. The specific objectives addressed in this thesis are outlined below.

Objective one: Identify SNPs specific to the Rangipo strain.

Objective two: Examine the consequences of genomic variation specific to the Rangipo strain at the molecular level using protein crystallography and biochemical approaches.

Objective three: Characterise the phylogenetic relationships and genetic variation within and between the Rangipo, Southern Cross and Otara clusters.

Objective four: Examine the evolutionary history of the Rangipo strain and identify its historical origin.

Chapter Two

Rangipo SNP-based analyses

2.1 Introduction

Colangeli *et al.* (2014) have sequenced the whole genomes of ten New Zealand Rangipo strain *Mycobacterium tuberculosis* clinical isolates spanning a 19-year period from 1991 to 2011, using the Applied Biosystems Inc. (ABI) SOLiD platform. These data were used to estimate *M. tuberculosis* mutation rates during latency and active disease (Colangeli *et al.* 2014) and for additional single nucleotide polymorphisms (SNP) analyses undertaken in this laboratory (Ruthe 2015). Isolates included five cases from a single source outbreak and five cases with no known epidemiological links to the outbreak (Figure 2.1). The ten isolates were found to carry little genetic variation, differing from each other by only 1–14 SNPs, and neighbour-joining trees also showed close phylogenetic relationships between the isolates (Colangeli *et al.* 2014). A detailed phylogeny in the context of other New Zealand or global *M. tuberculosis* isolates and full lineage classification was not performed. SNP differences between recently transmitted isolate pairs and reactivation cases estimated an approximately ten-fold slower mutation rate for latent *M. tuberculosis* compared with active disease (5.5×10^{-10} and 7.3×10^{-11} substitutions/site/generation, respectively (assuming a 20 hr generation time, μ(20hr))) (Colangeli *et al.* 2014). This was in contrast to the situation reported in rhesus macaques, whereby mutation rates estimated for latent and active disease were very similar (2.0×10^{-10} and 2.2×10^{-10} substitutions/site/generation, μ(20hr)). The rate reported by Colangeli *et al.* 2014 translates to a mean rate of $\sim 1.4 \times 10^{-7}$ substitutions/site/year (my own calculations) which is similar to other short-term mutation rate estimates estimated from clinical *M. tuberculosis* isolates (Eldholm *et al.* 2015; Eldholm *et al.* 2016; Ford *et al.* 2013; Pepperell *et al.* 2013; Roetzer *et al.* 2013; Walker *et al.* 2013a).

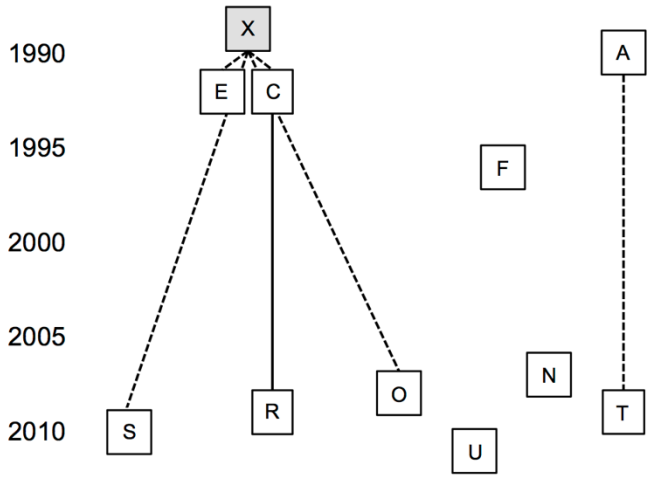


Figure 2.1. M. tuberculosis *Rangipo strain isolates sequenced by Colangeli et al. (2014)*. Chronological representation of epidemiological relationships among ten Rangipo isolates (1991–2011) sequenced on the ABI SOLiD platform. Each square represents a case at time of diagnosis, case "X" (grey box) and is assumed to be the index case for several other cases in the cluster. X was not available for sequencing. Broken lines represent known epidemiological links between cases. Isolates "C" and "R" were taken from the same individual 17 years apart and were shown to be the result of reinfection. Case "T" was the parent of "A", however SNP analysis indicated this was not a direct transmission (Colangeli *et al.* 2014). Modified from Colangeli *et al.* (2014).

Colangeli *et al.* (2014) identified a total of 747 SNPs between the reference strain H37Rv and the ten Rangipo isolates, and 247 high quality SNPs were shared amongst all Rangipo isolates. Ali Ruthe completed her PhD in this laboratory in 2015 and performed further analyses on the 247 common Rangipo SNPs (Ruthe 2015). Ruthe (2015) cross-checked all 247 SNPs against 38 *M. tuberculosis* complex (MTBC) genomes in public databases and *Mycobacterium canettii* to determine if other strains carry the same variants. Ninety-nine SNPs were not identified in any other strains and were classified as 'Rangipo specific' SNPs (RS-SNPs). Three putative RS-SNPs (1380G>A in Rv1821/*secA*; 691G>A in Rv2504c/*scoA*; and 334A>G in Rv3119/*moaE1*) were used to develop a rapid cost-effective diagnostic assay using multiplex polymerase chain reaction (PCR)-restriction fragment length polymorphism (RFLP), allowing classification of strains as Rangipo or non-Rangipo (Ruthe 2015). This assay was found to quickly and accurately distinguish Rangipo strain cases using DNA extracted from cultured isolates with higher discriminatory power than conventional MIRU-VNTR typing (Ruthe 2015). Region of difference PCR analysis identified the Rangipo strain as belonging to the L4/Euro-American lineage of the MTBC but the L4 sublineage remained unassigned (Ruthe 2015).

2.1.1 Objectives

Since the original analysis of Rangipo strain SNP data was performed (Colangeli *et al.* 2014; Ruthe 2015) the number of *M. tuberculosis* strains available in public databases for comparison has increased enormously. The primary objective of this work was to reclassify the 247 Rangipo SNPs and identify candidates of interest for structural and biochemical investigation. It was also important to assign the Rangipo strain to an L4 sublineage and re-evaluate the previous Rangipo PCR-RFLP diagnostic (Ruthe 2015) in light of an updated SNP analysis.

While more thorough phylogenetic and variant analyses were performed once Illumina sequencing of New Zealand cluster isolates was completed (Chapter Four), the analyses described in this chapter are included as they were pivotal to the direction of much of the work undertaken as part of this thesis.

2.1 Methods

2.1.1 Previous Rangipo strain outbreaks

Contact tracing results for three large Rangipo strain outbreaks have previously been reported (Calder 2013; De Zoysa *et al.* 2001; McElnay *et al.* 2004). The number of cases identified with active and latent tuberculosis were analysed for each outbreak. The proportion of total screened contacts with active and latent tuberculosis was calculated and compared to average estimates of tuberculosis prevalence in contacts reported in a large meta-analysis of 108 studies in high-income and 95 studies in low- to middle-income settings (Fox *et al.* 2013).

2.1.2 *M. tuberculosis* genomic DNA

Genomic DNA (gDNA) from clinical New Zealand *M. tuberculosis* cluster isolates was available from previous work in this laboratory (Ruthe 2015). DNA had been either extracted by Roberto Colangeli (University of Medicine and Dentistry of New Jersey, U.S.A.) as in Colangeli *et al.* (2014), or provided as crude clinical isolate extracts from LabPLUS (Auckland, N.Z.). LabPLUS samples included

gDNA from Rangipo, Otara and Southern Cross cluster isolates, and additional smaller New Zealand clusters identified by related MIRU-VNTR typing patterns (NZ_041, NZ_069, NZ_037, NZ_094) (Appendix A.1). Additional gDNA samples, including samples from various MTBC lineages and L4 sublineages, were provided by either LabPLUS or our collaborator Htin Aung (University of Otago, N.Z.).

DNA was quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific, U.S.A.) by measuring absorbance at 260 nm. The ratio of absorbance at 260 nm to 280 nm provides an indication of sample purity and a 260/280 ratio of ~1.8 was considered ideal indicating a DNA sample free from protein, RNA, phenol or other contaminants. Working *M. tuberculosis* genomic DNA samples were prepared by diluting a small amount of stock DNA solution with TE buffer (10 mM Tris-HCL, 1 mM EDTA, pH 8.0) to a concentration of approximately 10 $ng.\mu l^{-1}$.

2.1.3 Identification of putative Rangipo specific SNPs

Ten Rangipo genomes (Figure 2.1) were previously sequenced by collaborators Roberto Colangeli and David Alland at the University of Medicine and Dentistry of New Jersey (U.S.A.). Isolates from 2002 onwards were typed by IS*6110*-RFLP and six isolates (A, C, F, N, O) by MIRU-VNTR 24-locus typing, and share the same molecular typing patterns (IS*6110*-RFLP, 13/008RP; MIRU-VNTR, 233325153324-341444223362) (Ruthe 2015). Isolates were sequenced on an ABI SOLiD 5500XL instrument (Applied Biosystems Inc., U.S.A.) as described in Colangeli *et al.* (2014). Briefly, each isolate was sub-cultured and two individual colonies from each original culture were sequenced. SNPs relative to H37Rv were called using the CLC genomics workbench and SNPs supported by <80% of reads were discarded (Colangeli *et al.* 2014). 247 high quality SNPs present in all 20 genomes were identified.

A list of all 247 SNP positions with the base change and associated gene was provided in excel format. In this thesis, SNPs are named with the prefix “Rv” followed by the position in the H37Rv reference genome (NC_000962.3) and the reference base and nucleotide change as described in Fenner *et al.* (2013). For example, the 1380G>A in the Rv1821/*secA* gene is named Rv2067836GA as it changes the H37Rv nucleotide at position 2067836 from G to A. To identify SNPs

specific to the Rangipo strain, the position of each SNP was cross checked against 1084 global *M. tuberculosis* genomes in the GMTV database to determine if the same variant is present in other isolates (http://mTB.dobzhanskycenter.org) (Chernyaeva *et al.* 2014). Putative non-synonymous RS-SNPs (RS-nsSNPs) identified from this comparison were further checked by submitting protein sequences containing Rangipo variants to Protein BLASTP (BLASTP) and searching against the non-redundant protein sequence database (https://blast.ncbi.nlm.nih.gov) (Altschul *et al.* 1990).

2.1.3.1 In silico *SNP functional assessment*

Functional categories of genes harbouring RS-SNPs were assigned using the Tuberculist database (Lew et al. 2011). To identify SNPs that may impact on protein function the program SNAP2 was used (**S**creening for **N**on-**A**cceptable **P**olymorphisms) (Hecht *et al.* 2015). The H37Rv reference sequence of proteins harbouring RS-nsSNPs were input into the SNAP2 webserver (https://rostlab.org/services/snap2web/), the residue of interest was searched, and the result extracted for the variant amino acid corresponding to the SNP of interest.

2.1.4 Validation of SNPs by Sanger sequencing

2.1.4.1 Primers

Sanger sequencing was used to confirm the presence of RS-nsSNPs. Geneious R8 (Biomatters, N.Z.) and Primer3 (http://primer3.ut.ee) were used to design primers to amplify short genomic regions (168–387 bp) spanning SNPs of interest (Appendix A.3). Primer specificity was checked in Primer-BLAST (Ayres *et al.* 2012). Primers were ordered from Integrated DNA Technologies, Inc. (U.S.A.) and resuspended in 100 μM TE buffer (10 mM Tris-HCL, 1 mM EDTA, pH 8.0). Working stock solutions were prepared at 10 μM in MQ water.

2.1.4.2 Polymerase Chain Reaction

PCR was performed in 20 μl reactions containing 0.5 units (U) HOT FIREPol® DNA polymerase (Solis BioDyne, Estonia), the supplied B1 buffer at 1X

concentration, 2.0 mM $MgCl_2$, 250 µM dNTPs and 0.3 µM of each primer. Approximately 10 ng of gDNA from Rangipo isolate "A" was used as the PCR template and all reactions were run alongside a negative control. A touchdown PCR protocol was employed to reduce the potential for non-specific priming.

Thermal cycling conditions used for amplification were as follows:

95 °C	15:00	(min:sec)
95 °C	0:20	┐
68 °C	0:20	x 5 -1 °C per cycle
72 °C	0:20	┘
95 °C	0:20	┐
63 °C	0:20	x 30
72 °C	0:20	┘
72 °C	1:00	

5 µl of each PCR product was run on an agarose gel (Section 2.1.4.3) to check for a single band of the correct size. PCR products were then ExoISAP treated (Section 2.1.4.4) to remove any remaining primers and free dNTPs before sequencing.

2.1.4.3 Agarose Gel Electrophoresis

DNA fragments were separated by agarose gel electrophoresis. Samples were mixed with 5X DNA loading dye (25% (v/v) glycerol, 0.2% (w/v) bromophenol blue) and separated on 1–2% (w/v) agarose TAE (40 mM TRIS-acetate, 2 mM EDTA) gels made with SYBR safe™ DNA gel stain (Invitrogen, U.S.A.) and run in TAE running buffer. Gels were visualised by blue light and band sizes were determined by visual comparison against a 1 Kb Plus DNA ladder (Thermo Fisher Scientific, U.S.A.).

2.1.4.4 ExoISAP treatment and Sanger sequencing of PCR products

20 U Exonuclease I (Thermo Fisher Scientific, U.S.A.) and 1 U rAPid Alkaline Phosphatase (Roche, Switzerland) were gently mixed with 5 µl of PCR reaction and incubated for 30 min at 37 °C, followed by inactivation at 80 °C for 15 min. ExoISAP treated PCR products were sent to the Massey Genome Service

(Palmerston North, New Zealand) for sequencing on an ABI3730 DNA Analyzer (Applied Biosystems Inc., U.S.A.).

2.1.5 Sublineage classification

2.1.5.1 MIRU based sublineage classification

The MIRU-VNTR*plus* database website (http://www.miru-vntrplus.org) (Allix-Beguec *et al.* 2008) was used perform MIRU-based lineage assignment for the New Zealand clusters based on MIRU-VNTR 24-locus typing patterns (Appendix A.1). The "Identification by similarity search" tool was used with the default categorical weighting of 1 and strict distance cut off of 0.17 to classify lineage based on the best match with reference strains in the database. "Identification by phylogenetic tree" was performed using the UPGMA method to classify any strains without matches and as further verification of the classification.

2.1.5.2 S lineage PCR-RFLP assay

The S lineage, also known as 'S-type', is a sublineage of the L4/Euro-American lineage. Published S lineage marker SNPs are listed in Table 2.1. A PCR-RFLP assay was developed to type for the S-type marker SNP from Homolka *et al.* (2012) (Section 2.2.3.1). The S-type clade identified in Homolka *et al.* (2012) is congruent with the L4.4.1.1 sublineage described in Coll *et al.* (2014).

Table 2.1. *Phylogenetically informative S lineage marker SNPs.*

	SNP	aa change	Locus	Reference
S-type	Rv648992CG	synonymous	Rv0557/*mgtA*	Homolka et al. (2012)
L4.4.1.1/S	Rv355181GA	synonymous	Rv0291/*mycP3*	Coll et al. (2014)

PCR-RFLP enables SNPs to be easily and cheaply detected if one of the variant bases create a recognition sequence for cleavage by a restriction enzyme. WatCut (http://watcut.uwaterloo.ca/template.php) was used to screen both the S-type and the L4.4.1.1/S marker SNPs (Table 2.1) to identify restriction enzymes that would recognise only one of the variant bases. The BstNI restriction enzyme was identified

as a suitable restriction enzyme for the detection of the S-type SNP (Homolka *et al.* 2012) as it is expected to cut non-S samples [CC^WGG] but not samples carrying the S-type marker SNP [CGWGG]. Primers were designed to amplify a 343 bp product containing this marker SNP and include an additional BstNI cut site to act as an internal digestion control (Table 2.2). Expected fragments sizes for S lineage isolates following BstNI digestion are 138 and 205 bp; and for non-S, 138, 151 and 54 bp (Table 2.2). PCR was performed using the same thermal cycling parameters as in Section 2.1.4.2 and 5 µl of product was run on an agarose gel to check for amplification. The remaining PCR product was then digested in a total 25 µl reaction volume with 3 U BstNI (New England Biolabs, U.S.A.), NEBuffer 3.1 and MQ water. Digestion reactions were incubated at 60 °C for 1 hr in a PCR machine and fragments were separated by gel electrophoresis on a 2% gel (Section 2.1.4.3).

Table 2.2. *S lineage PCR-RFLP assay.* PCR primers for the amplification of the S-type Rangipo marker SNP (Rv648992CG) from Homolka *et al.* (2012) and expected fragments following restriction enzyme digestion with BstNI.

Primer name	Primer sequence (5' to 3')	Product size (bp)	Digested fragments (bp)	
			S lineage	Non-S
Rv0557(L153V)_fwd	CTGCTTGGCTACGGTGGACT	343	138, 205	138, 151, 54
Rv0557(L153V)_rev	GAGCAAGCCGACCCACAAAG			

2.1.6 DS6^{Quebec} deletion PCR assays

The DS6^{Quebec} deletion removes an ~11.4 kb region (positions 1987457 to 1998849) replacing it with an IS*6110* element, and is characteristic of the DS6^{Quebec} strain family endemic in Canada (Nguyen *et al.* 2004). A multiplex PCR assay was designed to screen isolates for the presence or absence this deletion using three primers (Table 2.3). A pair of external primers bind upstream and downstream of the DS6^{Quebec} deletion. If the isolate has the deletion, a large 1836 bp PCR product spanning the deletion and its associated IS*6110* element is amplified. An internal primer binds within the DS6^{Quebec} region to pair with the external upstream primer to amplify a small 250 bp product if the region is present as in H37Rv (the external primer pair will not produce a product as the genomic distance between them is too large (11.9 kb)).

A DS6^{Quebec}-IS*6110* insertion-specific PCR using primers described in Nguyen *et al.* (2004) was then performed on isolates testing positive for the deletion to verify the location is exactly the same as in the Canadian DS6^{Quebec} family. Internal IS*6110* primers were used with primers that span the deletion junction so that amplification will only occur if the deletion/IS*6110* insertion occurred at exactly the same site.

Table 2.3. *DS6^{Quebec} deletion assays.* Primer sequences and expected PCR products for the multiplex PCR to screen for the deletion (1), and DS6^{Quebec}-IS*6110* insertion-specific verification of the DS6^{Quebec} deletion (2 and 3). NA = no amplification.

	Primer name/ target region	Primer sequence (5' to 3')	Position in H37Rv[1]	Product (bp) DS6^{Quebec}	Product (bp) H37Rv
1	DS6 5′ upstream	TCAGGCACCGTGACATATTCG	1987278		
	DS6 5′ internal	TCTTAGCCAATAGACTGCCGC	1987527	1836	250
	DS6 3′ downstream	CGCAGTAACTATCGCTGACCTAT	1999149		
2[2]	IS*6110* 3′ junction	AAGCCCCGGCCGCGGCTGGATGAAC	1987438	161	NA
	IS*6110* 3′ internal	TTCAACCATCGCCGCCTCT	N/A		
3[2]	IS*6110* 5′ internal	GGTACCTCCTCGATGAACC	N/A	99	NA
	IS*6110* 5′ junction	AGCCAGCCAACCCCGGCCCTTGAAC	1998868		

[1] Position of 5′ end of primer in the H37Rv reference genome.
[2] Primer sequences from Nguyen *et al.* (2004). Note that due to the direction of the IS*6110* element, the IS*6110* 3′ primers amplify the 5′ flank of the deletion and the IS*6110* 5′ primers amplify the 3′ flank.

PCR reactions were prepared as per Section 2.1.4.2. Touchdown PCR was used to reduce the potential for non-specific priming for all PCR reactions. Thermal cycling conditions used for the DS6^{Quebec} multiplex PCR were as follows:

95 °C 15:00 (min:sec)
95 °C 0:20 ┐
68 °C 0:20 │ x 5 -1 °C per cycle
72 °C 2:00 ┘
95 °C 0:20 ┐
63 °C 0:20 │ x 30
72 °C 2:00 ┘
72 °C 3:00

Thermal cycling parameters used for DS6^{Quebec}-IS*6110* insertion-specific PCRs:

95 °C	15:00	(min:sec)
95 °C	0:20	┐
70 °C	0:20	│ x 5 -1 °C per cycle
72 °C	0:20	┘
95 °C	0:20	┐
65 °C	0:20	│ x 30
72 °C	0:20	┘
72 °C	3:00	

PCR products were analysed by agarose gel electrophoresis (Section 2.1.4.3). Sanger sequencing of PCR products as in Section 2.1.4.4 to confirm the correct genomic regions had been amplified and as additional confirmation of the deletion.

2.1.7 Contig mapping and analysis

SNP analyses identified SNPs shared between the Rangipo strain and Canadian SUMu strain isolates (Section 2.2.2). Genomic contigs for 12 Canadian SUMu *M. tuberculosis* isolates (SUMu001–SUMu012) were downloaded as fasta files from the Broad Institute (TB Natural Mutation Rate Sequencing Project, Broad Institute of Harvard and MIT, http://www.broadinstitute.org/). Contigs were mapped to the H37Rv reference genome (NC_000962.3) in Geneious R8 (Biomatters, N.Z.) using default settings. Variant positions of interest and the DS6^{Quebec} deletion region (positions 1987457 to 1998849) were manually examined.

2.1.8 Rangipo diagnostic assay

A Rangipo specific SNP-based multiplex PCR-RFLP diagnostic assay has previously been developed in this laboratory (Ruthe 2015). Based on the updated analyses performed in this work, the Rangipo diagnostic was revised to type for a single RS-SNP, Rv2067836GA in the Rv1821/*secA2* gene.

2.1.8.1 Rangipo PCR-RFLP assay

The Sau3AI restriction enzyme and its isoschizomer MboI were previously identified as suitable restriction enzymes for the detection of the Rangipo Rv1821/*secA2* SNP as they are expected to cut non-Rangipo samples [^GATC], but not samples carrying the Rangipo SNP [AATC] (Ruthe 2015). A new forward primer was designed to be used with the original reverse primer to incorporate an additional Sau3AI/MboI cut site to act as in internal digestion control. Primers were designed and prepared as in Section 2.1.4.1. Primers are expected to amplify a 455 bp PCR product spanning the SNP and restriction enzyme digestion of this product is expected to produce 386 and 69 bp fragments for Rangipo strain isolates, and 215, 171 and 69 bp fragments for non-Rangipo (Table 2.4).

Table 2.4. *Rangipo specific PCR-RFLP diagnostic assay.* PCR primers for the amplification of the Rv1821/*secA2* Rangipo marker SNP (Rv2067836GA) and expected fragments following restriction enzyme digestion with Sau3AI or its isoschizomer MboI.

Primer name	Primer sequence (5' to 3')	Product size (bp)	Digested fragments (bp)	
			Rangipo	Non- Rangipo
Rv1821 Fwd new	GAGGCCAAGGAAGGTATCGAGAC	455	386, 69	215, 171, 69
Rv1821 Rev[1]	CGCCCGGCCATTTGAGTTGA			

[1] Reverse primer from Ruthe (2015).

PCR reactions were carried out in 20 µL volumes with 1 U HOT FIREPol® DNA Polymerase (Solis BioDyne, Estonia), supplied buffer B1 at 1X concentration, 2.0 mM $MgCl_2$, 250 µM dNTPs and 0.3 µM of each primer. Approximately 10 ng of *M. tuberculosis* template gDNA of was added to each reaction tube and all reactions were run alongside a no template negative control.

The following thermal cycling conditions were used for amplification:

95 °C	15:00	(min:sec)
95 °C	0:20	┐
55 °C	0:20	│ x 30
72 °C	0:30	┘
72 °C	1:00	

PCR products were digested with MboI in order to distinguish between Rangipo and non-Rangipo isolates. 1 µl MboI (New England Biolabs, U.S.A.), NEB buffer 4.1 at 1X final concentration, and MQ H_2O were added to the PCR reaction to a final volume of 50 µl. This was incubated at 37 °C for 1 hr in a PCR machine, followed by 65 °C for 20 min. Digested products were separated on 1.5% agarose gels (Section 2.1.4.3).

2.1.8.2 Sputum samples

Clinical sputum specimens were provided by the Waikato Hospital (Hamilton, N.Z.) to assay with the Rangipo PCR-RFLP diagnostic. Samples had been decontaminated using the BD BBL MycoPrep Kit (BD, U.S.A.), heat inactivated at 95 °C for 20 min and then spun down at 4,500 rpm for 10 minutes by hospital laboratory staff. PCR was performed using 0.5 µl of the supernatant as template and products were digested as in Section 2.1.8.2.

2.1.9 Ethical Considerations

Ethical approval is not required for sequencing and DNA analysis of New Zealand clinical *M. tuberculosis* bacterial isolates. Ethical approval was not required for Rangipo diagnosis from sputum samples as this was a public health investigation seeking to improve the turnaround time for pathogen strain typing and samples were used solely for the purpose they were collected for (diagnosis and strain typing of mycobacterial pulmonary infection), and destroyed immediately after. No identifiable health information was provided to researchers.

2.2 Results

2.2.1 Previous Rangipo strain outbreaks

Contact tracing summary data was available for three large Rangipo strain outbreaks that occurred between 1996 and 2012 (Calder 2013; De Zoysa *et al.* 2001; McElnay *et al.* 2004). Collectively these outbreaks resulted in at least 94 cases of

active tuberculosis, of which 68 where identified through contact tracing (additional cases were notified to public health services) and 253 cases of latent infection. In total, 4.6% (68/1482) of traced contacts were diagnosed with active tuberculosis (95% CI, 4.0–5.0%). This is over three times higher than the estimated average prevalence of active tuberculosis in traced contacts in high-income settings (1.4%; 95% CI, 1.1–1.8%) and 1.5 times higher than for low- and middle-income settings (3.1%; 95% CI, 2.2–4.4%) (Fox *et al.* 2013). Furthermore, this excludes additional cases identified by notification to public health services. The prevalence of latent infection in traced contacts was 17.1% (253/1482) (95% CI, 11.7–33.5%), which is lower than the estimated average for low- to middle-income settings (51.5%; 95% CI, 47.1–55.8%) and lower but within the 95% CI for high-income settings (28.1%; 95% CI, 24.2–32.4%) (Fox *et al.* 2013). In each Rangipo strain outbreak, the proportion of infected contacts with active disease exceeded the 5–10% expected to progress to active disease, and overall 21.2% of infected individuals (68/321) were diagnosed with active tuberculosis (95% CI, 11.7–33.5%).

Table 2.5. *Contact tracing of previous Rangipo strain outbreaks*. Percentage of traced contacts is shown in brackets.

	North Island 1996-2000[1]	Hawkes Bay 2002[2]	Hawkes Bay 2012[3]	Total
Traced contacts	733	397	352	1482
Active tuberculosis	35 (4.8%)	19 (4.8%)	14 (4.0%)	68 (4.6%)
Latent tuberculosis	119 (16.2%)	40 (10.1%)	94 (26.7%)	253 (17.1%)
Total infected	154 (21.0%)	59 (14.9%)	108 (30.70%)	321 (21.7%)
Infected contacts with active disease	22.7%	32.2%	13.0%	21.2%
Contact types	Community and prison contacts	Household, close associates, and co-workers	Household, and shearing gang workers	

[1] De Zoysa *et al.* (2001)
[2] McElnay *et al.* (2004)
[3] Calder (2013)

2.2.2 Rangipo SNPs

Two hundred and forty-seven SNPs common to the Rangipo cluster were previously identified in ten Rangipo genomes sequenced on the ABI SOLiD platform (Colangeli *et al.* 2014). I then checked all 247 SNP positions by searching the GMTV database for the presence of the same variation in >1000 MTBC

genomes (accessed March 2015). Eleven of the 247 SNPs were found not to be SNPs relative to H37Rv (Table 2.6, Appendix A.2). Of the 236 remaining positions, 210 SNPs were present in at least one other strain and 26 were not present in any other isolates, and thus represent putative RS-SNPs. Nearly two-thirds of all SNPs were non-synonymous SNPs (nsSNPs) (150/236, 63.5%), consistent with that typically reported for *M. tuberculosis* (Fleischmann *et al.* 2002; Hershberg *et al.* 2008; Rose *et al.* 2013). A similar proportion of nsSNPs were observed for the RS-SNP subset (16/26, 61.5%).

Table 2.6. *Classification of 247 common Rangipo SNPs.* SNPs classified as Rangipo specific were not present in >1000 M. tuberculosis isolates in the GMTV database.

Classification	sSNPs	nsSNPs	Total
Not a SNP relative to H37Rv	–	–	(11)
Present in at least one other strain	76	134	210
Rangipo specific SNP (RS-SNPs)	10	16	26
Total	86	150	236

Amino acid sequences of the 16 proteins with RS-nsSNPs were submitted to BLAST to search other *M. tuberculosis* proteins for the same amino acid changes. This identified two amino acid changes (Rv0071 R67L and Rv1860/*apa* T146N) shared by the Rangipo strain and three *M. tuberculosis* isolates from Canada – SUMu002, SUMu004 and SUMu005.

The Canadian SUMu genomes carrying the Rv0071 and Rv1860 variants are from three of 12 *M. tuberculosis* isolates from Aboriginal communities in Western Canada sequenced using Roche 454 whole genome shotgun methodology (Pepperell *et al.* 2013). All 12 SUMu genomes are publicly available as contigs (raw fastq were files not available). These were downloaded and mapped to the H37Rv reference genome in Geneious R8 and all 26 putative RS-SNP positions were manually examined. SNPs in Rv0071 and Rv1860 were found to match those in the Rangipo strain, and two additional synonymous Rangipo SNPs; Rv1088676CT in Rv0976c and Rv2491388CT in Rv2221c/*glnE*, were also present. These four SNPs were all found in the SUMu002, SUMu004 and SUMu005 genomes, but not in any other Canadian SUMu genomes.

Excluding these four SNPs, a final list of 22 RS-SNPs was produced (Appendix A.2). Fourteen of these were non-synonymous and thus have potential effects on protein function and bacterial fitness (Table 2.7).

Table 2.7. *Non-synonymous Rangipo specific SNPs.*

SNP	Locus	aa change	Gene product
Rv3253CG	Rv0002/*dnaN*	P401R	DNA polymerase III beta sliding clamp subunit dnaN
Rv489437CT	Rv0405/*pks6*	S1236L	Probable membrane bound polyketide synthase Pks6
Rv550620GT	Rv0458	D316Y	Probable aldehyde dehydrogenase
Rv1289731TC	Rv1161/*narG*	Y802H	Respiratory nitrate reductase alpha chain NarG
Rv1836099TG	Rv1631/*coaE*	Y363D	Dephospho-CoA kinase CoaE
Rv2807374GC	Rv2492	G33R	Hypothetical protein
Rv3202633GA	Rv2893	G72S	Possible oxidoreductase
Rv3283879TC	Rv2941/*fadD28*	V182A	Long-chain-fatty-acid-AMP ligase FadD28
Rv3366098GA	Rv3007c	P118L	Possible oxidoreductase
Rv3561770AG	Rv3193c	L468S	Probable conserved transmembrane protein
Rv3895925GA	Rv3479	A36T	Possible transmembrane protein
Rv3980075CA	Rv3540c/*ltp2*	E195D	Probable lipid transfer protein or keto acyl-CoA thiolase ltp2
Rv4085870GA	Rv3646c/*topA*	T463M	DNA topoisomerase I topA
Rv4377908CT	Rv3894c/*eccC2*	G849S	EESX-2 type VII secretion system protein eccC2

The final RS-SNP list excluded two SNPs previously used as diagnostic markers; Rv2819180CT in Rv2504c/*scoA* and Rv3485475AG in Rv3119/*moaE1* (Ruthe 2015), as these were identified in other isolates in the GMTV database. The third diagnostic marker SNP Rv2067836GA in Rv1821/*secA2*, was not identified in any other isolates and remains classified as an RS-SNP. This was further validated by comparison against 220 *M. tuberculosis* L4.4 genomes in Chapter Four (Section 4.3.2.2).

One RS-nsSNP (Rv489437CT in Rv0405/*pks6*) and one SUMu shared SNP (Rv79685GT in Rv0071) have previously been validated in the Rangipo strain (Ruthe 2015). In this work, the remaining 13 RS-nsSNPs were confirmed by Sanger sequencing (Section 2.1.4).

2.2.2.1 Gene functional categories

Functional categories for genes carrying all 236 Rangipo SNPs were assigned using the Tuberculist database (Lew et al. 2011). Of the SNPs common to the Rangipo strain, genes in cell wall and cell processes were the most highly represented accounting for 27.5% of all SNPs (65/236), and 30.0% of nsSNPs (45/150) (Table 2.8). This was followed by the intermediary metabolism and respiration category, which accounted for 21.2% of all SNPs (50/236) and 20.7% of nsSNPs (31/150).

Genes in the intermediary metabolism and respiration category were the most highly represented among the RS-SNPs accounting for 31.8% of all RS-SNPs (7/22) and 35.7% of RS-nsSNPs (5/14); followed by cell wall and cell processes, 22.7% of all RS-SNPs (5/22), and 21.4% of RS-nsSNPs (3/14). The proportion of genes involved in lipid metabolism and in intermediary metabolism and respiration were enriched ~1.7-fold in the RS-nsSNP subset compared to the full set of nsSNPs. These two functional categories collectively accounted for 57.1% (8/14) of variant genes unique to the Rangipo strain; an approximate two-fold enrichment relative to their overall proportion of 28% in the *M. tuberculosis* genome (Camus *et al.* 2002) (Table 2.8). Genes involved in lipid metabolism were particularly over represented. This category accounting for 12.0% of common Rangipo nsSNPs (18/150) and 21.4% RS-nsSNPs (3/14), a two- and 3.5-fold enrichment, respectively, compared an overall proportion of just 6% in the *M. tuberculosis* H37Rv genome.

Although no RS-SNPs were identified in genes classified in the virulence, detoxification, adaptation functional category, four RS-nsSNPs occurred in genes encoding virulence factors listed in the PATRIC_VF and/or the VFDB databases: Rv0405/*pks6*, Rv1161/*narG*, Rv3540c/*ltp2*, and Rv2941/*fadD28*.

Table 2.8. *Functional categories of genes harbouring Rangipo SNPs.*

Gene Functional Category	Common Rangipo SNPs			Rangipo specific SNPs			H37Rv[1]
	sSNPs	nsSNPs	All SNPs	sSNPs	nsSNPs	All SNPs	
Cell wall and cell processes	20 (23.3%)	45 (30.0%)	65 (27.5%)	2 (25.0%)	3 (21.4%)	5 (22.7%)	18%
Conserved hypotheticals	16 (18.6%)	24 (16.0%)	40 (16.9%)	2 (25.0%)	1 (7.1%)	3 (13.6%)	26%
Information pathways	7 (8.1%)	17 (11.3%)	24 (10.2%)	1 (12.5%)	2 (14.3%)	3 (13.6%)	6%
Insertion seqs and phages[2]	2 (2.3%)	1 (0.7%)	3 (1.3%)	0	0	0	4%
Intermediary metabolism and respiration	19 (22.1%)	31 (20.7%)	50 (21.2%)	2 (25.0%)	5 (35.7%)	7 (31.8%)	22%
Lipid metabolism	10 (11.6%)	18 (12.0%)	28 (11.9%)	1 (12.5%)	3 (21.4%)	4 (18.2%)	6%
Regulatory proteins	7 (8.1%)	9 (6.0%)	16 (6.8%)	0	0	0	5%
Virulence, detoxification, adaptation	5 (5.8%)	5 (3.3%)	10 (4.2%)	0	0	0	2%
Total	86	150	236	8	14	22	(89%)

[1] Percentage of genes in each functional category in the H37Rv reference genome as reported in Camus *et al.* (2002)
[2] This gene category is underrepresented as the majority were excluded from SNP calling analyses due to the difficulties with mapping repetitive regions of the genome.

2.2.2.2 SNP functional effect prediction

SNAP2 is a freely available machine learning based classifier trained to distinguish between effect and neutral amino acid variations (Hecht *et al.* 2015). It considers a variety of sequence and variant features, including automatically generated multiple sequence alignments (which provide the most important signal), structural features, and annotations if available, to predict whether or not a single amino variant changes molecular function. For each residue and possible substitution, the program calculates an output score ranging from -100 (strong neutral) to +100 (strong effect), reflecting the likelihood that a specific mutation alters protein function, as well as the expected accuracy of the prediction.

All 14 RS-nsSNPs were analysed with SNAP2 to identify those with predicted functional effects. Eight SNPs were predicted to have an effect, including those in the virulence-associated proteins Rv0405/*pks6* and Rv1161/*narG* (Table 2.9). Around two-thirds of SNPs predicted to have functional consequences were found in genes involved in intermediary metabolism and respiration (Rv0458, Rv1161/*narG*, Rv1631/*coaE*, Rv2893 and Rv3007c). This gene functional category was also over represented among RS-nsSNPs compared with the proportion of genes in category in the *M. tuberculosis* genome (Section 2.2.2.2). The remaining three SNPs predicted to have an effect were found in genes involved in cell wall and cell processes (Rv3894c/*eccC2*), conserved hypotheticals (Rv2492), and in lipid metabolism (Rv0405/*pks6*).

The highest ranked SNP in terms of predicted effect, Y363D in Rv1631/*CoaE,* was further examined *in silico* (Section 2.2.2.3). The Rv2893 G72S SNP was also further examined (Section 2.2.2.4) and is the subject of biochemical and structural investigation in Chapter Three.

Table 2.9. *SNAP2 functional effect predictions for RS-nsSNPs.* SNAP2 output is shown in the predicted effect, score and expected accuracy columns. Known virulence factor genes are listed in bold type. SNPs are ranked according to effect prediction score.

Locus	SNP	Predicted Effect	Score	Expected Accuracy	Gene Functional Category
Rv1631/*coaE*	Y363D	effect	87	91%	Intermediary metabolism and respiration
Rv1161/*narG*	Y802H	effect	80	91%	Intermediary metabolism and respiration
Rv2492	G33R	effect	79	85%	Conserved hypotheticals
Rv3007c	P118L	effect	59	75%	Intermediary metabolism and respiration
Rv0405/*pks6*	S1236L	effect	46	71%	Lipid metabolism
Rv2893	G72S	effect	28	63%	Intermediary metabolism and respiration
Rv3894c/*eccC2*	G849S	effect	20	63%	Cell wall and cell processes
Rv0458	D316Y	effect	14	59%	Intermediary metabolism and respiration
Rv3646c/*topA*	T463M	neutral	-5	53%	Information pathways
Rv3193c	L468S	neutral	-17	57%	Cell wall and cell processes
Rv0002/*dnaN*	P401R	neutral	-26	61%	Information pathways
Rv3540c/*ltp2*	E195D	neutral	-67	82%	Lipid metabolism
Rv2941/*fadD28*	V182A	neutral	-77	87%	Lipid metabolism
Rv3479	A36T	neutral	-80	87%	Cell wall and cell processes

2.2.2.3 Rv1631/coaE Y363D SNP

The Y363D SNP in Rv1631/*coaE* was predicted as the most significant SNP in SNAP2 analyses (Section 2.2.2.2). Examination of *M. tuberculosis* CoaE domain architecture in PFAM shows it is comprised of an N-terminal catalytic dephosphocoenzyme A kinase (CoaE) domain and a C-terminal GrpB domain (Figure 2.2A). The Y363D SNP is located in the GrpB domain 45 residues from the C-terminus. The GrpB PFAM HMM logo reveals that the Y363D SNP changes one of four highly conserved residues comprised of two tyrosines and two lysines (Y-K-Y-K) near the C-terminal end of the domain (Figure 2.2B).

Figure 2.2. *Domain structure of* M. tuberculosis *CoaE and position of the Y363D RS-nsSNP.* (A) PFAM domain architecture shows Y363D is located in the GrpB domain 45 bp from the C-terminal end of the protein. (B) The PFAM HMM logo for the GrpB domain shows four highly conserved Y-K-Y-K residues at the C-terminal end of this domain. Y363D changes the first of these conserved sites and is highlighted by a red box.

The GrpB domain family is suggested to belong to the nucleotidyltransferase superfamily (Kuchta *et al.* 2009). This is a large and diverse family of proteins, nearly all of which catalyse the transfer of a nucleotide to an acceptor hydroxyl group (Kuchta *et al.* 2009). There are no structures available for *M. tuberculosis* CoaE and the only crystal structure of a GrpB domain is that of conserved protein GrpB of unknown function from *Enterococcus faecalis* (PDB ID: 2NRK). 2NRK shares 24% identity with the GrpB domain of *M. tuberculosis* CoaE, including the four conserved Y-K-Y-K residues and two sites identified as catalytic residues on the basis of structural homology (Asp52 and His99) (Kuchta *et al.* 2009) (Figure 2.3A). The Rangipo Y363D SNP position corresponds to Tyr131 in the 2NRK

structure. Examination of 2NRK shows the conversed Y-K-Y-K residues are positioned adjacent to one another on two neighbouring alpha helixes forming a highly conserved motif (Figure 2.3). The highly conserved nature and chemical properties of the Y-K-Y-K motif and its position across from the active site Asp52 and His99 residues suggests this motif might be important for substrate binding in the GrpB protein family. However, in *M. tuberculosis* CoaE the only currently assigned function of the GrpB domain is in the proper protein folding of the enzyme and the last 50 residues at the C-terminus (including the Y363D SNP and all four conserved Y-K-Y-K residues) are dispensable for proper folding and full catalytic activity *in vitro* (Walia *et al.* 2009). Therefore this SNP may be less likely to affect CoaE activity than expected based on the prediction from SNAP2.

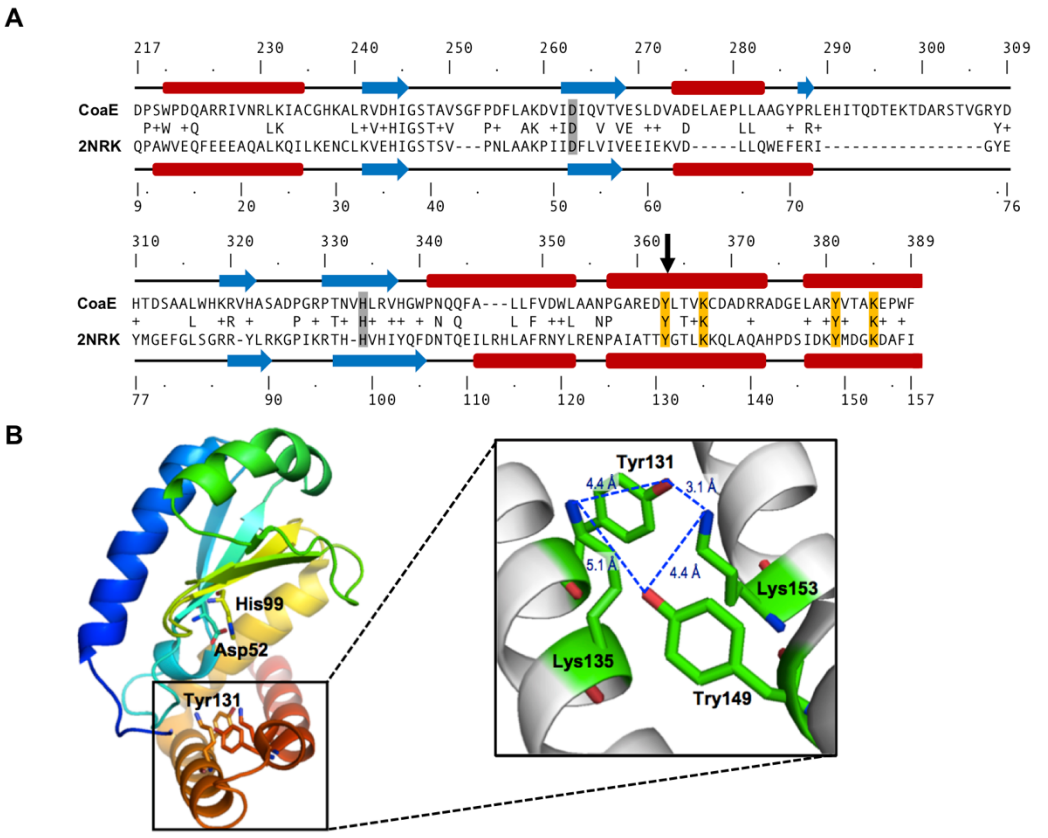


Figure 2.3. *The conserved protein GrpB from* E. faecalis *(2NRK).* The position of conserved Y-K-Y-K residues are indicated. The position of the Y363D RS-SNP corresponds to residue Tyr131 in 2NRK. (A) Sequence alignment and secondary structure elements of the *M. tuberculosis* CoaE GrpB domain and 2NRK. A black arrow indicates the position of Y363D. Catalytic residues Asp52 and His99 are highlighted in grey and the conserved Y-K-Y-K residues in orange. Secondary structure elements (alpha helixes red, beta sheets blue) were predicted using JPred4. (B) The 2NRK structure showing the position of the Y-K-Y-K motif and catalytic residues.

2.2.2.4 Rv2893 G72S SNP

Rv2893 is predicted to be a F_{420}-dependent oxidoreductase and belongs to the luciferase like hydride transferase (LLHT) family of enzymes (previously known as luciferase like monooxygenases) (Selengut and Haft 2010). A PDB protein sequence search was performed to identify deposited structures with sequence similarity to Rv2893 (accessed 2015). Four of the top five unique protein hits were for F_{420}-dependent LLHTs; F_{420}-dependent glucose-6-phosphate dehydrogenase from *M. tuberculosis* (mtbFGD, PDB ID: 3B4Y and 3C8N) (Bashiri *et al.* 2008), methylenetetrahydromethanopterin reductase from *Methanosarcina barkeri* (bMer, 1Z69) (Aufhammer *et al.* 2005), Mer from *Methanobacterium thermoautotrophicum* (tMer, 1F07), and Mer from *Methanopyrus kandleri* (kMer, 1EZW) (Shima *et al.* 2000). These share 21–32% protein sequence identity with Rv2893. All remaining proteins identified were non-F_{420}-dependent LLHTs.

A sequence alignment of Rv2893 with mtbFGD, bMer, tMer, kMer and MSMEG_2516 (the *Mycobacterium smegmatis* homologue of Rv2893) shows Rv2893 has the conserved glycine in position 179 that enables binding of the phosphate group of F_{420} without steric hindrance (Aufhammer *et al.* 2004), consistent with it being an F_{420}-dependent protein (Figure 2.4). Rv2893 also retains the conserved active site histidine (His48) and a second histidine occupies the site of the catalytic glutamate (His109) (His40 and Glu109 in mtbFGD (Oyugi *et al.* 2016)). The Rangipo G72S SNP site corresponds to Gly71 in mtbFGD and is conserved in the alignment. In mtbFGD, Gly71 is positioned two residues from Ser73 and Val74, which are connected by a rare non-prolyl *cis*-peptide bond (Bashiri *et al.* 2008). Examination of the mtbFGD structure shows Gly71 is positioned at the end of a beta sheet next to the base of the non-prolyl *cis*-peptide bond (Figure 2.5). This *cis*-peptide bond is found in other F_{420}-LLHTs and is important for F_{420} binding and catalysis. It creates a bulge that packs against the *Re*-face of F_{420} that helps induce a butterfly bend in the ring system of F_{420}, increasing the reactivity of the C5 atom (which is oxidised or reduced in the reaction) (Aufhammer *et al.* 2004; Aufhammer *et al.* 2005; Bashiri *et al.* 2008; Shima *et al.* 2000). The location of the G72S mutation near at the base of this bulge could have important consequences for F_{420}-binding and catalysis via influence on the bulge and its interaction with F_{420}.

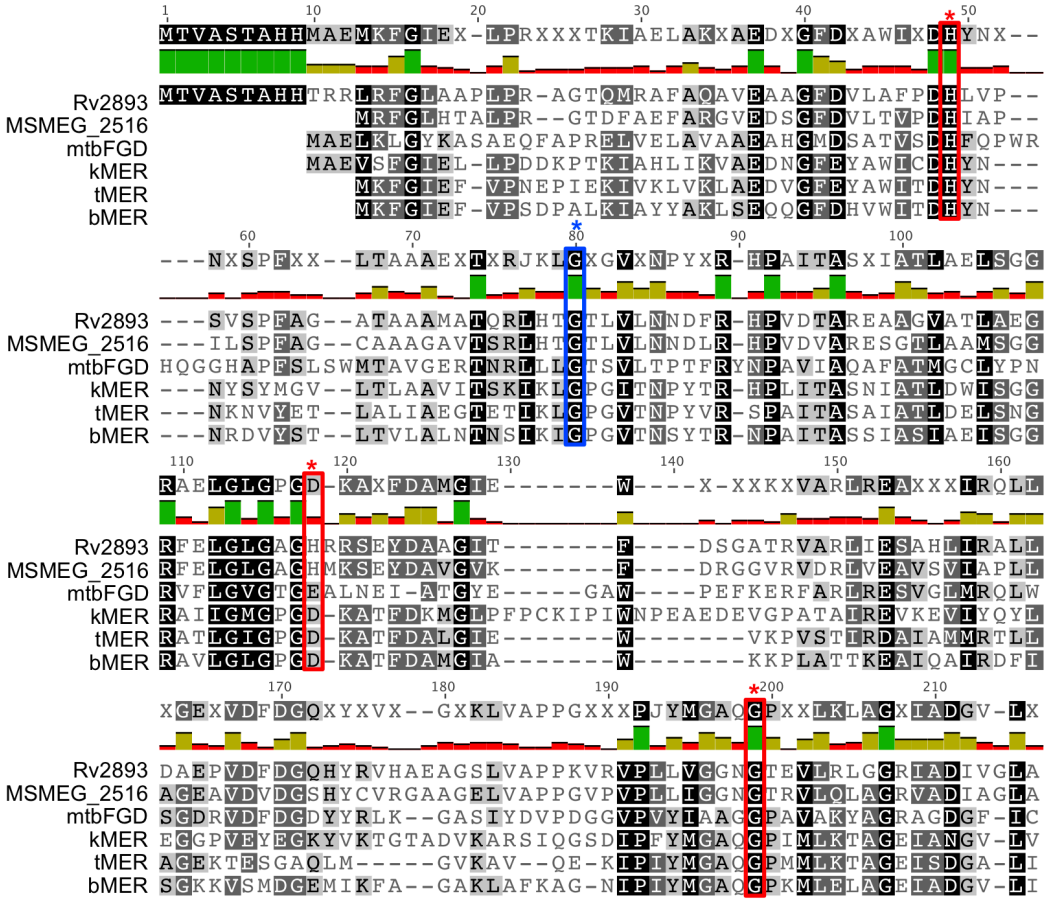


Figure 2.4. *Multiple sequence alignment of F_{420}-dependent of bacterial LLHT family proteins.* Sequences are shown for *M. tuberculosis* Rv2893, the *M. smegmatis* orthologue MSMEG_2516, and three F_{420}-dependent LLHTs identified by sequence similarity searches in the PDB (mtbFGD, kMER, tMer and bMER). Catalytic resides (mtbFGD His40 and Glu109) and the conserved glycine that differentiates F_{420} -binding LLHTs are marked with red boxes and asterisks. The Rangipo specific G72S mutation is marked by a blue box and asterisk. Only the first 196 residues of Rv2893 are shown.

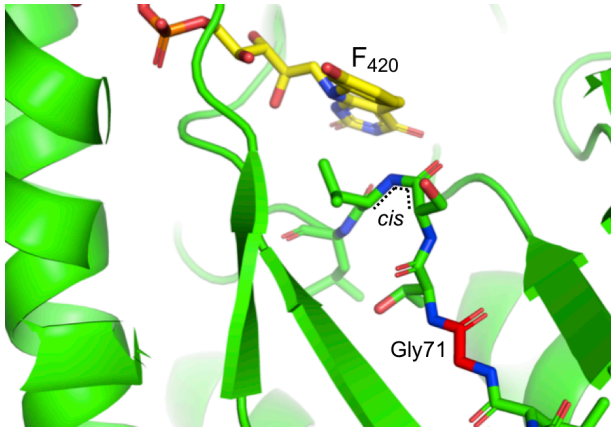


Figure 2.5. *Corresponding position of the Rangipo G72S SNP site* in M. tuberculosis *FGD (3B47).* The non-prolyl *cis*-peptide bond linking Ser73 and Val74 is indicated and Gly71, (position of the Rangipo SNP), is shown in red and labelled. F_{420} is coloured yellow.

2.2.3 Lineage classification

MIRU-VNTR-based lineage assignment was performed using the MIRU-VNTR*plus* database. This classified the Rangipo, Otara and NZ_094 clusters as belonging to the S lineage, Southern Cross and NZ_037 as LAM, and NZ_041 and NZ_069 as H37Rv-like.

2.2.3.1 S lineage classification PCR-RFLP assay

The S lineage, or 'S-type', was originally classified based on traditional molecular typing methods and is equivalent to the L4.4.1.1/S sublineage described in a recent high-resolution SNP phylogeny (Coll *et al.* 2014). L4.4.1.1/S belongs to the L4.4 sublineage of the L4/Euro-American lineage and L4.4 comprises three clades; L4.4.1.1/S, L4.4.1.2 and L4.4.2. L4.4 is reported to account for 10% of global L4 isolates and is the most common L4 sublineage in New Zealand, accounting for 43% of L4 cases (Stucki *et al.* 2016b).

To confirm the assignment of the Rangipo, Otara and NZ_094 clusters to the S lineage and to provide a fast and affordable method to identify other S lineage isolates, a SNP-based RFLP-PCR assay was designed. WatCut identified that the S-type marker SNP Rv648992CG described by Homolka *et al.* (2012) is selectively recognised using the restriction enzyme BstNI. A 343 bp product spanning the S-type Rv648992CG marker SNP was amplified by PCR and digested with BstNI. Sanger sequencing of the PCR product confirmed the correct 343 bp fragment had been amplified and the presence of the marker SNP in the Rangipo strain. Digestion with BstNI gave the expected fragments of 205 and 138 bp for S-type, and 138, 151 and 54 bp for non-S-type, allowing classification of isolates as S lineage or non-S (Figure 2.6A,B). Due to size similarity, the 138 and 151 bp fragments for non-S isolates migrated as a single band on the gel. This assay was used to screen New Zealand *M. tuberculosis* cluster isolates (Figure 2.6C). The Rangipo, Otara and the NZ_094 clusters were all classified as belonging to the S lineage and Southern Cross and the remaining clusters as non-S, consistent with MIRU-VNTR based lineage assignment.

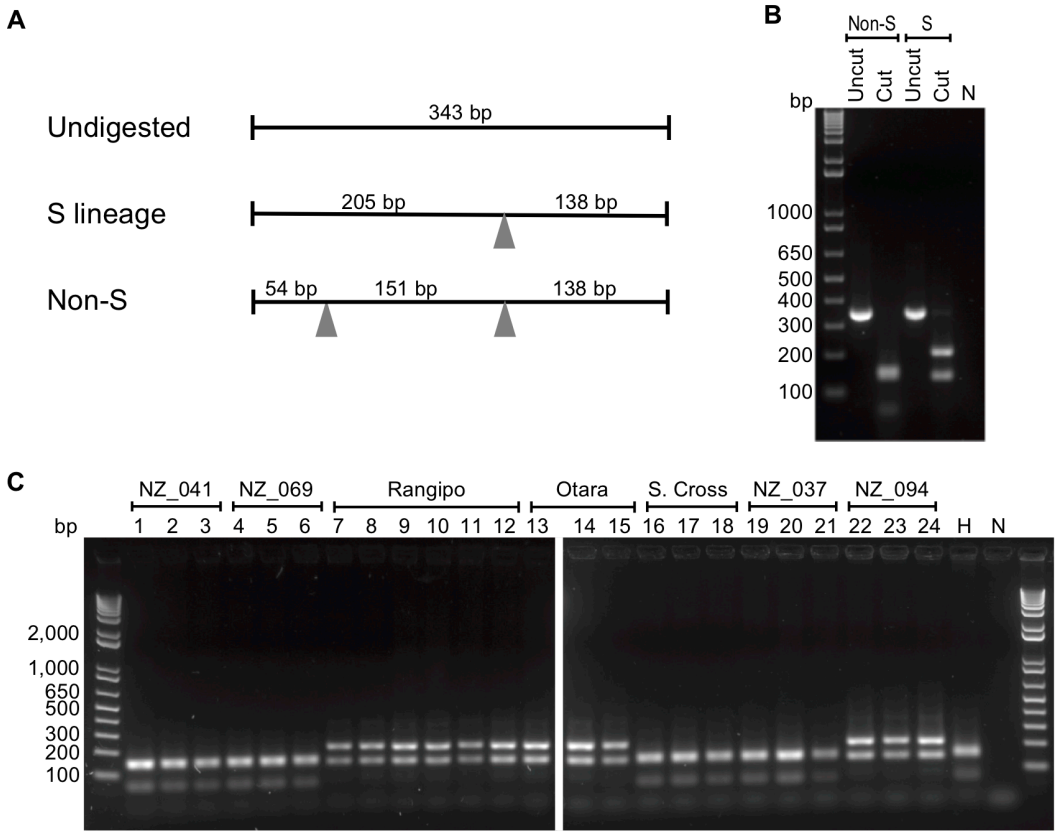


Figure 2.6. *S lineage PCR-RFLP assay.* (A) Schematic showing positions the BstNI recognition sequences (arrows) and expected fragment sizes. (B) Agarose gel showing uncut PCR products and digested fragments for non-S (H37Rv) and S lineage (Rangipo A) isolates. (C) S lineage typing of New Zealand *M. tuberculosis* cluster isolates. MIRU-VNTR cluster ID is shown above the lane number (lanes 1-3, NZ_041; 4-6, NZ_069; 7-12, Rangipo; 13-15, Otara; 16-18, Southern Cross; 19-21, NZ_037; 22-24, H, reference strain H37Rv; N, PCR negative control).

2.2.4 DS6^{Quebec} deletion

BLAST searches identified two SNPs shared only with Rangipo and three *M. tuberculosis* SUMu genomes from Aboriginal communities in Canada (Section 2.2.1). The most common *M. tuberculosis* lineage in Western Canadian Aboriginal populations and in French Canadian populations in Quebec is the DS6^{Quebec} strain family (Pepperell *et al.* 2011). Spoligotyping patterns of DS6^{Quebec} isolates show that this strain family belongs to the S lineage (Nguyen *et al.* 2003). The DS6^{Quebec} family is characterised by the presence of a large genomic deletion called the DS6^{Quebec} deletion. This removes ~11.4 kb between positions 1987457 and 1998849, replacing it with an IS*6110* element (Nguyen *et al.* 2004) (Figure 2.7). A similar but evolutionarily independent ~12 kb deletion (positions 1986636 to

1998621) known as the RD152 deletion is also found in the hypervirulent Beijing/W strain (Tsolaki *et al.* 2005).

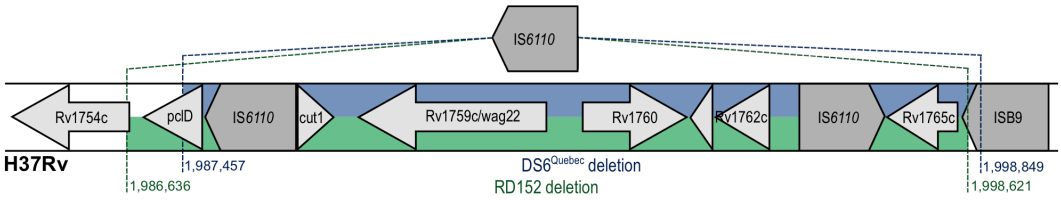


Figure 2.7. *Schematic of the DS6^{Quebec} and RD152 deletions.* Genomic regions present in H37Rv that are deleted by the DS6^{Quebec} deletion and the RD152 deletion in the Beijing/W strain are shown in blue and green, respectively.

All 12 *M. tuberculosis* SUMu genomes were examined for the presence of S lineage marker SNPs (Table 2.1) and evidence of the DS6^{Quebec} deletion, as indicated by absence of coverage over the region. Eight of these genomes, including the three sharing SNPs with the Rangipo strain, had the DS6^{Quebec} deletion and belong to the S lineage. The remaining four genomes were non-S lineage and did not have the DS6^{Quebec} deletion. The identification of Rangipo SNPs shared with a subset of S lineage/DS6^{Quebec} SUMu isolates suggested that the Rangipo strain might have a close phylogenetic relationship with the Canadian DS6^{Quebec} lineage and harbour the DS6^{Quebec} deletion.

PCR assays were used to screen Rangipo and other New Zealand *M. tuberculosis* cluster isolates for the DS6^{Quebec} deletion (Figure 2.8). A multiplex PCR was used to first identify isolates with deletions in the DS6 region using forward and reverse primers external to the deletion together with an internal reverse primer (Figure 2.8A,C,D). If an isolate has the DS6^{Quebec} deletion, a large 1836 bp product is amplified from the external primers. If the region is not deleted, a short 250 bp product is amplified from the external upstream and the internal primer, and the external primers will not amplify as the genomic distance between them is too large (11.9 kb). Two separate DS6^{Quebec}-IS*6110* insertion-specific PCRs as described in Nguyen *et al.* (2004) where then performed (Figure 2.8B,C,E). These use external deletion flanking primers that include 5 bp of the end of the IS*6110* sequence and 20 bp of the sequence flanking the deletion, so that amplification will only occur if IS*6110* insertion has occurred at exactly the same site (Figure 2.8B,E).

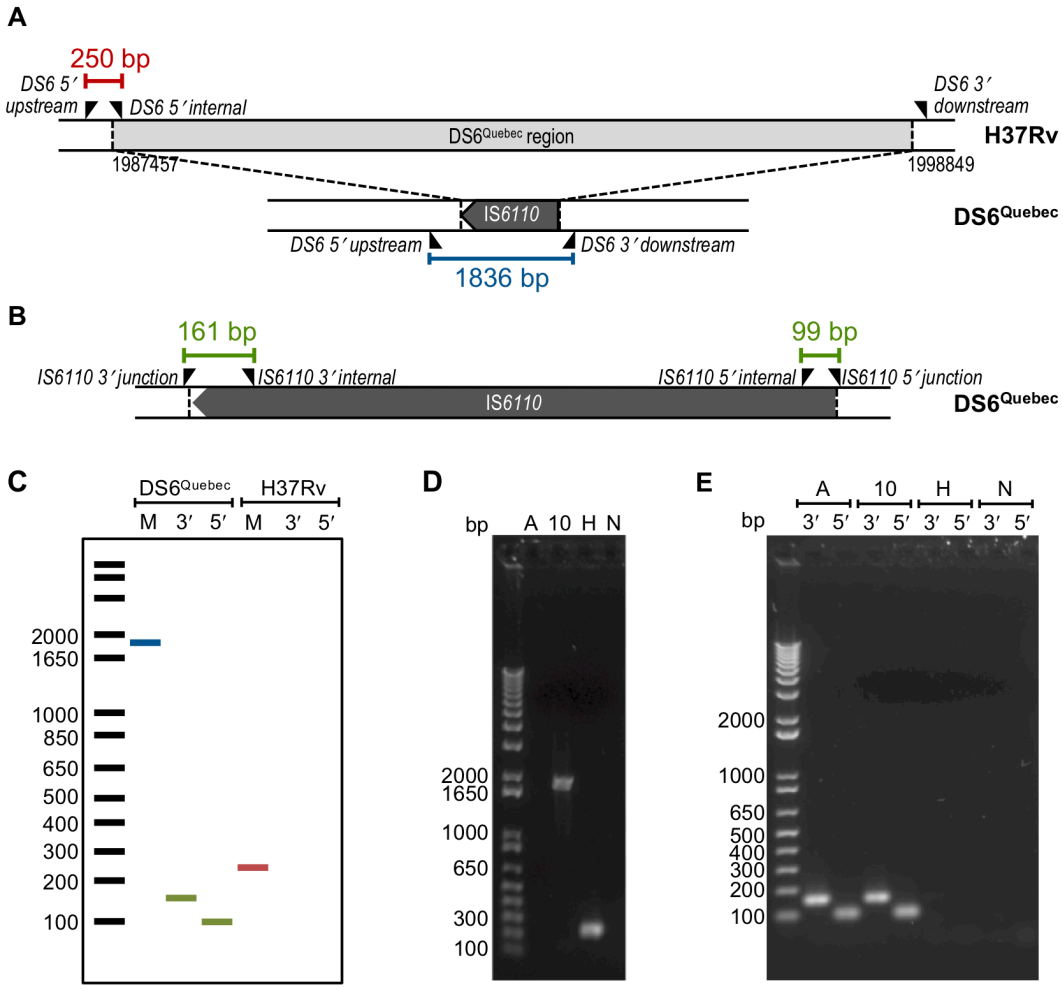


Figure 2.8. *DS6^{Quebec} deletion PCR assays.* (A) DS6^{Quebec} multiplex PCR. Expected PCR product indicating deletion of the DS6^{Quebec} region is shown in blue and in red if the region is not deleted. Primer binding sites are indicated by black triangles. Not to scale. (B) DS6^{Quebec}-IS*6110* insertion-specific verification of the DS6^{Quebec} deletion. Note that due to the direction of the IS*6110* insertion, the IS*6110* 3′ primers amplify the 5′ flank of the deletion and the IS*6110* 5′ primers amplify the 3′ flank. (C) Schematic showing theoretical PCR products for the DS6^{Quebec} multiplex (M) and DS6^{Quebec} IS*6110* insertion-specific PCRs (3′ and 5′) for strains harbouring the DS6^{Quebec} deletion and for H37Rv. (D) DS6^{Quebec} multiplex PCR amplification products. A, Rangipo A; 10, Rangipo 10; H, H37Rv; N, PCR negative control. (E) DS6^{Quebec}-IS*6110* insertion-specific PCR amplification products. The absence of a PCR product for Rangipo A in the multiplex PCR is likely due to DNA degradation preventing amplification of the large 1836 bp product.

Sanger sequencing confirmed the correct products had been amplified and the presence of the DS6^{Quebec} deletion in the Rangipo strain. Rangipo isolate A produced bands of the correct sizes for both insertion-specific PCRs but failed to amplify in the multiplex PCR (Figure 2.8D,E). This is likely due to DNA degradation preventing amplification of the large 1836 bp product as LabPLUS Rangipo samples amplified in the large multiplex PCR product as expected for isolates carrying the deletion (Figure 2.8D, Figure 2.9A).

New Zealand *M. tuberculosis* cluster isolates were screened with the multiplex PCR. Isolates from the Rangipo, Otara and NZ_094 clusters all gave a band of the expected size for the DS6^{Quebec} deletion and insertion-specific PCRs confirmed the deletion occurred at the exact same site as in the DS6^{Quebec} strain family (Figure 2.9). All Southern Cross and NZ_037 isolates and a single NZ_041 isolate failed to amplify in the multiplex PCR, despite repeated attempts. The remaining isolates from NZ_041 and all NZ_069 isolates gave the expected sized band for the absence of the deletion. The NZ_041 isolate that failed to amplify, one Southern Cross, and one NZ_037 isolate were also screened with the insertion-specific PCRs to ensure amplification failure was not the result of DNA degradation impairing amplification of long product. Neither the Southern Cross nor NZ_037 isolates amplified in either PCR reaction, confirming these clusters do not carry the DS6^{Quebec} deletion. The NZ_041 isolate did not amplify for the 5′ insertion-specific PCR, but gave a band of the correct size for the 3′ insertion-specific PCR. This genomic region contains multiple mobile elements and is known to be highly variable and associated with frequent IS*6110* insertion (Ho *et al.* 2000). The failure of these isolates to amplify in either a DS6^{Quebec} or H37Rv-like manner suggests an alternative deletion or IS*6110* insertion in this region.

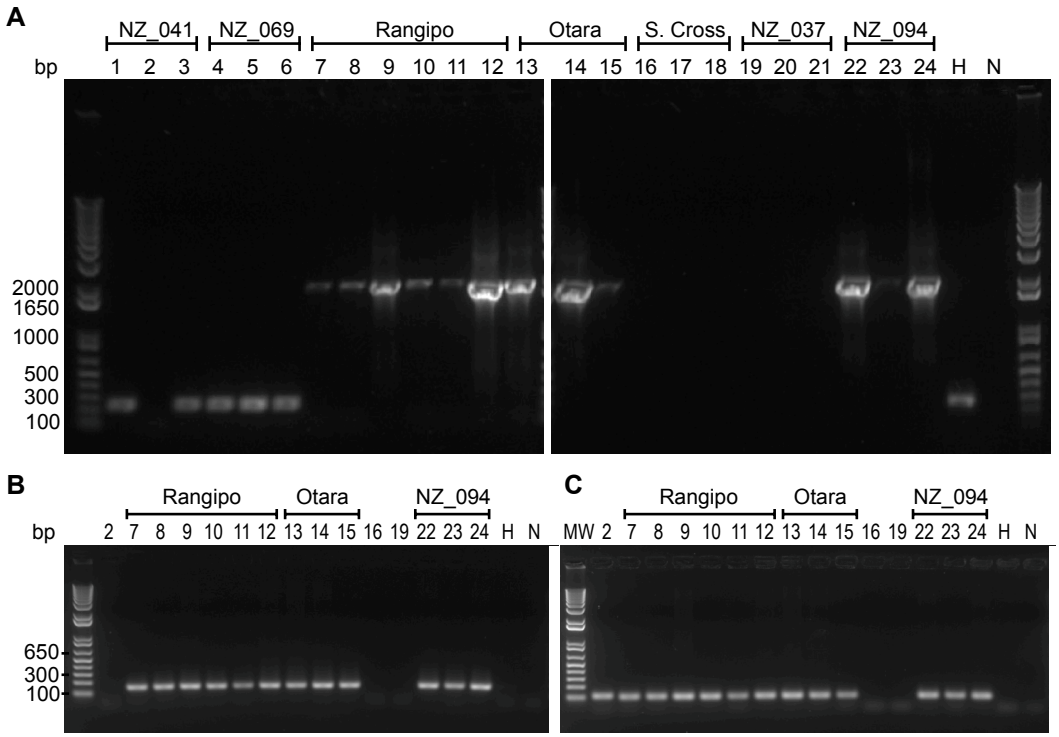


Figure 2.9. *DS6^{Quebec} deletion PCR assay typing of New Zealand* M. tuberculosis *cluster isolates.* MIRU-VNTR cluster ID is shown above the lane number. (A) DS6^{Quebec} deletion multiplex PCR (lanes 1-3, NZ_041; 4-6, NZ_069; 7-12, Rangipo; 13-15, Otara; 16-18, Southern Cross; 19-21, NZ_037; 22-24, H, H37Rv; N, PCR negative control). (B,C) DS6^{Quebec}-IS*6110* insertion-specific PCR for the 5′ (B) and 3′ (C) ends of the deletion. Lane numbers correspond to isolates assayed in the multiplex PCR.

2.2.5 Rangipo diagnostic

Based on the reclassification of Rangipo SNPs (Section 2.2.1) and WGS of additional New Zealand cluster isolates (Chapter Four), a Rangipo specific SNP-based multiplex PCR-RFLP diagnostic assay previously developed in this laboratory (Ruthe 2015) was re-evaluated and updated. Reclassification of Rangipo SNPs identified the original diagnostic marker 1380G>A in Rv1821/*secA2* (Rv2067836GA) as specific to the Rangipo strain. Analysis of Illumina WGS data confirmed the presence of this SNP in a further 18 Rangipo genomes and its absence in 220 non-Rangipo L4.4 genomes (Section 4.3.2.2). An updated search in the GMTV database did not identify this SNP in 2501 global MTBC genomes (accessed August 2018).

2.2.5.1 Rangipo PCR-RFLP assay

A revised RFLP-PCR diagnostic for the Rangipo specific Rv1821/*secA2* SNP was developed to enable affordable and rapid identification of this strain. A 455 bp product PCR spanning the SNP was amplified and Sanger sequencing confirmed the correct fragment had been amplified. Digestion of this PCR product with the restriction enzyme MboI gave the expected fragments of 386 and 69 bp for Rangipo, and 215, 171 and 69 bp for non-Rangipo, allowing Rangipo and non-Rangipo isolates to be easily distinguished (Figure 2.10A,B).

This assay was experimentally validated on 28 isolates sequenced on the ABI SOLiD and/or Illumina platforms and 39 additional *M. tuberculosis* isolates including New Zealand cluster isolates and isolates from various MTBC lineages and L4 sublineages (Figure 2.10C,D). One sample (Figure 2.10D, lane 12) classified as Rangipo by MIRU-VNTR typing produced the non-Rangipo banding pattern. This isolate was from a case that likely contracted tuberculosis overseas (Ruthe 2015), supporting the non-Rangipo diagnosis and highlighting the greater discriminatory power of SNP-based typing over conventional MIRU typing. All remaining isolates gave the expected fragments for Rangipo and non-Rangipo isolates.

A total of 39 Rangipo isolates have been typed by WGS or PCR-RFLP for the Rv1821/*secA2* SNP in this study and in previous work (Ruthe 2015) and all harbour the Rangipo specific variant. These isolates span a 26-year period (1991–2017) and include isolates from a range of geographic locations throughout New Zealand, and are expected to provide a good representative sample of this cluster. The phylogenetic analyses performed in Chapter Four show this SNP is a phylogenetically informative SNP specific to the Rangipo strain, further supporting its suitability for use as a diagnostic marker.

A paper describing this diagnostic assay has been published in Diagnostic Microbiology and Infectious Disease (Mulholland *et al.* 2017) (Appendix D).

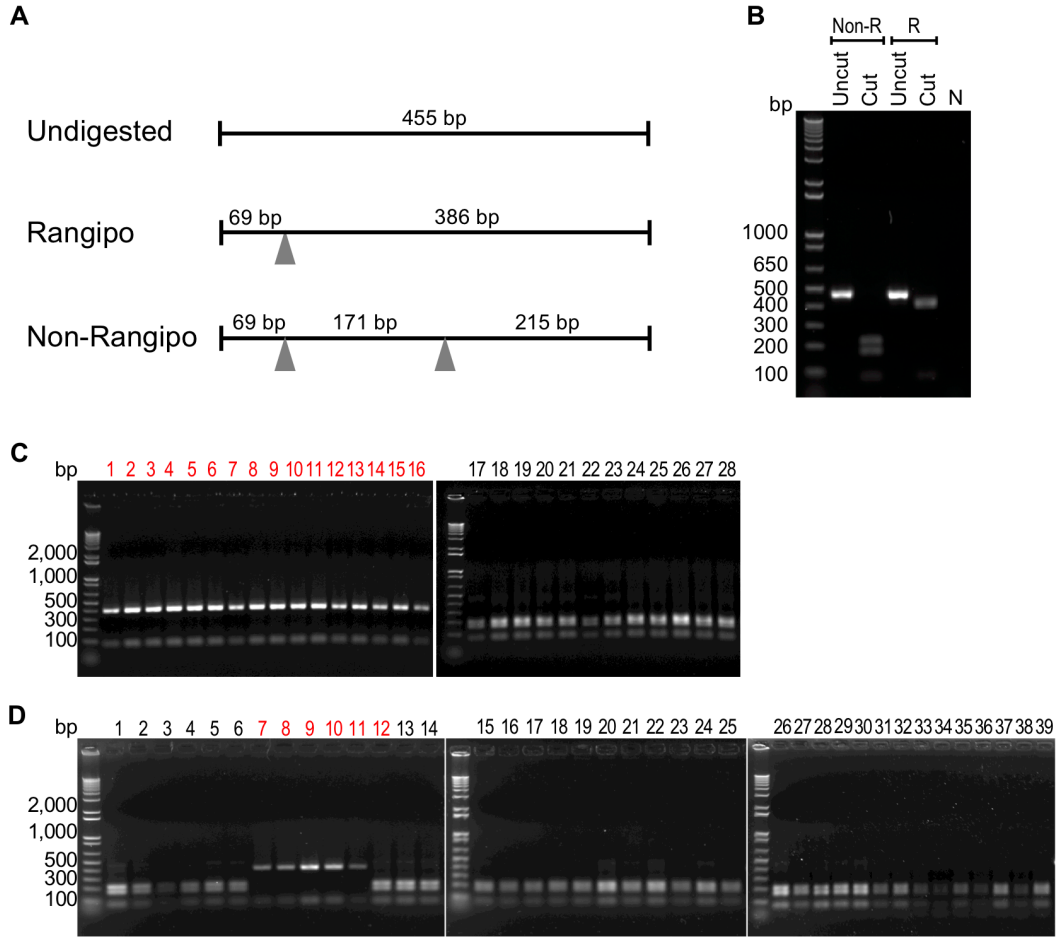


Figure 2.10. *Rangipo strain specific PCR-RFLP diagnostic assay.* (A) Schematic showing positions of the MboI recognition sequences (arrows) on the PCR product and the resulting fragment sizes for Rangipo and non-Rangipo samples. (B) PCR-RFLP assay showing uncut PCR products and digested fragments of non-Rangipo (Non-R) and Rangipo (R) PCR products. (C) Digested fragments from New Zealand clinical *M. tuberculosis* isolates sequenced on the Illumina and/or ABI SOLiD platform. Isolates with Rangipo MIRU-VNTR typing profiles are shown in red text. (Lanes 1-16, Rangipo; 17-21, Southern Cross; 22-28, Otara). (D) New Zealand clinical *M. tuberculosis* isolates and isolates from major MTBC lineages and other from lineage 4 sublineages. Isolates with Rangipo MIRU-VNTR typing profiles are shown in red text. (Lanes 1-3, NZ_041; 4-6, NZ_069; 7-12 Rangipo; 13-15, Otara; 16-18 Southern Cross; 19-21, NZ_037; 22-25, NZ_094; 26, L1/EAI; 27-29, L2/Beijing; 30, L3/CAS; 31, Cameroon (L4); 32, Ural (L4); 33, Iran (L4); 34, Uganda (L4); 35, PGG3 (L4); 36, X (L4); 37, LAM (L4); 38, Haarlem (L4); 39, H37Rv (L4)).

2.2.5.2 Evaluation of the Rangipo diagnostic in a clinical setting

The Rangipo diagnostic was evaluated using culture independent DNA from heat inactivated sputum samples from a small local Rangipo strain outbreak in 2017/2018. Four samples were provided by the Waikato Hospital (Hamilton, New Zealand), including two non-Rangipo control cases. The PCR-RFLP assay was performed as for extracted gDNA, with the exception that heat inactivated supernatant from sputum specimens was directly input into the PCR instead of gDNA (Section 2.1.8.2). Samples 1 and 2 were both diagnosed as Rangipo and samples 3 and 4 as non-Rangipo (Figure 2.11). Consistent with these results, MIRU-VNTR typing at LabPLUS confirmed Sample 1 and 2 as Rangipo (MIRU-VNTR 24-locus code, 233325153324–341444223362) and 3 and 4 as non-Rangipo (223325173533–445643423382). PCR-RFLP results were available within 24 hrs, however this could be further reduced for samples received early in the day.

These results have been published in the New Zealand Medical Journal (Mulholland *et al.* 2018) (Appendix D).

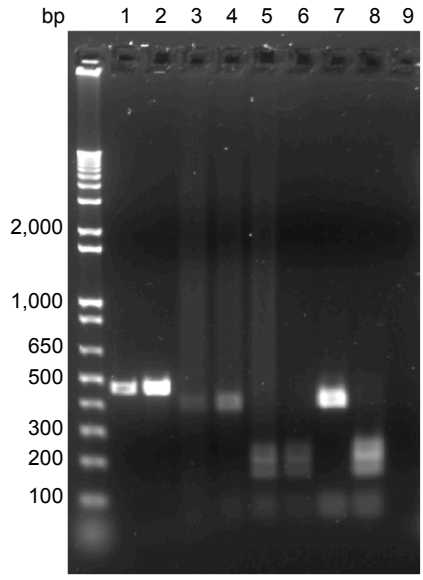


Figure 2.11. *Rangipo PCR-RFLP diagnostic screening of clinical sputum samples*. Lane 1, uncut Rangipo control; 2, uncut non-Rangipo control (H37Rv); 3 and 4, clinical sputum samples diagnosed as Rangipo by this assay (samples 1 and 2); 5 and 6, clinical sputum samples diagnosed as non-Rangipo (samples 3 and 4); 7, Rangipo control isolate after digestion; 8, Non-Rangipo control after digestion; 9, PCR negative control.

2.3 Discussion

The Rangipo strain has been the source of several large outbreaks over the past 30 years and anecdotal evidence suggests that it is highly transmissible (Colangeli *et al.* 2014; De Zoysa *et al.* 2001; McElnay *et al.* 2004). Analysis of contact tracing results of previous Rangipo strain outbreaks revealed a higher proportion of traced contacts with active disease and a lower proportion with latent infection compared to average contact tracing estimates (Section 2.2.1). The portion of infected cases with active disease also exceeded the 5–10% of infections typically expected to progress to active disease. The skew towards active disease and the relatively low proportion of contacts with latent infection suggests that the Rangipo strain might not necessarily be more infectious, but that a greater proportion of infected contacts progress to active disease.

Bacterial, host and environmental factors can all influence the outcome of *M. tuberculosis* infection and as active pulmonary tuberculosis is required for transmission to occur, progression to active disease is relevant when considering tuberculosis transmission. De Jong *et al.* (2008) have reported strain specific differences in progression to active disease, demonstrating a role of bacterial genetic background in disease progression and the outcome of infection. While this preliminary analysis hints that Rangipo strain infections may be more likely to progress to active disease supporting the notion that this strain is highly virulent, it is limited by the absence of a control group/strain and heterogeneity among contact-tracing methodologies, including those used to estimate average tuberculosis incidence among contacts (Fox *et al.* 2013). Latent infection could also potentially be underestimated due to limitations of methods for identifying latent infections (Al-Orainey 2009).

Host factors also undoubtedly play an important role in the success of this strain. For example, prison incarceration is a common factor noted among Rangipo strain cases (De Zoysa *et al.* 2001) and is a known tuberculosis risk factor (Baussano *et al.* 2010). The Rangipo strain is predominantly found in Māori and inequalities facing this population such as household crowding, socioeconomic deprivation and cigarette smoking are also likely to influence transmission and outcome of infection. Further work should seek to examine transmission and rates of progression to active disease compared to other circulating strains to better

understand its epidemiology, along with laboratory testing to investigate its presumed high virulence.

2.3.1 Rangipo SNPs

To investigate potential bacterial genetic factors that may influence the pathogenicity of the Rangipo strain and to identify targets for further biochemical analysis, I sought to identify nsSNPs specific to this strain. Colangeli *et al.* (2014) identified 247 SNPs shared by ten Rangipo strain isolates collected from 1991–2011 sequenced on the ABI SOLiD platform. All of these SNPs have previously been crosschecked against 38 MTBC genomes in publicly available databases to identify SNPs specific to the Rangipo strain (Ruthe 2015). Since that analysis was performed, the number of strains available for comparison has increased dramatically. In this work, reclassification against >1000 *M. tuberculosis* genomes decreased the total number of RS-SNPs from 99 to 22, and the number of RS-nsSNPs from 61 to 14. Having a large number of strains available for comparison is particularly important for the identification of diagnostic marker SNPs as it is essential SNPs are specific to the strain of interest to prevent misclassification. Of three putative RS-SNPs that have previously been proposed as diagnostic makers for the Rangipo strain (Ruthe 2015) only one, 1380G>A in Rv1821/*secA*, remained classified as a RS-SNP in this work and was validated as a suitable diagnostic marker for Rangipo.

SNPs unique to the Rangipo strain, particularly those with *in silico* predicted functional effects, were predominantly located in genes involved in intermediary metabolism and respiration and lipid metabolism. A recent study by Gautam *et al.* (2017) sequenced nine Rangipo strain isolates on the Illumina platform and found nsSNPs common to Rangipo isolates were also predominantly located in genes with metabolic functions (Gautam *et al.* 2017). In the host granuloma, the bacterium experiences various stresses such as nutrient and oxygen depletion, requiring metabolic changes to circumvent nutritional shortages and to adapt to hypoxic conditions. The ability of *M. tuberculosis* to adapt to such hostile environmental changes during the course of infection underlies much of its success as a pathogen (Cook *et al.* 2009). Genes involved in basic cellular metabolism are generally not considered virulence factors, however other metabolic processes such as lipid

metabolism are important virulence determinants as during the course of infection the bacterium shifts from utilising carbon sources such as glucose, to fatty acids and host lipids (Schnappinger *et al.* 2003). Various lipids and glycolipids are also incorporated into the cell envelope where they are thought to be important for bacterial pathogenicity (Neyrolles and Guilhot 2011).

Determining the functional consequences of SNPs is important in understanding how genetic variation in the MTBC impacts bacterial fitness and virulence. *In vitro* biochemical and *in vivo* characterisation of variants provide the ultimate experimental evidence of variant effects. However, these methods are laborious and the functions of many MTBC proteins remain unknown making such investigations challenging. Computational approaches can identify SNPs with putative molecular effects enabling large numbers SNPs to be examined *in silico*. Such approaches do not connect effect at the molecular level to "impact on organism", but rather can be used to inform the selection of candidate variants for experimental analysis (Hecht *et al.* 2015).

Eight SNPs unique to the Rangipo strain were predicted *in silico* to affect molecular function including two in the virulence factors *narG* and *pks6*. *narG* encodes the catalytic α-subunit NarG of the membrane bound multi-subunit nitrate reduction complex NarGHJI. This complex catalyses the respiratory reduction of nitrate (NO_3^-) to nitrite (NO_2^-), enabling nitrate to be used in place of oxygen as the final electron acceptor in the electron transport chain. *M. tuberculosis* is one of the strongest nitrate reducers amongst mycobacteria and the nitrate reductase system is implicated in *M. tuberculosis* virulence (reviewed in Khan and Sarkar 2012). *pks6* encodes a predicted polyketide synthase Pks6. Polyketide synthases are large multi-domain proteins that catalyse complex reactions for the synthesis of complex lipids and secondary metabolites, and Pks6 is suggested to be involved in the production of novel polar lipids (Waddell *et al.* 2005). *pks6* gene inactivation attenuates *M. tuberculosis* in mouse (Camacho *et al.* 1999) and human macrophage models (Rosas-Magallanes *et al.* 2007), implicating *pks6* in virulence, however its exact role in pathogenesis remains unclear.

The Y363D SNP in *coaE* had the strongest prediction for a functional effect of all 14 RS-SNPs. Further investigation showed this SNP is located in the GrpB domain and is part of highly conserved Y-K-Y-K motif. Currently the only assigned

function of this domain is for the correct folding of the full enzyme and the last 50 residues at the C-terminus, including the Y363D SNP and all four conserved Y-K-Y-K residues, are dispensable for proper folding and catalytic activity (Walia *et al.* 2009). Therefore, if the sole role of the GrpB is for the folding of CoaE, the G72S SNP is unlikely to have an effect on enzyme activity. However, if this domain has an additional as yet unknown function, considering the chemical nature of the amino acid change and the highly conserved nature and structural confirmation of the Y-K-Y-K motif, this SNP could have a dramatic effect on activity.

Rv2893 encodes a putative oxidoreductase and is predicted to utilise the unusual deazaflavin cofactor F_{420} (Selengut and Haft 2010). F_{420} is an unusual low-redox potential flavin derivative cofactor synthesised by methanogenic archaea and several aerobic bacterial phyla including the Actinobacteria (Daniels *et al.* 1985; Ney *et al.* 2017). F_{420} is thought to enhance the metabolic flexibility of mycobacteria and in *M. tuberculosis* plays an important role in survival under hypoxic conditions, and protection from oxidative, nitrosative and antibiotic stress (Gurumurthy *et al.* 2013; Hasan *et al.* 2010; Purwantini and Mukhopadhyay 2009). Rv2893 belongs to the LLHT family of F_{420}-dependent enzymes. There are 14 predicted F_{420}-dependent LLHTs in *M. tuberculosis* (Selengut and Haft 2010), the most well characterised of which is F_{420}-reducing glucose-6-phosphate dehydrogenase (FGD) (Bashiri *et al.* 2008; Oyugi *et al.* 2016; 2018). The functions of the majority of these, including Rv2893, remains unknown.

One of the main aims of the SNP analysis presented in this chapter was to identify candidate SNPs for biochemical and structural investigation. While the *pks6* and *narG* SNPs are of interest due to these being known virulence associated genes, for practical reasons these variants were not selected for further study in this work. Pks6 is a large multi-domain enzyme and NarG is part of a multi-subunit complex anchored to the membrane. These features make proteins very difficult to work with for *in vitro* biochemical studies and are expected to make them difficult to crystallise. Soluble *M. tuberculosis* CoaE has been recombinantly expressed in *E. coli* and enzyme activity assayed by spectrophotometry (Walia *et al.* 2009) making it a suitable candidate from a practical perspective. However, further investigation revealed the mutation is located in a region of the protein dispensable for CoaE catalytic activity *in vitro* (Walia *et al.* 2009).

Based on the role of cofactor F_{420} in *M. tuberculosis*, further *in silico* analysis and practical considerations, the Rv2893 G72S SNP was selected as a suitable candidate for further investigation in Chapter Three of this thesis. Rv2893 is a cytoplasmic protein and is not part of a multimeric complex, making it a more suitable candidate for protein crystallography than other genes of interest harbouring Rangipo SNPs. The intrinsic fluorescence of F_{420} and the differences in its absorbance spectrum between the reduced and oxidised states also provides a simple way to monitor enzyme activity for future assay development. Furthermore, little is known about the functions of most members of the F_{420}-dependent LLHTs in *M. tuberculosis*, making this protein of general interest for further study.

Since this SNP analysis was performed, additional Rangipo strains have been sequenced using Illumina technology (as part of this research and by Gautam *et al.* (2017)) leading to the identification of additional Rangipo specific SNPs (Section 4.3.2.2). While these identify further candidates of interest, the selection of candidate SNPs was based on the analysis presented in this chapter as Illumina data was not available at the time.

In addition to genomic variation in the form of SNPs, differential carriage of virulence-associated genes may also impact bacterial fitness and pathogenicity. This work shows that the Rangipo strain harbours the $DS6^{Quebec}$ deletion that results in the loss of several virulence-associated genes (discussed in Section 2.3.3). Gautam *et al.* (2017) have also recently identified several genes present in the Rangipo strain that are absent in H37Rv, including three virulence-associated genes. Although these large genomic variants are not restricted to the Rangipo strain, they potentially also affect strain virulence and further work is needed to determine the consequences of this genomic diversity in *M. tuberculosis*.

2.3.2 The S lineage

The Rangipo strain has previously been assigned to the L4/Euro-American lineage of the MTBC but the sublineage remained unclassified (Ruthe 2015). Here, Rangipo, and additionally the Otara and NZ_094 strains, are classified as belonging to the L4.4.1.1/S sublineage. The ‘S lineage’ was originally classified based on its spoligotyping signature, which is characterised by the absence of the direct repeat spacers 9–10 and 33–36 (Warren *et al.* 2002a). Sequence based phylogenetic

analyses show that S isolates group into the same SNP defined phylogenetic clade congruent with MIRU-VNTR and spoligotyping data (Coll *et al.* 2014; Homolka *et al.* 2012). A high-resolution WGS SNP phylogeny constructed from 1601 global MTBC genomes shows that the S lineage is equivalent to the L4.4.1.1 sublineage, which belongs to the larger L4.4 sublineage (Coll *et al.* 2014). Stucki *et al.* (2016b) have SNP-typed >3000 global L4 isolates from 100 countries and found L4.4 to be the fourth most frequent of the ten L4 sublineages globally accounting for 10% of isolates, and has an intermediate global distribution relative to other sublineages. In New Zealand, L4.4 is the most common L4 sublineage and accounts for 43% of L4 isolates from New Zealand-born cases (Stucki *et al.* 2016b).

Isolates belonging to the S lineage were first identified around the same time independently in Canada ('DS6^{Quebec}') (Cheng *et al.* 2000), Italy (Sola *et al.* 2001) and South Africa ('F28') (Warren *et al.* 2002b). The DS6^{Quebec} family is endemic in Canada and is the most prevalent lineage in French Canadians in Quebec (48% of isolates) and in Aboriginal populations in Ontario, Saskatchewan and Alberta (38–62% of isolates) (Pepperell *et al.* 2011). This lineage assumed to have been introduced to Quebec by a French migrant in the 17th century and was dispersed to indigenous populations in the 18–19th centuries via the Canadian fur trade (Pepperell *et al.* 2011). In New Zealand, the S lineage Rangipo and Otara strains are most prevalent in indigenous Māori and Pacific People and despite extensive social admixture of New Zealand Europeans and Māori, the Rangipo strain remains strongly associated with Māori. Different MTBC lineages are hypothesised to be adapted to specific human host populations (Gagneux 2012), raising the intriguing prospect that the S lineage may be locally adapted to these populations. Alternatively, the high prevalence of the Rangipo strain in Māori is also consistent with a founder effect or population bottle neck.

A small number of SNPs were found to be shared between the New Zealand Rangipo strain and a subset of Canadian DS6^{Quebec} family isolates, suggesting these strains might share a close phylogenetic relationship. This was further explored in a detailed phylogenetic analysis including Otara and other global L4.4 lineage isolates in Chapter Four of this thesis.

2.3.3 DS6^{Quebec} deletion

Nguyen *et al.* (2004) have previously postulated that the 11.4 kb DS6^{Quebec} deletion that characterises the Canadian DS6^{Quebec} family may be a marker of a more globally distributed lineage. We have confirmed that the Rangipo strain and other New Zealand S lineage isolates harbour the DS6^{Quebec} deletion, corroborating this hypothesis. South African F28 strain isolates also have an IS*6110* insertion in this region (Chihota 2011), which together with our results suggest the DS6^{Quebec} deletion might be more widely characteristic of the S lineage. This was further investigated in Chapter Four (Section 4.3.3.3).

The similar but evolutionarily independent RD152 deletion is characteristic of the highly successful Beijing/W strain (Tsolaki *et al.* 2005). Both the DS6^{Quebec} and RD152 deletions remove or interrupt the seven genes from Rv1755c/*plcD* (which is truncated in H37Rv) to Rv1765c, and RD152 additionally interrupts Rv1754c (Figure 2.7). Some of these deleted genes have important roles in pathogenicity. Rv1759c/*wag22* for example, is a PE-PGRS family fibronectin-binding protein that elicits an antibody response in human tuberculosis patients (Espitia *et al.* 1999) and is suggested to play a role in the immune response maintaining latent infection (Campuzano *et al.* 2007). Rv1760 is adjacent to Rv1759c/*wag22* and encodes a possible triacylglycerol synthase. Triacylglycerol is an important energy source for sustaining *M. tuberculosis* during latent infection and this gene is highly upregulated in a hypoxic macrophage model mimicking the microenvironment within the granuloma (Daniel *et al.* 2011). It is interesting to note that this deletion removes genes that may play a role in latency. Increased progression from latent infection to active tuberculosis has been reported for contacts exposed to Beijing lineage strains compared to other *M. tuberculosis* strains (De Jong *et al.* 2008) and analysis of Rangipo outbreak contact tracing data performed here suggests relatively high rates of progression to active disease in Rangipo strain infections.

While the genes deleted by the DS6$^{Quebec-}$ deletion are evidently not essential for virulence, the impact of this deletion on virulence and disease outcome remains unclear. Furthermore, additional deletions of variable lengths in this region have also been reported (Alonso *et al.* 2013; Kato-Maeda *et al.* 2001; Tsolaki *et al.* 2005). H37Rv also has two deletions in this region relative to *M. bovis* BCG and other MTBC strains (RvD2 and RvD3, both spanned by the DS6^{Quebec} and RD152

deletions) (Brosch *et al.* 1999; Gordon *et al.* 1999). The highly variable nature of this region is associated with frequent IS*6110* insertion, and homologous recombination between IS*6110* elements provides a likely mechanism for frequent deletion events (Ho *et al.* 2000).

2.3.4 Rangipo diagnostic

WGS based *M. tuberculosis* strain-typing will likely become routine in the future but is not yet performed in New Zealand. WGS-directed SNP based assays such as the Rangipo PCR-RFLP diagnostic presented here, offers a rapid and affordable alternate approach for epidemiological typing in the interim. This diagnostic is able to distinguish between Rangipo and non-Rangipo isolates with higher discriminatory power than MIRU typing providing results within one day at little cost. Clinical isolates were able to be typed using culture-independent *M. tuberculosis* DNA directly from clinical sputum samples substantially faster than the standard turnaround time of 3-4 weeks for MIRU typing. Using a PCR-RFLP assay for strain typing offers the advantage of being cheap and easy to perform with basic lab equipment and requires little optimisation. As this is a SNP-based approach it could also be automated, for example by incorporation into the GeneXpert system. If the Rangipo strain is indeed more virulent, close supervision ensuring treatment adherence and broadening contact tracing networks may be beneficial to ensure secondary cases are completely and quickly detected to prevent potential later reactivation and further spread. The ability to rapidly identify Rangipo strain cases upon patient presentation will enable prompt intervention to help prevent further transmission of this strain.

2.3.5 Conclusions

The New Zealand Rangipo and Otara strains belong to the L4.4.1.1/S sublineage of the L4/Euro-American lineage of the MTBC. New Zealand S lineage strains harbour the DS6^{Quebec} deletion and a putative relationship between the Rangipo strain and Canadian DS6^{Quebec} family isolates was identified. Twenty-two SNPs were classified as specific to the Rangipo strain, including eight nsSNPs with predicted functional effects, providing interesting candidates for further investigation. In light of these results, a previously developed SNP-based Rangipo

diagnostic assay was re-optimised. This diagnostic enables rapid and affordable typing of Rangipo strain isolates directly from clinical sputum samples, offering higher resolution and faster results than conventional MIRU-VNTR typing.

Chapter Three

Structure of the F_{420}-dependent oxidoreductase Rv2893

3.1 Introduction

The New Zealand *Mycobacterium tuberculosis* Rangipo cluster harbours a strain specific single nucleotide polymorphism (SNP) encoding a G72S mutation in Rv2893 predicted *in silico* to affect protein function (Section 2.2.2). The Rv2893 gene encodes a putative oxidoreductase that belongs to the luciferase-like hydride transferase family of flavin/deazaflavin proteins. Rv2893 is predicted to utilise the 5′-deazaflavin cofactor F_{420}, a functionally versatile redox co-factor that is important for energy metabolism in *M. tuberculosis* and contributes to its ability to persist in challenging environments (Gurumurthy *et al.* 2013; Purwantini and Mukhopadhyay 2009). Bioinformatics analyses indicate the presence of at least 28 F_{420}-dependent proteins in *M. tuberculosis* (Selengut and Haft 2010). F_{420}-dependent proteins have been implicated in a range of processes in mycobacteria, including synthesis of cell wall lipids (Purwantini *et al.* 2016; Purwantini and Mukhopadhyay 2013), fatty acid modification, and heme degradation (Ahmed *et al.* 2015), however, the biological function of most of these oxidoreductases remains unknown.

There is no structure available for Rv2893 and its function is unknown. To investigate the consequences of the G72S mutation on protein structure and help elucidate the biological role of Rv2893, structures for both the wild-type protein ($Rv2893^{H37Rv}$) and the Rangipo G72S variant ($Rv2893^{G72S}$) in both the apo and F_{420}-bound forms were solved using protein X-ray crystallography.

3.1.1 Cofactor F_{420}

F_{420} is a low redox potential 5′-deazaflavin cofactor named for the strong fluorescence of the oxidised cofactor upon excitation at 420 nm (emission at 470 nm). It consists of an isoalloxazine tricyclic ring system with a side chain comprised of ribitol and phospholactate moieties and oligoglutamate tail of varying length (Figure 3.1). In archaea, two glutamate residues are present on the tail (Eirich *et al.* 1978), whereas up to nine are observed in mycobacteria (Bashiri *et al.* 2008). While its structure is similar to those of the universal flavin cofactors FAD and FMN, the N5 of the isoalloxazine ring is substituted for a carbon in F_{420} resulting in different electrochemical properties. Unlike the flavins, which can act as one or two electron carriers, F_{420} is an obligate two-electron hydride carrier and has a lower redox potential (−340 mV) (Jacobson and Walsh 1984) than FAD and FMN (−220 mV and 190 mV, respectively), and the nicotinamide cofactor NAD(P) (–320 mV) (Thauer *et al.* 1977). These properties make F_{420} useful for reducing a wide range of reactants, enhancing the metabolic flexibility of microorganisms that produce it.

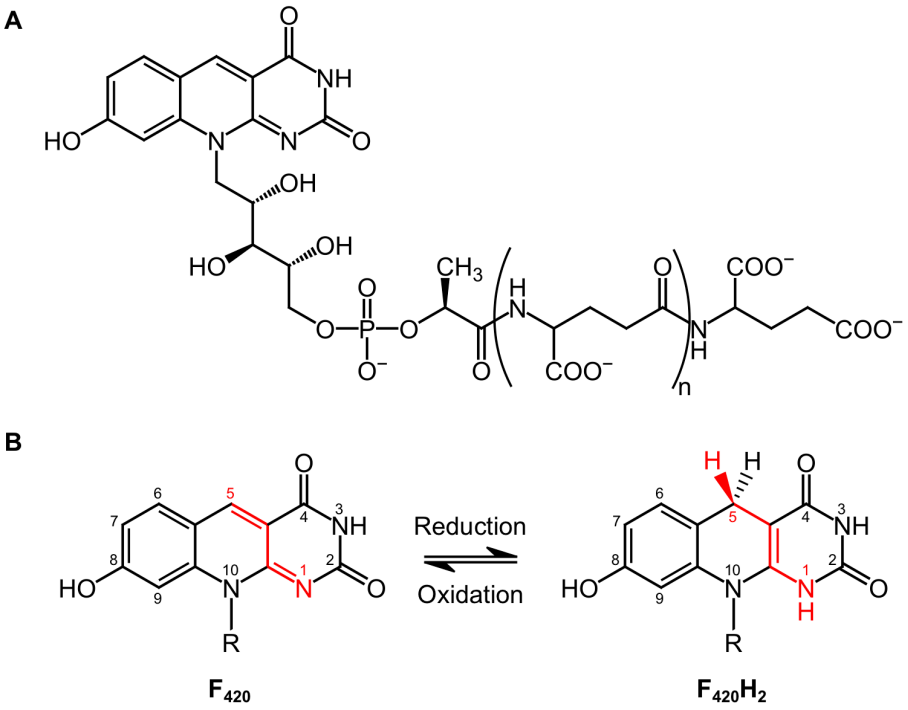


Figure 3.1. *Molecular structure of cofactor F_{420}.* (A) Schematic representation of the chemical structure of F_{420}. The length of the polyglutamate tail varies among microorganisms; in *M. tuberculosis*, the major species contain 5–6 residues (Bair *et al.* 2001). (B) Oxidised and reduced forms of cofactor F_{420} (F_{420} and $F_{420}H_2$, respectively).

F_{420} was formally identified and first isolated in the 1970s from methanogenic archaea (Cheeseman *et al.* 1972) and its chemical structure proposed six years later (Eirich *et al.* 1978). It was formally identified in bacteria in 1980 (Eker *et al.* 1980) and discovered in *M. tuberculosis* in 1985 (Daniels *et al.* 1985). F_{420} is considered to be taxonomically restricted to the Euryarchaeota and Actinobacteria, although recently it has been shown to be more widely distributed in aerobic soil bacteria than previously thought (Ney *et al.* 2017).

F_{420} is synthesised universally amongst Mycobacteria (Greening *et al.* 2016). The ability to synthesize and reduce F_{420} is non-essential in mycobacteria under ideal conditions, however, loss of F_{420} biosynthesis and reduction results in impaired survival under oxidative and nitrosative stress and following exposure to antimicrobials, and impairs its ability to recover following hypoxia-induced dormancy (Gurumurthy *et al.* 2013; Hasan *et al.* 2010; Jirapanjawat *et al.* 2016; Purwantini and Mukhopadhyay 2009). Mechanisms that enable the bacterium to withstand these stresses are important for pathogenesis and persistence of *M. tuberculosis* (Ehrt and Schnappinger 2009), pointing to a likely role for F_{420} in the pathogenesis of *M. tuberculosis.* Furthermore, activation of the anti-tuberculosis drug PA-824 occurs via a F_{420}-dependent mechanism (Manjunatha *et al.* 2006), a finding that has sparked additional interest in cofactor F_{420} and the enzymes which utilise it.

3.1.2 F_{420}-dependent enzymes

The majority of F_{420}-dependent enzymes fall into two major super families; the flavin/deazaflavin oxidoreductases (FDORs) and the luciferase-like hydride transferases (LLHTs). These have a diverse range of catalytic activities and vary in their co-factor preference for the different flavins (FMN, FAD and F_{420}).

The FDORs have a split β-barrel protein fold and can be divided into two major families (FDOR-A and FDOR-B) which share ~30% sequence identify (Ahmed *et al.* 2015; Taylor *et al.* 2010). FDOR-A family proteins are exclusively F_{420}-binding and are taxonomically restricted to the *Actinobacteria* and *Chloroflexi*. Conversely, FDOR-B proteins variously utilise different cofactors (F_{420}, FAD, FMN, and heme)

and are more widely distributed, including in non-F_{420} producing organisms. (Reviewed in Greening *et al.* 2016).

Members of the LLHT protein super family have a TIM barrel fold. The LLHTs were previously known as luciferase-like monooxygenases (LLMs), however these are better defined as LLHTs as the reaction mechanisms of these enzymes is O_2 independent (Greening *et al.* 2016). The co-factor preference of these enzymes varies and they are found in both F_{420} and non-F_{420} producing organisms (Selengut and Haft 2010). A conserved glycine in the F_{420}-binding site allows binding of the phosphate group without steric hindrance and distinguishes F_{420}-dependent LLHTs from non-F_{420} dependent enzymes (Aufhammer *et al.* 2004). Fourteen of the seventeen LLHT family genes in *M. tuberculosis* have been predicted as being F_{420}-dependent *in silico* (F_{420}-LLHTs), including Rv2893 (Selengut and Haft 2010). F_{420}-LLHTs are implicated in a range of processes in actinobacteria and most are thought to serve as reductases (reviewed in Greening *et al.* 2016). The most well characterised mycobacterial F_{420}-LLHT is F_{420}-dependent glucose-6-phosphate dehydrogenase (FGD, previously FGD1) from *M. tuberculosis* (mtbFGD) (Bashiri *et al.* 2007; Bashiri *et al.* 2008; Oyugi *et al.* 2016; 2018). FGD uses F_{420} as the hydride acceptor to catalyse the conversion of glucose-6-phosphate (G6P) to 6-phosphogluconolactone, producing reduced F_{420} ($F_{420}H_2$) that is then used to reduce a diverse range of compounds by F_{420}-dependent reductases. FGD is thought to be the main source of $F_{420}H_2$ in mycobacteria (Jirapanjawat *et al.* 2016). FGD and other characterised F_{420}-LLHTs all have *Si*-face stereospecificity with respect to the C5 atom of the 5-deazaflavin group (Aufhammer *et al.* 2004; Aufhammer *et al.* 2005; Bashiri *et al.* 2008; Shima *et al.* 2000). Other F_{420}-LLHTs in *M. tuberculosis* have been implicated in the cell wall lipid synthesis (Purwantini *et al.* 2016; Purwantini and Mukhopadhyay 2013), although the functional roles of most of these proteins in *M. tuberculosis* is unknown.

3.1.2.1 Rv2893

Rv2893 is an LLHT family protein that belongs to the likely F_{420}-dependent F420_MSMEG_2516 family (TIGR03621), which includes three members from *Mycobacterium smegmatis* and only Rv2893 in *M. tuberculosis* (Selengut and Haft 2010). Rv2893 belongs to the intermediary metabolism and respiration functional

category of genes (Tuberculist database assignment (Lew *et al.* 2011)) and is non-essential in *M. tuberculosis* for bacterial growth and for infection in a mouse model (Sassetti *et al.* 2003; Sassetti and Rubin 2003). The only phenotypic profiling data found for this protein shows that Rv2893 is repressed 1.5-fold during acid shock *in vitro* (Fisher *et al.* 2002).

Rv2893 is a 325 amino acid protein with a molecular weight of 34.6 kDa. It shares 65% identify with the *M. smegmatis* homologue MSMEG_2516 and 100% sequence identify with *M. bovis* Mb2917. The Rv2893 gene is located at position 3202420–3203397 in the *M. tuberculosis* genome and is not part of an operon (Figure 3.2).

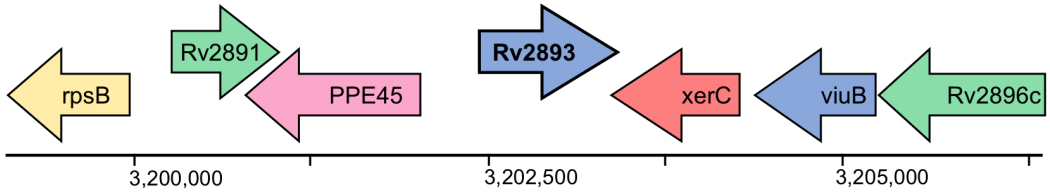


Figure 3.2. *Genomic context of Rv2893 in the* M. tuberculosis *H37Rv genome. Genes* are coloured by functional category: information pathways (yellow), conserved hypotheticals (green), PE/PPE (pink), intermediary metabolism and respiration (blue), insertion sequences and phages (red). Rv2893 belongs to the intermediary metabolism and respiration functional category and is not part of an operon.

3.1.3 Objectives

SNP analyses in Chapter Two identified 22 SNPs specific to the New Zealand *M. tuberculosis* Rangipo strain. Eight of these were non-synonymous (nsSNPs) predicted *in silico* to effect on protein function, including a G72S mutation in the predicted F_{420}-dependent LLHT Rv2893. Due to the role of F_{420} in *M. tuberculosis* and the relatively limited knowledge on F_{420}-LLHTs, as well as for practical reasons (Section 2.3.1), the Rv2893 G72S mutation was chosen for further investigation. There is no structure available for this protein and its function is unknown. To investigate the impact of this SNP on protein structure and provide clues to the function of Rv2893, I sought to solve the structure of Rv2893 for both apo- and F_{420}-bound $Rv2893^{H37Rv}$ and $Rv2893^{G72S}$ using protein X-ray crystallography.

3.2 Methods

3.2.1 Molecular cloning of Rv2893

Cloning primers were designed using Geneious R8 (Biomatters, N.Z.) for the ligation of Rv2893 variants into pYU28b (Bashiri *et al.* 2010) with an N-terminal His-tag (Appendix B.2). *M. tuberculosis* genomic DNA (gDNA) was kindly provided by LabPLUS (Auckland, New Zealand). Rv2893 was amplified from Rangipo strain gDNA ($Rv2893^{G72S}$) using KAPA HiFi^{TM} HotStart DNA polymerase (Kapa Biosystems, U.S.A.), and from H37Rv gDNA ($Rv2893^{H37Rv}$) using HOT FIREPol® Blend Master Mix (Solis BioDyne, Estonia). PCR products were either gel extracted and purified using the QIAquick Gel Extraction Kit (Qiagen, Netherlands), or, purified from solution using the QIAquick PCR Product Purification Kit (Qiagen, Netherlands). Purified PCR products and plasmid DNA were double digested with NdeI and BamHI using the NEB CutSmart® buffer system (New England Biolabs, U.S.A.). Digestion products were purified from solution using the QIAquick PCR Product Purification Kit (Qiagen, Netherlands) and ligated with T4 DNA ligase (Invitrogen, U.S.A.). Ligation reactions were transformed into electrocompetent *E. coli* TOP10 cells by electroporation using standard laboratory protocols. For electroporation, single aliquots of freshly thawed cells were mixed with 10 µl ligation reaction and 40 µl 10% ice cold glycerol. Electroporation was performed in 2 mm cuvettes at 2.5 kV, 25 µF capacitance and 200 Ω using a Bio-Rad Gene Pulser. Cells were immediately recovered in 1 ml of SOC media and incubated at 37 °C with shaking for 60 min before plating onto low salt LB agar plates with hygromycin B (50 $\mu g.ml^{-1}$) for selection. Positive *E. coli* transformants were screened for Rv2893 insertion into pYU28b using PCR. Transformants testing positive for the insert were grown in 5 ml low salt LB broth with hygromycin B (50 $\mu g.ml^{-1}$) overnight at 37 °C with shaking. Plasmid DNA was extracted from broth cultures using the QIAprep Spin Miniprep Kit (Qiagen, Netherlands) and sent to the Massey Genome Service (Palmerston North, N.Z.) for sequencing on an ABI3730 DNA Analyzer (Applied Biosystems Inc., U.S.A.) using T7 primers.

3.2.2 Protein expression in *M. smegmatis*

Sequence verified plasmids were transformed into *M. smegmatis* mc^2 4517 (Wang *et al.* 2010) by electroporation using established protocols (Cirillo *et al.* 1993). Single aliquots of freshly thawed cells were mixed with 1 µl plasmid DNA and 260 µl of 10% ice cold glycerol. Electroporation was performed in 2 mm cuvettes at 2.5 kV, 25 µF capacitance and 1000 Ω using a Bio-Rad Gene Pulser. Cells were immediately recovered in 1 ml of Middlebrook 7H9 broth supplemented with ADC and 0.05% (v/v) Tween 80, and incubated at 37 °C with shaking for 3 hrs before plating onto low salt LB agar plates containing hygromycin B (50 $\mu g.ml^{-1}$) and kanamycin (50 $\mu g.ml^{-1}$) for selection. Plates were incubated in a humidity chamber at 37 °C for three days and then used to inoculate a 10 ml starter culture of PA-0.5G defined media (Appendix B.1) with 0.05% (v/v) Tween 80, hygromycin B (50 $\mu g.ml^{-1}$) and kanamycin (50 $\mu g.ml^{-1}$). This was grown for 48 hrs at 37 ºC with shaking, then used at a 1:100 dilution to inoculate expression cultures of ZYP-5052 auto-induction media (Appendix B.1) with 0.05% (v/v) Tween 80, hygromycin B (50 $\mu g.ml^{-1}$) and kanamycin (50 $\mu g.ml^{-1}$). The optimal expression conditions for protein purification were determined to be growth for 24 hrs at 37 ºC and then for a further 4 days at 28 ºC (Section 3.3.1.1). Cells were harvested by centrifugation (4,500 x *g* 20 min, 4 °C), and either frozen at -80 °C or used immediately for protein purification (Section 3.2.3).

3.2.2.1 Small scale protein expression trails

Protein expression of Rv2893 was initially confirmed and optimised from small scale expression cultures using His-tag binding with Ni Sepharose™ High Performance beads (GE Healthcare, Sweden). Expression cultures (5 ml) were grown at either 37 ºC for 4 days; 28 °C for 5 days; or 37 °C for 24 hrs followed by 28 °C for 4 days. Cells were harvested by centrifugation and resuspended in 500 µl sodium phosphate lysis buffer (50 mM sodium phosphate buffer, 200 mM NaCl, 20 mM imidazole, pH 7.4). Cells were lysed on ice by sonication and the insoluble fraction separated by centrifugation. The soluble fraction was mixed with 25 µl Ni beads that had been pre-equilibrated with lysis buffer. Beads were incubated at room temperature for 15 min in a thermomixer at with shaking at 1,000 rpm, then

left to settle for 5 min, centrifuged (600 x *g*, 1 min), and the supernatant removed. Beads were washed twice with 1 ml of lysis buffer then analysed by SDS-PAGE. During the procedure, samples of whole cells, the soluble fraction, insoluble fraction and flow though were also set aside for analysis by SDS-PAGE.

3.2.3 Protein purification

Cell pellets from 500 ml expression cultures grown as in Section 3.2.2 were resuspended in 25 ml lysis buffer (50 mM HEPES, 200 mM NaCl, 20 mM imidazole, 10% glycerol, pH 7.4) and lysed on ice by sonication. Lysed cells were centrifuged (20,000 x *g*, 20 minutes, 4 °C) to separate the supernatant from insoluble cellular debris. Protein was isolated from the soluble fraction by immobilised metal affinity chromatography (IMAC) and anion exchange chromatography and excess salt removed either by dialysis or using a desalting column.

3.2.3.1 Immobilised metal affinity chromatography purification

IMAC was performed using a HisTrap HP column (GE Healthcare, Sweden) on a ÄKTA Basic or Purifier FPLC system (GE Healthcare, Sweden), or an NGC Quest™ 10 Plus Chromatography System (Bio-Rad, U.S.A.). The supernatant from lysed cells was successively filtered through 1.2 μm and 0.45 μm syringe filters and loaded onto the column. Weakly bound non-specific protein was removed by washing the column with 20 ml of 4% elution buffer (50 mM HEPES, 200 mM NaCl, 0.5 M imidazole, 10% glycerol, pH 7.4): 96% lysis buffer at a flow rate of 1 $ml.min^{-1}$. Rv2893 was eluted using a gradient of 0–100% elution buffer over 75 ml at a flow rate of 1 $ml.min^{-1}$ and collected in 2 ml aliquots. Rv2893 containing fractions were identified via the 280 nm absorbance trace and/or SDS-PAGE.

3.2.3.2 Anion exchange chromatography

Anion exchange was performed using a 5 ml HiTrap Q XL anion exchange column (GE Healthcare, Sweden) on a ÄKTA Basic or Purifier FPLC system (GE Healthcare, Sweden), or an NGC Quest™ 10 Plus Chromatography System (Bio-Rad, U.S.A.). IMAC purified Rv2893 fractions (Section 3.2.3.1) were pooled

and diluted with three volumes of anion dilution buffer (50 mM HEPES, 10% glycerol, pH 7.4), filtered through a 0.2 μm syringe filter and loaded onto the column. Unbound protein was removed by washing the column with 10 ml start buffer (50 mM HEPES, 50 mM NaCl, 10% glycerol, pH 7.4) at a 1 ml.min^{-1} flow rate. Rv2893 was eluted using a gradient of 0–100% elution buffer (50 mM HEPES, 1 M NaCl, 10% glycerol, pH 7.4) over 40 ml at a flow rate of 1 ml.min^{-1} and collected in 2 ml aliquots. Fractions containing Rv2893 were identified via the 280 nm absorbance trace and/or by SDS-PAGE.

3.2.3.3 Buffer exchange

For large scale dialysis, typically the most concentrated protein fraction from anion exchange (2 mL) was dialyzed against 500 ml target buffer (50 mM HEPES, 200 mM NaCl, 10% glycerol, pH 7.4) overnight at 4 ºC using 6–8 kDa molecular weight cut off Spectra Por® dialysis tubing (Spectrum Laboratories, U.S.A.). Alternatively, a 5 ml HiTrap Desalting column (GE Healthcare, Sweden) was used to desalt protein samples into target buffer as per the manufacturer's instructions.

3.2.3.4 Size exclusion chromatography

Prior to the establishment of anion and buffer exchange protocols, Rv2893 was purified by size exclusion chromatography for crystallography screens. Protein containing fractions from IMAC purification were pooled and concentrated by centrifugation in a 2 ml or 20 ml Vivaspin concentrator with a 10 kDa molecular weight cut off (Sartorius AG, Germany). Protein was filtered through a 0.2 μm Minisart syringe filter and 0.5 ml was loaded on to a Superdex^{TM} 200 10/300 GL column (GE Healthcare, Sweden) with size exclusion buffer (50 mM HEPES, 200 mM NaCl, 10% glycerol, pH 7.4). Protein fractions were separated and eluted in size exclusion buffer at a flow rate of 0.35 ml.min^{-1} and collected in 0.5 ml aliquots. Fractions containing Rv2893 were identified from the 280 nm absorbance trace and/or SDS-PAGE.

3.2.3.5 Determination of protein concentration by A_{280}

Protein concentration was determined by measuring absorbance at 280 nm using a Nanodrop 2000 UV-vis spectrophotometer (Thermo Fisher scientific, U.S.A.) and correcting by the theoretical extinction coefficient (ε) as calculated using the ExPASy ProtParam tool (Gasteiger *et al.* 2005).

3.2.4 Protein characterisation

3.2.4.1 Molecular weight determination by size exclusion chromatography

A Superdex^{TM} 200 Increase 10/300 GL analytical size exclusion column (GE Healthcare, Sweden) was calibrated using six protein standards from high & low molecular weight gel filtration calibration kits (GE Healthcare, Sweden), as per kit instructions. The column void volume (V_0) was determined with 1 mg.ml^{-1} blue dextran dye. Standards and dye were eluted with size exclusion buffer (50 mM HEPES, 200 mM NaCl, 10% glycerol, pH 7.4) at a flow rate of 0.35 ml.min^{-1} using an NGC Quest™ 10 Plus Chromatography System (Bio-Rad, U.S.A.). A sample of Rv2893^{G72S} purified by IMAC and anion exchange chromatography was separately eluted at a flow rate of 0.35 ml.min^{-1} and the retention volume determined from the 280 nm absorbance trace. Retention volumes of protein standards were used to construct a calibration curve by plotting K_{av} (Equation 3.1) versus molecular weight and the molecular weight of Rv2893 was calculated using the standard curve.

$$\mathbf{K_{av} = (V_e - V_o) / (V_c - V_o)} \qquad \text{(Equation 3.1)}$$

Where K_{av} = gel phase distribution coefficient, V_e = elution volume; V_o = column void volume; and V_c = geometric column volume

3.2.4.2 Melt temperature determination by differential scanning fluorimetry

Differential scanning fluorimetry (DSF) using the SYPRO method (Lo *et al.* 2004) was used to compare the melting temperature (T_m) of Rv2893^{G72S} and Rv2893^{H37Rv}. SYPRO dye (Life Technologies, U.S.A.) was diluted to 300X in buffer (50 mM HEPES, 200 mM NaCl, 10% glycerol, pH 7.4) and 7.5 µl of this was added to protein to a final volume of 25 µl with a protein concentration of 0.1 mg.ml^{-1}. Melts were performed in a Rotor-Gene^{TM} real time PCR machine (Corbett Life Science,

Australia) using a temperature gradient of 25–99 ºC with continuous fluorescence monitoring at excitation and emission wavelengths of 470 and 555 nm, respectively. Protein samples were assayed in triplicate alongside a blank containing no protein. Data was analysed in GraphPad Prism (www.graphpad.com). The T_m was taken as the inflection point of the raw data ($d^2F/dT^2 = 0$) and mean T_m values for triplicate samples were calculated.

3.2.5 Protein Crystallography

3.2.5.1 Crystallisation robot screens

Initial determination of crystallisation conditions was performed using the sitting drop method. Screens were set up using a Mosquito® Crystal robotic system (TTP Labtech Ltd, UK), using the PEGRx HT and Index HT crystallisation screens (Hampton Research, U.S.A.). Conditions were set up with 200 nl drops of equal volumes of protein and reservoir solution, against a 100 µl reservoir in Intelli-Plates (Hampton Research, U.S.A.), and stored at 18 ºC.

3.2.5.2 Crystallisation fine screens

Crystallisation conditions were optimised by hanging-drop vapour diffusion in 24-well plates (Hampton Research, U.S.A.). In general, fine screens varied the precipitant concentration and pH of conditions identified in robot screens. Drops of 2–6 µl containing protein and reservoir solution in a 1:1 ratio were pipetted onto a siliconised coverslip which was then inverted and sealed over 0.5 ml reservoir solution. Plates were stored at either 18 °C, 21 °C, or 25 °C. Final crystallisation conditions used to grow crystals for data collection are shown in Table 3.1.

3.2.5.3 Heavy atom derivatisation

Electrophoretic mobility shift assays were employed to identify promising heavy atoms for experimental phasing, (Boggon and Shapiro 2000). Mixtures containing different heavy metal compounds at 2 mM and $Rv2893^{G72S}$ at 1 $mg.ml^{-1}$ were prepared in protein buffer (50 mM HEPES, 200 mM NaCl, 10% glycerol, pH 7.4) to a final volume of 10 µl and incubated on ice for 15 minutes. The protein was then

analysed by Native PAGE on a 12% polyacrylamide gel to identify compounds that produced a band shift indicating heavy atom derivatisation. Promising compounds were soaked into crystals to create derivates for experimental phasing. A 2 µl drop of reservoir solution containing the heavy metal compound at 1–2 mM was pipetted onto a coverslip and a crystal was transferred to this drop. This was placed over the reservoir and left overnight before looping and freezing the crystal for testing.

3.2.5.4 Ligand soaks

A sample of F_{420} was kindly provided by Ghader Bashiri (University of Auckland, N.Z.). This had been purified from *M. smegmatis* $mc^2$4517 cells engineered to produce high levels of F_{420} (Bashiri *et al.* 2010) using solvent extraction and anion exchange as previously described (Bashiri *et al.* 2010; Isabelle *et al.* 2002), followed by high performance liquid chromatography. F_{420} concentration was calculated from an extinction coefficient of 25 $mm^{-1}.cm^{-1}$ at 400 nm (Jacobson and Walsh 1984) in 25 mM Sodium phosphate buffer, pH 7.0. A 2 µl drop of reservoir solution containing 1 mM F_{420} was pipetted onto a coverslip and a crystal was transferred to this drop. This was placed over the reservoir and left overnight before looping and freezing the crystal for testing.

3.2.6 X-Ray diffraction data collection and structure determination

3.2.6.1 Data collection

Crystals were removed from crystallisation drops with a cryo-loop and then briefly submerged in cryo-protectant solution (crystallisation reservoir solution with either 35% polyethylene glycol (PEG) 3350 or 10% PEG 3350 and 20% glycerol) before being flash frozen in liquid nitrogen. X-ray diffraction data were collected at the Australian Synchrotron (Melbourne, VIC). Data for apo crystals were collected at the macromolecular crystallography MX1 beamline using a Quantum 210r detector (ADSC, U.S.A.). Data from heavy metal and F_{420} soaked crystals were collected on the MX2 beamline using an EIGER x 16 M detector (Dectris, Switzerland).

3.2.6.2 Structure solution using single-wavelength anomalous dispersion

Phase information was obtained experimentally using the anomalous dispersion of gold bound to a $Rv2893^{H37Rv}$ crystal derivatised in $KAuCl_4$ as per Section 3.2.5.3. Multiple datasets were collected from the same crystal at 0.9537 Å as in Section 3.2.6.1. Datasets were auto-processed at the Australian synchrotron in XDS (Kabsch 2010) then merged using *phenix.scale_and_merge* with optimisation for anomalous differences (Akey *et al.* 2016). Gold sites were identified and refined using the Auto-Rickshaw platform (Panjikar *et al.* 2005) and a solution was determined by single-wavelength anomalous diffraction (SAD) phasing in the AutoSol wizard in PHENIX (Terwilliger *et al.* 2009) using these sites and the merged $KAuCl_4$ dataset as input. Automated model building with refinement was done using AutoBuild in PHENIX (Terwilliger *et al.* 2008), with the inclusion of a higher resolution isomorphous data set (obtained from a Hg acetate soaked crystal) to assist model building and refinement.

3.2.6.3 Structure solution by molecular replacement

The apo-$Rv2893^{G72S}$ structure was solved by molecular replacement in PHASER (McCoy *et al.* 2007) using the experimentally phased solution (Section 3.3.5) as the search model. Automated model building and refinement was done using the AutoBuild wizard in PHENIX (Terwilliger *et al.* 2008) and manual building performed in COOT (Emsley and Cowtan 2004; Emsley *et al.* 2010). The model was refined as in Section 3.2.6.4, and then used as the search model for molecular replacement for subsequent structures.

Diffraction data for molecular replacement collected using a CCD detector (native crystals, Section 3.2.6.1) and were integrated in iMOSFLM (Battye *et al.* 2011) then scaled and merged in the CCP4 software suite (Winn *et al.* 2011) using the Data Reduction task. This sequentially runs POINTLESS to determine the Laue group and possible space group (Evans 2006), AIMLESS to scale and merge multiple observations of reflections (Evans and Murshudov 2013), and then CTRUNCATE to convert intensities to structure factor amplitudes, and provides merging and other data processing statistics. Diffraction data collected using an EIGER detector (F_{420} soaked crystals, Section 3.2.6.1) were indexed, integrated and scaled using XDS (Kabsch 2010) and then merged using AIMLESS (Evans and

Murshudov 2013). For all datasets, 5% of reflections were randomly marked for the R_{free} set for validation during structural refinement.

3.2.6.4 Model Refinement

Models were initially refined using real-space and rigid body refinement and simulated annealing in *phenix.refine* (Afonine *et al.* 2012). Manual model building was performed in COOT (Emsley and Cowtan 2004; Emsley *et al.* 2010) with the electron density 2|FO|-|FC| and |FO|-|FC| maps contoured to 1 σ and 3 σ respectively. Ligands were manually added and fitted in COOT following identification of regions of unmodeled density. Model refinement using REFMAC5 (Murshudov *et al.* 2011) and *phenix.refine* was performed regularly during manual building and ligand fitting to ensure alterations improved the model fit to the data. Waters were added in the final stages of refinement in *phenix.refine*.

3.2.6.5 Structure Analysis

Final refinement was performed in REFMAC5. Model statistics were generated using *phenix.table_one* and Ramachandran analysis performed using RAMPAGE (Lovell *et al.* 2003). Structure analysis and visualisation was performed in COOT and PyMOL (The PyMOL Molecular Graphics System, Version 2.2 Schrödinger, LLC) and PyMOL used for image generation. Root mean square differences (RMSD) between protein molecules were calculated in PROsmart (Nicholls *et al.* 2014) in the CC4 suite. The dimer interface was analysed and surface area determined using PDBePISA (Krissinel and Henrick 2007). The Protein-Ligand Interaction Profiler (Salentin *et al.* 2015) and LigPlot+ (Laskowski and Swindells 2011) tools were used to assist the examination of the F_{420} binding mode. Hydrogen bonds were inferred where bond distances were within 2.5–3.2 Å for strong bonds and 3.2–4.0 Å for weak bonds (Jeffrey and Jeffrey 1997).

3.3 Results

3.3.1 Heterologous expression and purification of Rv2893

The Rv2893 genes from *M. tuberculosis* H37Rv ($Rv2893^{H37Rv}$) and the Rangipo G72S variant ($Rv2893^{G72S}$) were cloned into the mycobacterial expression vector pYUB28b (Bashiri *et al.* 2010) with an N-terminal His-tag (Section 3.2.1). Ligation reactions were first transformed into *E. coli* TOP10 cells for transformant selection, vector purification and sequence validation. Sequence validated plasmids were then transformed in *M. smegmatis* mc^2 4517 cells (Wang *et al.* 2010) for heterologous protein expression and purification (Section 3.3.2).

3.3.1.1 Small scale expression trails

Small scale expression trails were first performed to evaluate Rv2893 protein expression in *M. smegmatis* mc^2 4517 and his-tag binding affinity using Ni beads (Section 3.2.2.1). Small scale cultures expressing $Rv2893^{G72S}$ were grown at either 37 °C for 4 days; 28 °C for 5 days; or at 28 °C for 4 days following an initial growth period at 37 °C for 24 hrs. $Rv2893^{G72S}$ predominantly expressed insolubly at 37 °C with a small amount of protein binding to the Ni beads (Figure 3.3A). A subtle improvement was seen at 28 °C with a stronger band of the correct size band in the soluble fraction, however, expression overall was lower (Figure 3.3B). Dropping the temperature to 28 °C after growing for 24 hrs at 37 °C had a marked effect on both soluble expression and overall yield (Figure 3.3C). $Rv2893^{H37Rv}$ expression was then trailed under the same conditions resulting in soluble expression with ample binding to the Ni beads (Figure 3.3D). Two bands of similar size and intensity were also observed on the Ni beads when Rv2893 was expressed at 28 °C and the same bands were also apparent at 37 ºC, although the smaller band was weaker. These two bands were both found to be Rv2893 and eluted as separate peaks in IMAC purification (Section 3.3.2).

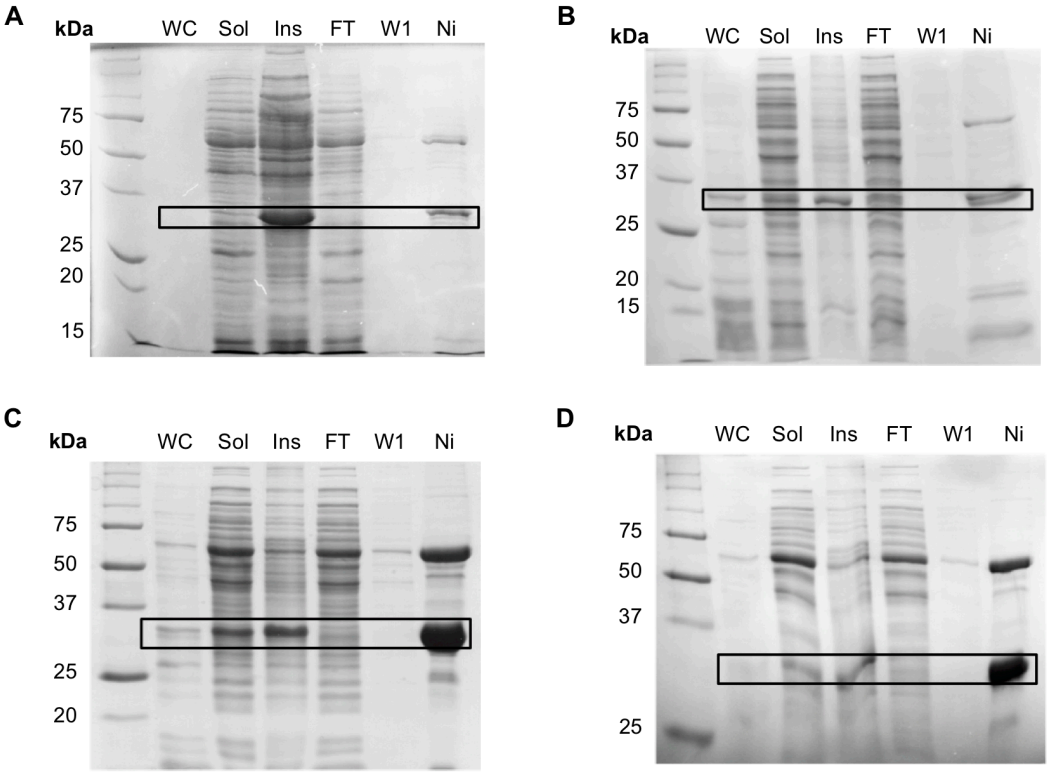


Figure 3.3. *Small scale protein expression trails and his-tag binding of Rv2893.* (A–C) $Rv2893^{G72S}$ and (D) $Rv2893^{H37Rv}$. $Rv2893^{G72S}$ cultures were grown at (A) 37 °C for 4 days; (B) 28 °C for 5 days; (C) 37 °C for 24 hrs then 28 °C for 4 days; and (D) $Rv2893^{H37Rv}$ grown at 37 °C for 24 hrs then 28 °C for 4 days. Samples were separated on 12% SDS-PAGE gels. Black boxes indicate expressed Rv2893. Labels: WC, whole cell sample; Sol, soluble fraction; Ins, insoluble fraction; FT, flow through; W1, wash one; and Ni, Ni beads with protein bound after wash steps.

3.3.2 Large scale expression and protein purification

Having established soluble expression for Rv2893^{G72S} and Rv2893^{H37Rv} at small scale, *M. smegmatis* expression was scaled up to large volume (500 mL) cultures for protein purification. Soluble protein suitable for crystallography was successfully purified using a two-step protocol comprised of IMAC (Section 3.2.3.1) followed by anion exchange (Section 3.2.3.2). Representative chromatograms are shown in Figure 3.4. Excess salt was removed either by dialysis or using a desalting column (Section 3.2.3.3). Anion purified and desalted protein was put directly into crystallisation screens. Using this protocol, protein suitable for crystallisation was purified at concentrations typically ~5–7 mg.ml^{-1}, with concentrations as high as 8.7 mg.ml^{-1} obtained.

Rv2839^{G72S} eluted in two peaks by IMAC purification with a major peak at ~220 mM imidazole and a second minor peak at ~400 mM imidazole (Figure 3.4A). SDS-PAGE bands from each peak were both similar in size to that expected for Rv2893 and corresponded to those observed during small-scale expression trails (Figure 3.3). Rv2839^{H37Rv} also produced the same two bands, however these eluted in three peaks by IMAC purification; the first peak predominantly contained the lower weight band, the second peak was a mixed population of the two bands, and the third peak was predominantly the higher weight band (Figure 3.4B). Bands from Rv2893^{G72S} were sent for analysis by mass spectroscopy at MS^3 Solutions (Hamilton, N.Z.) and both were confirmed to be *M. tuberculosis* Rv2893 with no detectable differences between the two bands. Analysis by absorbance and fluorescence scans failed to detect cofactor F_{420} or any differences between the two peaks. The structures of Rv2893 also confirmed the absence of F_{420} in crystals grown from native purified protein. Because of the requirement for highly pure homogeneous samples for protein crystallisation, only the first peak from both Rv2839^{G72S} and Rv2839^{H37Rv} was carried through to further purification by anion exchange chromatography. Both Rv2893^{G72S} and Rv2893^{H37Rv} eluted in a single sharp peak at ~330 mM NaCl by anion exchange and the protein obtained was highly pure and of suitable concentration for crystallography (Figure 3.4C–D).

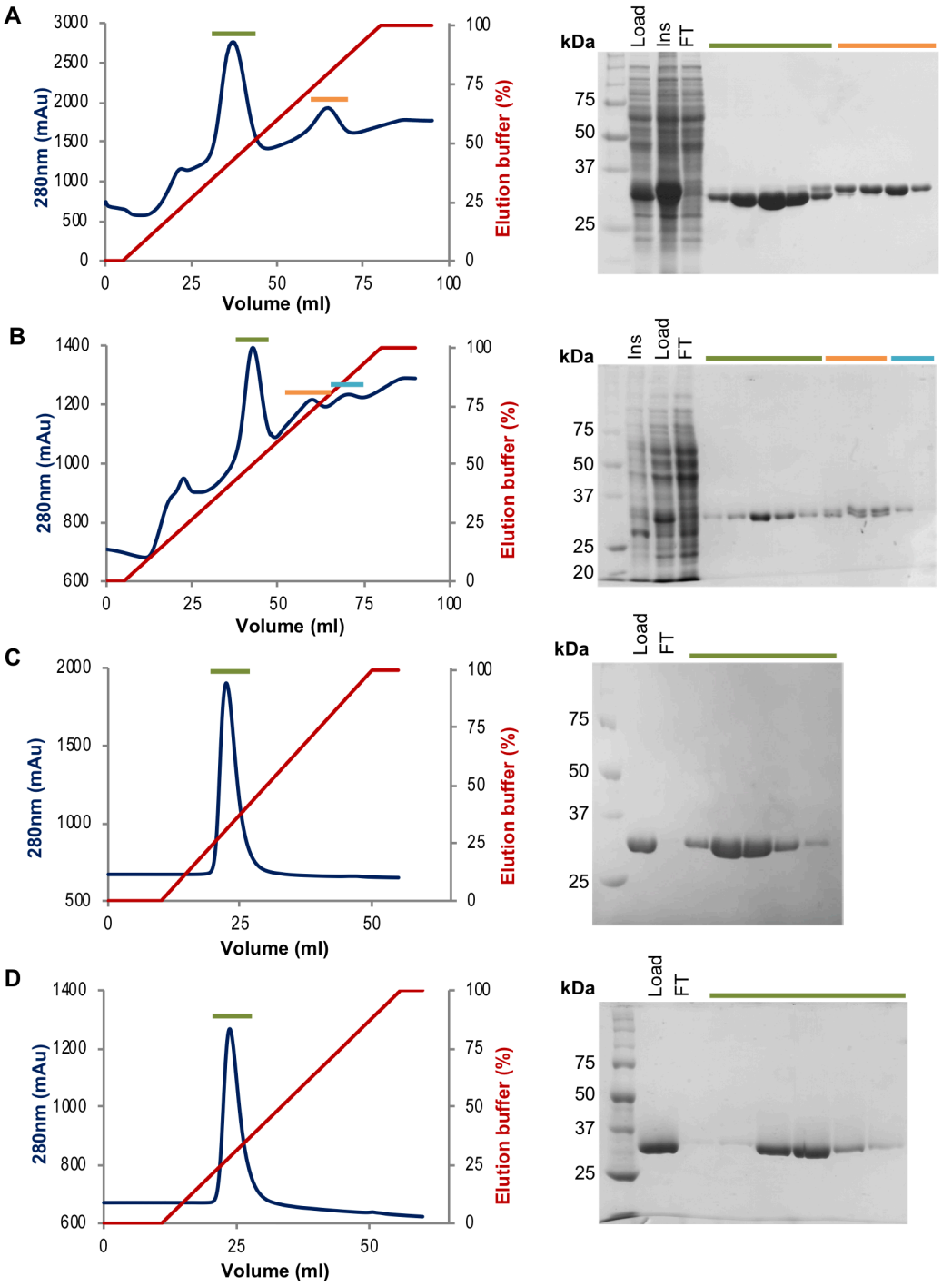


Figure 3.4. *Rv2893 protein purification by immobilised metal affinity chromatography (IMAC) and anion exchange chromatography.* (A–B) IMAC purification of (A) Rv2893^{G72S} and (B) Rv2893^{H37Rv}. Chromatograms depict 280 nm UV absorbance and elution profile over a 50 mM to 0.5 M imidazole elution gradient. Corresponding 12% SDS-PAGE gels show the soluble fraction loaded onto the column (Load), insoluble fraction (Ins), flow through (FT) and eluted fractions corresponding to each coloured bar. IMAC purifications shown contained 0.01% DDM in buffers. (C–D) Anion exchange chromatography purification of (C) Rv2893^{G72S} and (D) Rv2893^{H37Rv}. Chromatograms depict 280 nm UV absorbance and elution profile over a 50 mM to 1 M NaCl elution gradient. Corresponding 12% SDS-PAGE gels show the sample that was loaded onto the column (Load), the column flow through (FT) and eluted fractions.

One of the major challenges encountered establishing purification protocols for Rv2893 was that the protein was unstable and typically precipitated out of solution within 24 hours. Concentrating Rv2893 was also challenging as it precipitated on protein concentrator spin columns. Anion exchange provided a means to circumvent this problem and also removed the imidazole from IMAC purification. The most concentrated anion exchange fraction eluted was typically 150–200% more concentrated from that eluted by IMAC, resulting in concentrations sufficient for crystallisation without the need for additional concentration steps.

Various approaches were trailed to improve the stability of Rv2893 including different additives, storage at different temperatures, and buffer screening using different lysis buffers, small scale dialysis and DSF thermal shift assays (Ericsson *et al.* 2006). The addition of glycerol had the most notable effect on stability and was included at 10% in all buffers. The inclusion of 1 mM dithiothreitol (DTT) in purification buffers also improved stability and this effect was retained when DTT was included only in lysis buffers and then re-added to purified protein for storage. This reduced the rate of protein aggregation by one or two days and was therefore included in all subsequent purifications. Earlier purifications included the addition of 0.01-0.03% n-Dodecyl β-D-maltoside (DDM). DDM is a non-ionic detergent that is effective for protein solubilisation and initially appeared promising at improving the stability of Rv2893. However, it still failed to retain protein in solution for >24 hrs, even at higher concentrations (1–2 %), and so was removed from purification buffers. Considering these challenges, the focus was made on laying down crystal screens using protein within 24 hrs of purification as freshly purified protein consistently produced the highest quality crystals.

The oligomeric state of purified Rv2893 was determined using analytical size exclusion chromatography (Section 3.2.4.1). The elution volume corresponded to 62.4 kDa, which is consistent with a dimer in solution when compared to the molecular weight of the monomer of 36.8 kDa (calculated using ProtParam, http://web.expasy.org/protparam).

3.3.3 Melt temperature of Rv2893^{G72S} and Rv2893^{H37Rv}

The melt temperatures of Rv2893^{H37Rv} and Rv2893^{G72S} were estimated using DSF to determine if the Rangipo G72S mutation affects the thermal stability of the protein. The T_m was determined from the highest peak of the first derivative of the melting curve (the dF/dT maxima) as the second derivate was too noisy. This corresponds to the inflection point of the melting curve, which is the temperature at which half the protein is in the unfolded state. The T_m of Rv2893^{H37Rv} was estimated to be 54.9 ºC and the Rv2893^{G72S} variant had a 4 ºC higher T_m of 58.9 ºC (average of three replicates) (Figure 3.5). This indicates that that the Rangipo Rv2893^{G72S} variant is more temperature stable than Rv2893^{H37Rv}.

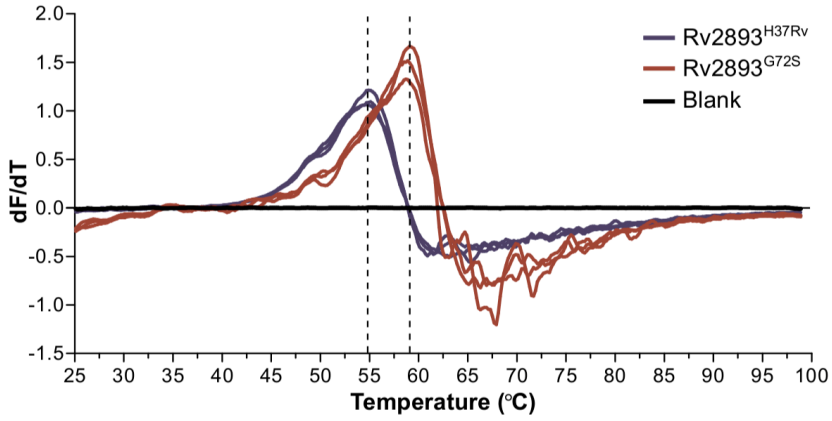


Figure 3.5. *Melting temperature (T_m) comparison of Rv2893^{H37Rv} (blue) and Rv2893^{G72S} (red).* The first derivative of the fluorescence output is shown. T_m values were determined from the dF/dT maxima and are indicated by dashed lines.

3.3.4 Crystallisation of Rv2893

High throughput crystallisation trails were performed using the sitting-drop vapour diffusion method (Section 3.2.5.1) with Rv2893^{G72S} at 2.6 mg.ml^{-1} and 2.2 mg.ml^{-1} purified by IMAC and size exclusion chromatography (Section 3.2.3, with 0.01% DDM in buffers). The best condition (0.2 M ammonium acetate, 0.1 M Bis-tris pH 5.5–6.5, 25% PEG 3350) was pursued using hanging-drop vapour diffusion fine screens (Section 3.2.5.2) using protein purified by IMAC and anion exchange chromatography directly eluted from anion exchange and/or following the removal of excess salt by dialysis or using a desalting column (Section 3.2.3).

Crystals appeared overnight and reached their maximum size within a few days. In addition to varying the precipitant concentration and pH of crystallisation conditions, a range of other approaches were trialled to improve crystal quality and diffraction. Additive screening and crystal seeding (batch and streak seeding) failed to improve crystal morphology and/or diffraction. Increasing the crystallisation temperature from 18 ºC to 25 ºC produced more single crystals and thicker crystals which diffracted well (Figure 3.6C, Table 3.1). Larger drop sizes (up to 6 µl) and the addition of glycerol or DTT inconsistently improved crystal quality. The final conditions used to grow crystals for data collection are reported in Table 3.1.

Table 3.1. *Conditions used to grow Rv2893 crystals used for data collection.*

Dataset	Protein Conc.	Purification[1]	Crystallisation conditions	Temp.
$KAuCl_4$-$Rv2893^{H37Rv}$	2.8 mg.ml^{-1}	IMAC + anion exchange + desalting	0.1 M Bis-tris pH 6.0 0.2 M Ammonium acetate 16% PEG 3350	25 ºC
Hg acetate-$Rv2893^{H37Rv}$	2.7 mg.ml^{-1}	IMAC + anion exchange + desalting	0.1 M Bis-tris pH 6.2 0.2 M Ammonium acetate 25% PEG 3350	25 ºC
apo-$Rv2893^{G72S}$	5.0 mg.ml^{-1}	IMAC[2] + anion exchange + dialysis	0.1 M Bis-tris pH 6.6 0.2 M Ammonium acetate 25% PEG 3350	18 ºC
apo-$Rv2893^{H37Rv}$	5.7 mg.ml^{-1}	IMAC[2] + anion exchange	0.1 M Bis-tris pH 7.0 0.2 M Ammonium acetate 19% PEG 3350	21 ºC
F_{420}-$Rv2893^{G72S}$	7.2 mg.ml^{-1}	IMAC + anion exchange + dialysis	0.1 M Bis-tris pH 6.5 0.2 M Ammonium acetate 27% PEG 3350	25 ºC
F_{420}-$Rv2893^{H37Rv}$	1.8 mg.ml^{-1}	IMAC[3] + anion exchange + dialysis	0.1 M Bis-tris pH 6.3 0.2 M Ammonium acetate 18% PEG 3350, 1mM DTT	25 ºC

[1] Protein purified as in Sections 3.2.3.1–3.2.3.3.
[2] 1 mM DDM in lysis and IMAC buffers.
[3] 1 mM DTT in lysis buffer.

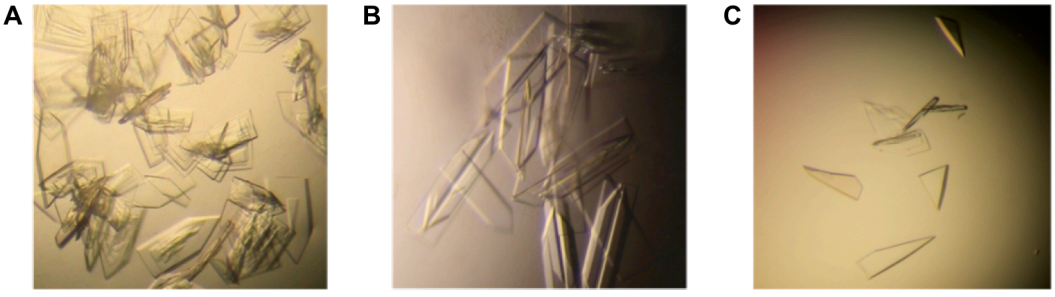


Figure 3.6. *Typical morphology of Rv2893 crystals*. (A) $Rv2893^{G72S}$ crystals grown at 18 ºC. (B) $Rv2893^{H37Rv}$ crystals grown at 18 ºC. (C) $Rv2893^{H37Rv}$ crystals grown at 25 ºC. These crystals were used for $KAuCl_4$ soaks to solve the structure by experimental phasing.

3.3.5 Experimental Phasing

Initial attempts to solve the Rv2893 structure by molecular replacement using *M. tuberculosis* FGD (3C8N) as the search model were unsuccessful and so phases were determined experimentally. To identify potential heavy-atom derivatives to use for experimental phasing, native PAGE gel shift assays were performed (Boggon and Shapiro 2000). $Rv2893^{G72S}$ was incubated with various heavy-atom compounds and then analysed by native PAGE gel electrophoresis to identify heavy atoms that bind to the protein, as indicated by a band shift (Section 3.2.5.3). Shifts were detected for $HgCl_2$, Hg acetate and K_2PtCl_4 (Figure 3.7). All gold compounds screened ($KAu(CN)_2$, $KAuCl_4$ and $KAuBr_4$), visibly precipitated Rv2893 and the protein did not run on a Native-PAGE gel. While this indicated denaturation of the protein it also implied that gold was interacting with Rv2893, therefore gold was pursued in addition to mercury and platinum for experimental phasing.

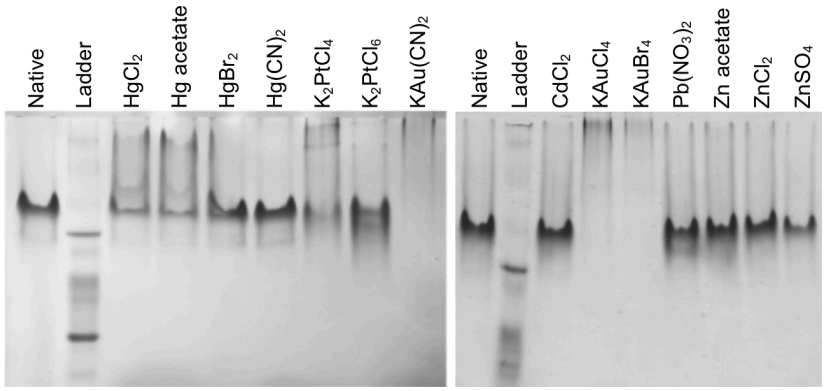


Figure 3.7. *Heavy metal native PAGE gel shift assay to identify heavy atom derivatives for experimental phasing of Rv2893.*

Rv2893 crystals soaked overnight in $HgCl_2$, Hg acetate, K_2PtCl_4, K_2PtCl_6 and $KAuCl_4$ (Section 3.2.5.3) were taken to the Australian synchrotron for screening and data collection. Mercury, platinum and gold were all found to bind to the protein as indicated by excitation scans and/or the presence of an anomalous signal in the diffraction data. The anomalous signal from mercury and platinum however, was not adequate to solve the structure and the final solution was obtained using gold-soaked crystals and SAD phasing.

Five datasets of 180 º/1800 frames were collected at 12.1 KeV (0.18 KeV above the L_{III} edge of gold, theoretical $f'' = 9.94$ e^- and $f' = -10.39$ e^-) from various points and starting angles on a single $KAuCl_4$ soaked crystal (Table 3.1). Data was processed

in the monoclinic space group $P12_11$ and the maximum resolution of individual datasets ranged from 3.14–3.68 Å. Individual datasets were scaled and merged together to optimise anomalous differences (Table 3.2). The anomalous signal was predicted to extended to 3.8 Å, and the overall anomalous correlation was 0.542. Twenty gold sites were identified using SHELX and refined with BP3 in the Autorickshaw server (Panjikar *et al.* 2005). The PHENIX AutoSol wizard (Terwilliger *et al.* 2009) identified an additional nine gold sites and successfully calculated phases from the anomalous differences. The calculated phases allowed model building and refinement and the inclusion of a higher resolution dataset collected from an isomorphous Hg acetate-soaked crystal (2.80 Å) was used to assist the model building (Table 3.2). The resultant model was not subject to further manual refinement as it was sufficient for molecular replacement with high resolution data. That we aware of, this was the first structure to be experimentally solved using the new EIGER detector on MX2 at the Australian Synchrotron.

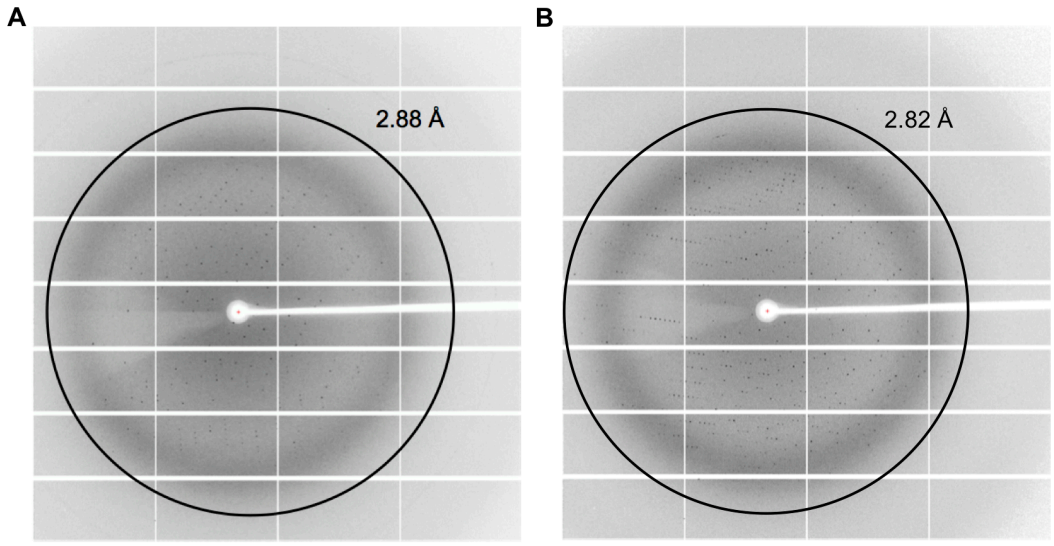


Figure 3.8. *Diffraction of Rv2893 heavy metal derivatives.* (A) $Rv2893^{H37Rv}$ derivatised with gold ($KAuCl_4$) was used for SAD phasing. (B) $Rv2893^{H37Rv}$ derivatised with mercury (Hg acetate) was used as a higher resolution isomorphous dataset to improve model building.

Table 3.2. *Data collection statistics for Rv2893 SAD-phasing and statistics for the experimentally phased model.* Statistics for the highest-resolution shell are shown in parentheses. Hg acetate-Rv2893^{H37Rv} data was used at the model building stage but not for phasing.

Data collection[1]	**$KAuCl_4$-Rv2893^{H37Rv}**	**Hg acetate-Rv2893^{H37Rv}**
Wavelength (Å)	1.0247	0.9999
Space group	P 1 21 1	P 1 21 1
Unit cell dimensions a/b/c (Å)	75.82/116.32/88.18	76.37/117.47/87.63
α/β/γ (°)	90.00/90.88/90.00	90.00/90.81/90.00
Resolution range (Å)	3.13–48.55 (3.13–3.35)	2.80–48.79 (2.80–2.93)
R_{merge}	0.259 (0.608)	0.110 (0.628)
R_{pim}	0.067 (0.381)	0.045 (0.254)
Mean I/σ	6.2 (1.3)	11.8 (3.1)
$CC_{(1/2)}$	0.987 (0.783)	0.997 (0.857)
Total number of observations	351907 (14044)	264418 (31420)
Total number unique	26313 (4119)	37887 (4529)
Completeness (%)	97.0 (84.3)	99.6 (97.2)
Multiplicity	13.4 (3.4)	7.0 (6.9)
Mosaicity	0.13	0.1
Phasing statistics[2]		
Anomalous Completeness (%)	91.25	
Mean anomalous difference	0.0976	
Anomalous CC between half-datasets	0.524 (0.104)	
Model		
R_{work}	0.302	
R_{free}	0.363	
Protein residues	947	

[1] Data collection statistics calculated in Aimless. For Au-Rv2893^{H37Rv}, data statistics are for five merged datasets collected from the same crystal.
[2] Phasing statistics calculated in *phenix.scale_and_merge* and *phenix.xtriage*

3.3.6 Structure determination by molecular replacement

High resolution data obtained from native and F_{420} soaked crystals were used to solve apo- and F_{420}-bound $Rv2893^{G72S}$ and $Rv2893^{H37Rv}$ structures by molecular replacement. The growth conditions of crystals used for final data collection are summarised in Table 3.1. To obtain structures complexed with F_{420}, apo crystals were soaked overnight in 1 mM F_{420} (Section 3.2.5.4). Data for apo crystals were collected on the MX1 beamline at the Australian synchrotron using a CCD detector and data from F_{420} soaked crystals on the MX2 beamline at the Australian synchrotron using an EIGER detector. Datasets were all collected at 13.0 KeV and frames from 360 º oscillation were collected for each crystal. A delta of 0.5 º was used for collecting MX1 data, and MX2 data was collected using a delta of 0.1 º.

All datasets were processed in space group $P12_11$ to thresholds of $CC_{(1/2)} > 0.500$ and $I/\sigma > 1.5$ in the outer/high resolution shell (Table 3.3). The maximum resolution of all four datasets were in a similar range of 2.17–2.31 Å. Both the $Rv2893^{G72S}$ and $Rv2893^{H37Rv}$ native datasets contained eight molecules in the asymmetric unit (ASU) and had strong evidence of pseudo-translational symmetry ($p = 1.48 \times 10^{-4}$ and $p = 1.49 \times 10^{-4}$, respectively). Datasets from F_{420} soaked crystals had four molecules in the asymmetric unit and no crystal pathologies were detected.

The best dimer from the model obtained from the SAD solution (Section 3.3.5) was used as the search model for molecular replacement with the high resolution native $Rv2893^{G72S}$ dataset. Following refinement, the best monomer from this model was used as the search model to solve the remaining structures. Final statistics for refined models are given in Table 3.4. The higher R-factors for the apo structures are likely in part due to the translational pseudosymmetry in these datasets as this can lead to difficulties in refinement (Read *et al.* 2013).

Table 3.3. *Data collection statistics for apo and F_{420} bound $Rv2893^{G72S}$ and $Rv2893^{H37Rv}$.* Statistics for the highest-resolution shell are shown in parentheses.

Data Statistic	**apo-Rv2893^{H37Rv}**	**apo-Rv2893^{G72S}**	**F_{420}-Rv2893^{H37Rv}**	**F_{420}-Rv2893^{G72S}**
Wavelength (Å)	0.9537	0.9537	0.9537	0.9537
Space group	P 1 21 1	P 1 21 1	P 1 21 1	P 1 21 1
Unit cell dimensions a/b/c (Å)	114.75/117.17/116.40	114.82/116.94/116.35	80.63/118.81/85.61	81.12/118.23/86.50
α/β/γ (°)	90.00/96.16/90.00	90.00/96.21/90.00	90.00/91.23/90.00	90.00/91.59/90.00
Resolution range (Å)	2.30–64.47 (2.30–2.34)	2.12–57.83 (2.12–2.16)	2.17–18.80 (2.17–2.21)	2.31–48.80 (2.31–2.36)
R_{merge}	0.240 (1.447)	0.212 (1.322)	0.078 (1.329)	0.078 (1.094)
R_{pim}	0.093 (0.559)	0.082 (0.512)	0.032 (0.526)	0.032 (0.454)
Mean I/σ	8.4 (1.5)	9.3 (1.7)	13.6 (1.5)	12.9 (1.5)
$CC_{(1/2)}$	0.992 (0.560)	0.994 (0.513)	0.999 (0.608)	0.999 (0.565)
Total number of observations	1030086 (51414)	1306478 (64022)	593159 (32844)	494858 (30363)
Total number unique	135812 (6760)	172401 (8481)	84901 (4491)	71447 (4548)
Completeness (%)	100.0 (100.0)	99.7 (99.3)	99.7 (99.1)	100.0 (99.9)
Multiplicity	7.6 (7.6)	7.6 (7.5)	7.0 (7.3)	6.9 (6.7)
Mosaicity	0.56	0.59	0.15	0.11
Monomers in the ASU	8	8	4	4

Table 3.4. *Refinement and model statistics for native and F_{420} bound $Rv2893^{G72S}$ and $Rv2893^{H37Rv}$.*

Model statistic	apo-$Rv2893^{H37Rv}$	apo-$Rv2893^{G72S}$	F_{420}-$Rv2893^{H37Rv}$	F_{420}-$Rv2893^{G72S}$
R_{work}	0.204 (0.296)	0.197 (0.293)	0.192 (0.293)	0.192 (0.303)
R_{free}	0.237 (0.312)	0.232 (0.315)	0.220 (0.312)	0.222 (0.334)
Protein residues	2463	2458	1265	1264
Total number of non-hydrogen atoms	19066	19576	9520	9415
Protein	17864	17853	9110	9132
Ligands	16	26	95	89
Solvent	1186	1697	315	194
RMS Bonds (Å)	0.014	0.014	0.015	0.014
Angles (°)	1.79	1.8	1.86	1.84
Average B value (Å^2)	28.78	28.26	51.30	57.85
Protein	28.87	27.87	51.36	57.93
Ligands	43.77	44.97	55.58	64.91
Solvent	27.2	32.07	48.24	50.75
Ramachandran analysis[1]				
Number of residues in favoured region (%)	2376 (97.9%)	2371 (98.1%)	1218 (97.8%)	1221 (98.2%)
Number of residues in allowed region (%)	51 (2.1%)	45 (1.9%)	27 (2.2%)	23 (1.8%)
Number of residues in outlier region (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

[1] Ramachandran plots are presented in Appendix B.5.

3.3.6.1 Rv2893 structural models

The crystal structures of apo-Rv2893^{H37Rv} and apo-Rv2893^{G72S} were solved at 2.30 Å and 2.12 Å, respectively, and models refined to R_{work}/R_{free} values of 0.204/0.237 and 0.197/0.232 (Table 3.4). Crystal structures of F_{420}-Rv2893^{H37Rv} and F_{420}-Rv2893^{G72S} were solved at 2.17 Å and 2.31 Å respectively, and models refined to R_{work}/R_{free} values of 0.192/0.220 and 0.192/0.222.

Both apo structures were solved with four dimers in the ASU and F_{420} bound structures with two dimers in the ASU. The apo- and F_{420}-bound structures for both Rv2893^{H37Rv} and Rv2893^{G72S} have very similar overall conformations, as evident from the low overall root mean square differences (RMSD) in the Cα atomic positions between monomers (RMSD 0.11–1.53 Å, mean 0.60 Å) (Table 3.5, Appendix B.6). In both apo structures, each of the eight chains had a similar overall conformation with RMSD values ranging from 0.11–1.20 Å (mean, 0.47 Å). For F_{420}-Rv2893^{H37Rv} and F_{420}-Rv2893^{H37Rv} structures, RMSD between chains ranged from 0.69–1.47 Å (mean, 1.3 Å).

Up to three loops were missing in the electron density and were unable to be modelled in all chains. These loops are Gly197–Phe212 (the β_6-α_6 loop), Arg250–Ser272 (β_7-α_7 loop), and Glu304–Arg308 (β_8-α_8 loop) (Appendix B.7). In all chains, the first 25–26 residues from the N-terminus of the protein (including the His-tag) were not modelled, and 0–1 residue was missing from the C-terminus. In addition to these missing residues, various sidechains throughout the structures were missing electron density for the side chain atoms and so were pruned from the model. Non-protein electron density that accommodated PEG was present in both apo structures and glycerol was present in F_{420} bound structures.

Table 3.5. *RMSD in the Cα atomic positions between Rv2893 structures.* Average RMSD for all combinations of chain comparisons are reported, and minimum and maximum shown in brackets. RMSD between chains belonging to the same structure are italicised.

	apo-Rv2893^{H37Rv}	**apo-Rv2893^{G72S}**	**F_{420}-Rv2893^{H37Rv}**	**F_{420}-Rv2893^{G72S}**
apo-Rv2893^{H37Rv}	*0.47 (0.11–1.20)*	0.44 (0.11–1.22)	0.69 (0.26–1.53)	0.67 (0.27–1.53)
apo-Rv2893^{G72S}		*0.47 (0.18–1.16)*	0.68 (0.22–1.53)	0.66 (0.29–1.53)
F_{420}-Rv2893^{H37Rv}			*1.04 (0.74–1.47)*	0.81 (0.17–1.47)
F_{420}-Rv2893^{G72S}				*0.95 (0.69–1.25*)

The electron density for F_{420} soaked crystals revealed F_{420} bound in two molecules of the same dimer for both F_{420}-Rv2893^{G72S} and F_{420}-Rv2893^{H37Rv} (chains A and B of the four chains in the ASU) (Figure 3.9A). A single monomer (chain D) in the other dimer had unexpected electron density that was neither protein, F_{420}, nor water, in the active site region (Figure 3.9B). Considering the buffer and crystallisation conditions, it is not possible to guess the identity of the ligand that occupies this density. The arrangement of the sidechains of His48, His109 and Glu113 suggest a bound cation with tetrahedral coordination occupies part of this density. The N1 atoms of the imidazole sidechains of His48 and His109 and the carboxylate of Glu113 provide three of the required coordination partners and the fourth would be provided by a non-protein ligand. Mn^{2+} has the required coordination number and the binding site is occupied by amino acid residues that are among the three major ligands for Mn^{2+} in proteins (His, Asp, and Glu) (Zheng *et al.* 2008). Modelling Mn^{2+} into this site shows the expected tetrahedral binding geometry and Mn^{2+} fits the electron density well (Figure 3.9B). Modelling in other cations with a tetrahedral coordination (Co^{2+}, Ni^{2+}, Cu^{2+} and Zn^{2+}) produced poorer electron density maps and a higher R_{free}. Attempts were made to identify the unexpected ligand using *phenix.ligand_identification*, but failed to identify ligands that fit the density appropriately. This tool automates the fitting of the 200 most frequently observed ligands found in the Protein Data Bank (PDB) to a provided electron density map (Terwilliger *et al.* 2007). Model analysis was therefore performed using models with Mn^{2+} included and remainder of the density left unmodeled.

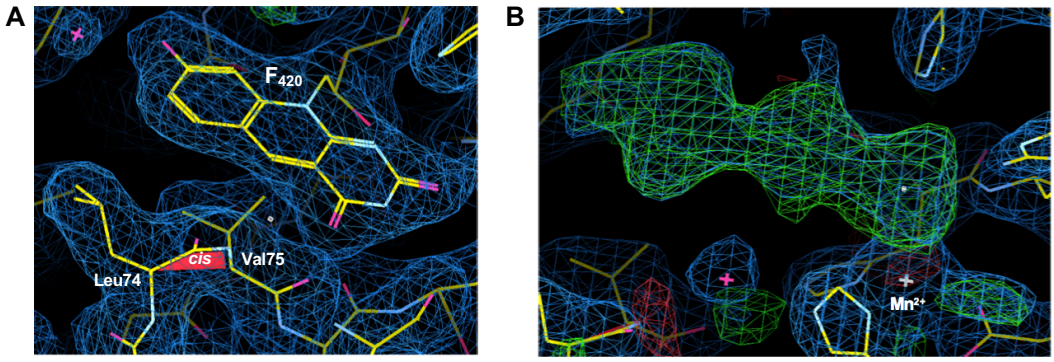


Figure 3.9. *COOT electron density maps for ligand binding in Rv2893.* 2|FO|-|FC| (blue) and |FO|-|FC| (red/green) maps are shown contoured to 1 σ and 3 σ, respectively. (A) Electron density supporting the placement of F_{420}. Density for the non-prolyl *cis*-peptide bond joining Leu74 and Val75 is also shown. (B) Non-protein electron density in the active site of chain D. The Mn^{2+} cation is shown as a grey cross. Density for the F_{420}-Rv2893^{H37Rv} is shown and the same unmodeled density was present in F_{420}-Rv2893^{G72S}.

3.4 Analysis of Rv2893 structures

Each Rv2893 monomer is formed by a single $(\alpha/\beta)_8$ TIM barrel domain and has a surface area of ~13400 Å^2 (Figure 3.10). Two monomers are arranged in an approximately 180 º rotation parallel to the central axis of the TIM-barrel to form the Rv2893 homodimer (Figure 3.11). The TIM barrel α-helixes α_2 and α_3 from each monomer pack closely at the dimer interface and the monomers also associate by interactions involving the α_1 helix and additional loops. The dimer interface buries ~1900 Å^2 (14% of each monomer) and involves four salt bridges and at least 14 hydrogen bonds. The active site is located at the C-terminal end of the TIM barrel (C-terminal end of the β-strands) as in other LLHTs. The substrate binding region has a relatively open conformation and three main accessory loops of 18–34 amino acids, loops β_4-α_4, β_6-α_6 and β_7-α_7, protrude upwards from the C-terminal end of the barrel. A fourth loop comprised of two β-strands (α_4-β_5, 28 amino acids) extends from the N-terminus of the TIM barrel and loops back around to sit adjacent to helix α_4.

A structural similarity search of the PDB using PDBeFold shows that Rv2893 matches most closely with other bacterial LLHTs as expected. The closest structural homologue to Rv2893 is F_{420}-dependent methylenetetrahydromethanopterin reductase (Mer) from the methanogenic archaea *Methanosarcina barkeri* (bMer) (Table 3.6). The top ten matches to Rv2893 included six different F_{420}-dependent LLHTs (F_{420}-LLHTs); three Mer enzymes from different methanogenic archaea (bMer, kMer and tMer); two actinobacterial FGD enzymes from *M. tuberculosis* (mtbFGD) and *Rhodococcus jostii* (rhFGD); and F_{420}-dependent alcohol dehydrogenase (Adf) from the methanoarchaea *Methanoculleus thermophilus*. Superimposition of the three Mer structures show these are all structurally very similar (RMSD 1.1–2.5 Å), as are the two FGD structures (RMSD 1.5 Å). bMer (PDB code 1Z69) and mtbFGD (3B4Y) were therefore selected for comparison with Rv2893, along with Adf (1RHC), as these all have F_{420} bound (Figure 3.12). The four remaining top matches were to three different non-F_{420} dependent bacterial LLHTs (Table 3.6).

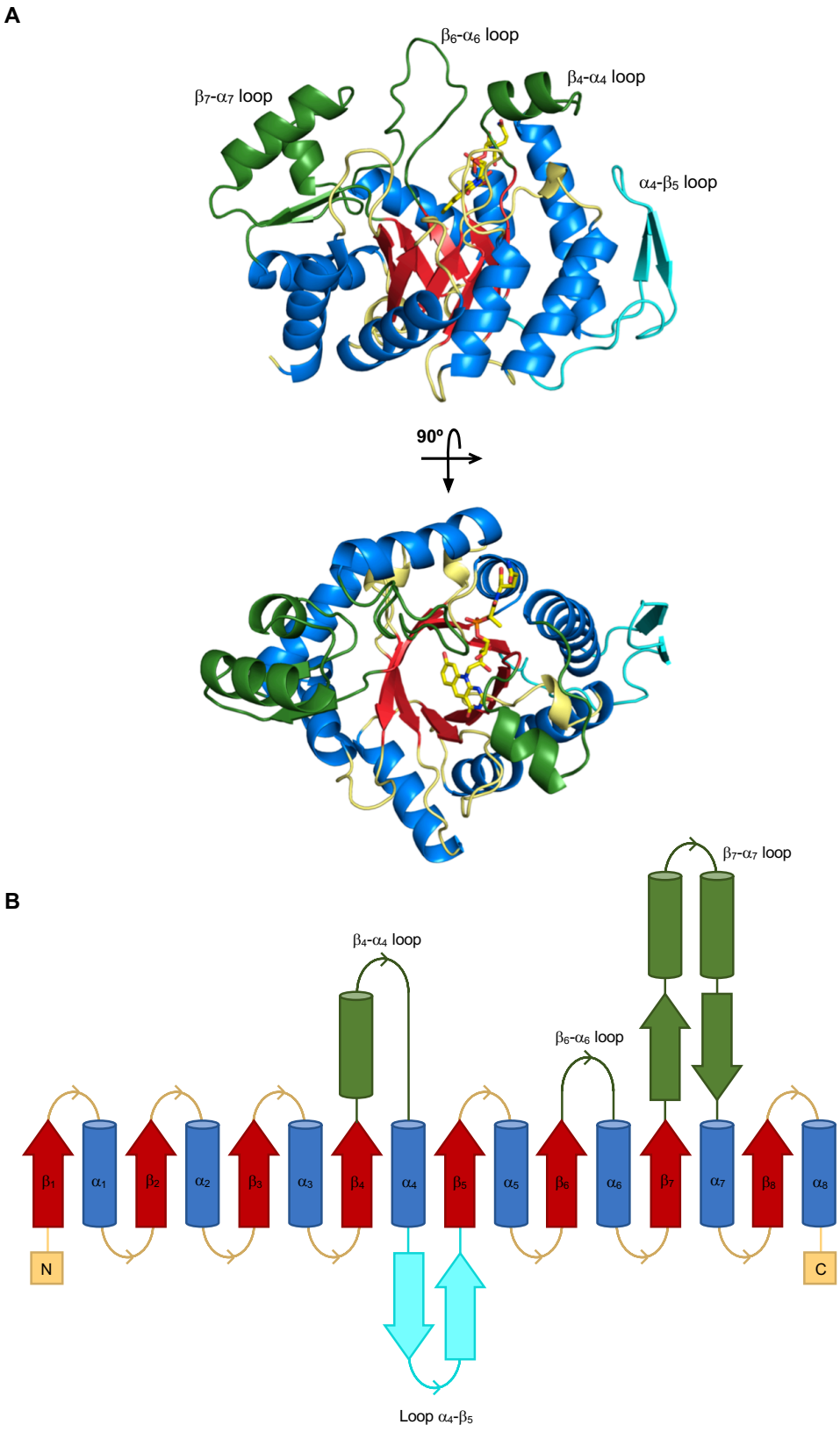


Figure 3.10. *Structure of Rv2893*. (A) Cartoon structure of Rv2893 with F_{420} bound (yellow). Colouring is based on secondary structure; red, TIM barrel β strands; blue, TIM barrel α helixes; pale yellow, short connecting loops (<18 aa); green, large accessory loops at the C-terminal end of the barrel; and cyan, large accessory loops at the N-terminal of the barrel. (B) Schematic depicting the topology of the Rv2893 structure. Colours correspond to the scheme in (A).

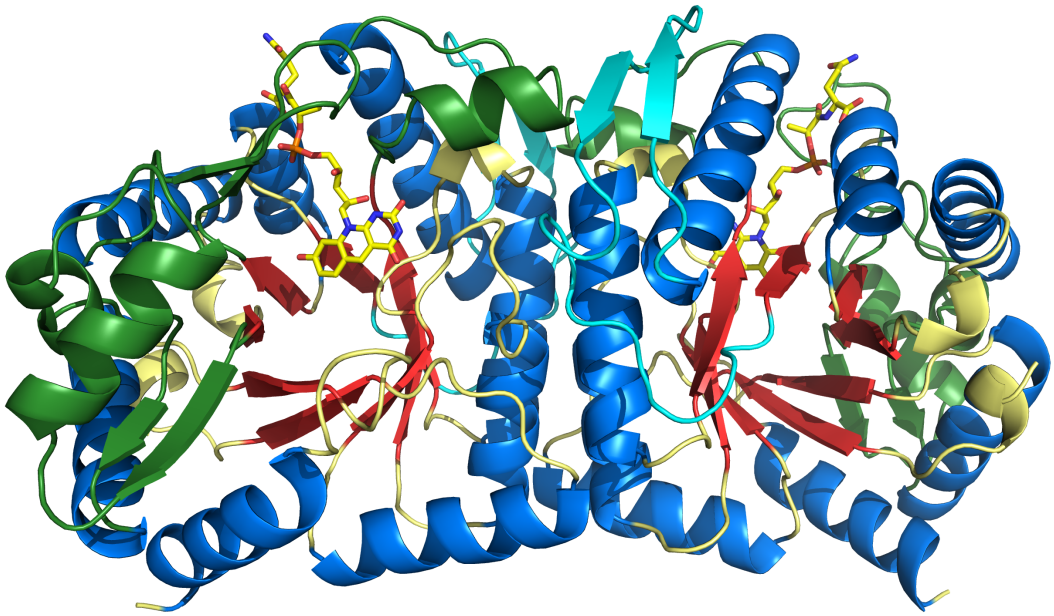


Figure 3.11. *Dimeric arrangement of Rv2893.* F_{420} (yellow) is bound to both chains in the dimer and is depicted as sticks.

A sequence alignment of Rv2893, MSMEG_2516, bMer, mtbFGD, Adf and the top two non-F_{420}-dependent LLHT hits (1M41 and 3RAO), was generated in STRAP using the Aligner3D method to consider both structural and sequence information (Gille and Frömmel 2001) (Figure 3.13). In Rv2893, His48 and His109 correspond to the mtbFGD active site residues His40 and Glu109 (Bashiri *et al.* 2008; Oyugi *et al.* 2016). Examination of the alignment and structures shows 1M41 and 3RAO lack the conserved non-prolyl *cis*-bond found in F_{420}-LLHTs, as well as the catalytic histidine. Examination of the loops and substrate binding region in all four non-F_{420} LLHTs structures showed no obvious commonalities with Rv2893, therefore further analyses focused on the three F_{420}-LLHTs with F_{420} bound.

Table 3.6. *Proteins in the PDB with structural similarity to Rv2893.* The top ten matches identified using PDBeFold are shown. Structures used for main comparisons with Rv2893 are shown in bold and those used included in the alignment in Figure 3.13 are marked with asterisks.

PDB ID	RMSD (Å)	$\%_{sseQ}$[1]	$\%_{sseT}$[1]	Protein	Abbr.[2]	Organism	F_{420}-dependent	F_{420} bound
1Z69*	2.35	88	81	Coenzyme F420-dependent N(5),N(10)-methylenetetrahydromethanopterin reductase	bMer	*Methanosarcina barkeri*	Yes	Yes
1M41*	2.15	79	83	FMNH2-dependent alkanesulfonate monooxygenase	Ssud	*Escherichia coli*	No	N/A
1F07	2.13	75	75	Coenzyme F420-dependent N(5),N(10)-methylenetetrahydromethanopterin reductase	tMer	*Methanobacterium thermoautotrophicum*	Yes	No
3RAO*	2.34	79	83	Putative Luciferase-like Monooxygenase		*Bacillus cereus*	No	N/A
1NQK	2.21	71	81	FMNH2-dependent alkanesulfonate monooxygenase	Ssud	*Escherichia coli*	No	N/A
1RHC*	2.24	83	87	F420-dependent alcohol dehydrogenase	Adf	*Methanoculleus thermophilus*	Yes	Yes
3B4Y*	2.26	71	74	F420-dependent glucose-6-phosphate dehydrogenase 1	mtbFGD	*Mycobacterium tuberculosis*	Yes	Yes
5LXE	2.34	71	74	F420-dependent glucose-6-phosphate dehydrogenase 1	rhFGD	*Rhodococcus jostii*	Yes	No
1EZW	2.47	88	81	Coenzyme F420-dependent N(5),N(10)-methylenetetrahydromethanopterin reductase	kMer	*Methanopyrus kandleri*	Yes	No
1LUC	2.44	71	81	Bacterial Luciferase		*Vibrio harveyi*	No	N/A

[1] $\%_{sse}$ is percent of matched secondary structure of the query chain identified in the target protein ($\%_{sseQ}$) and of target chain matched in the query ($\%_{sseT}$)
[2] Abbr. = abbreviation used in this thesis

Comparison of the overall topology of Rv2893 with other F_{420}-LLHT structures show all structures share the same TIM barrel fold and a similar α_4-β_5 loop at the N-terminal end of the barrel (Figure 3.12). The main differences are in the accessory loops at the C-terminal end of the barrel, particularly in the β_6-α_6 and β_7-α_7 loops. The β_6-α_6 loop is extended by 13–14 residues in Rv2893 compared to the other structures (discussed further in Section 3.4.1), whereas the β_7-α_7 loop is 20–36 residues shorter and does not fold over F_{420} to cap the barrel. Substrate binding cavities were also very different among the F_{420}-LLHTs.

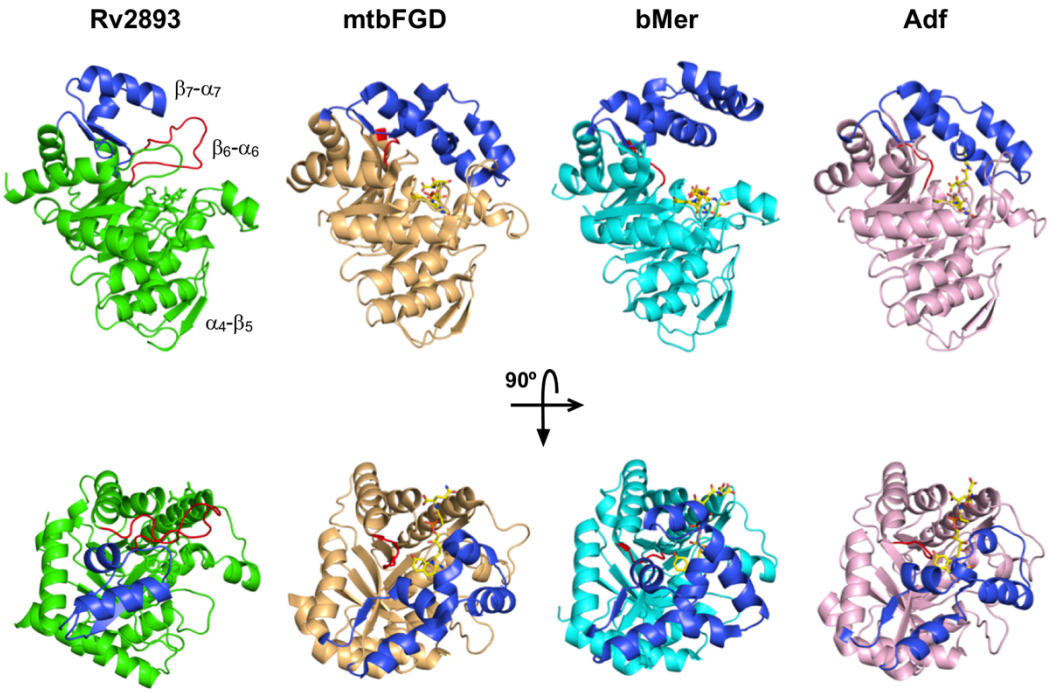


Figure 3.12. *Overall topology of different F_{420}-dependent LLHTs.* Structures of *M. tuberculosis* Rv2893, *M. tuberculosis* FGD (3B4Y), *M. barkeri* Mer (1Z69) and *M. thermophilus* Adf (1RHC) with F_{420} bound are shown. The β_7-α_7 loop is coloured blue, the β_6-α_6 loop red, and F_{420} is shown as yellow sticks. Compared with mtbFGD, bMer and Adf, the Rv2893 β_7 strand is interrupted and instead of forming a single extended strand, a short beta strand at the N-terminal end of the β_7-α_7 loop follows the β_7 strand and is coloured as part of the β_7-α_7 loop in Rv2893.

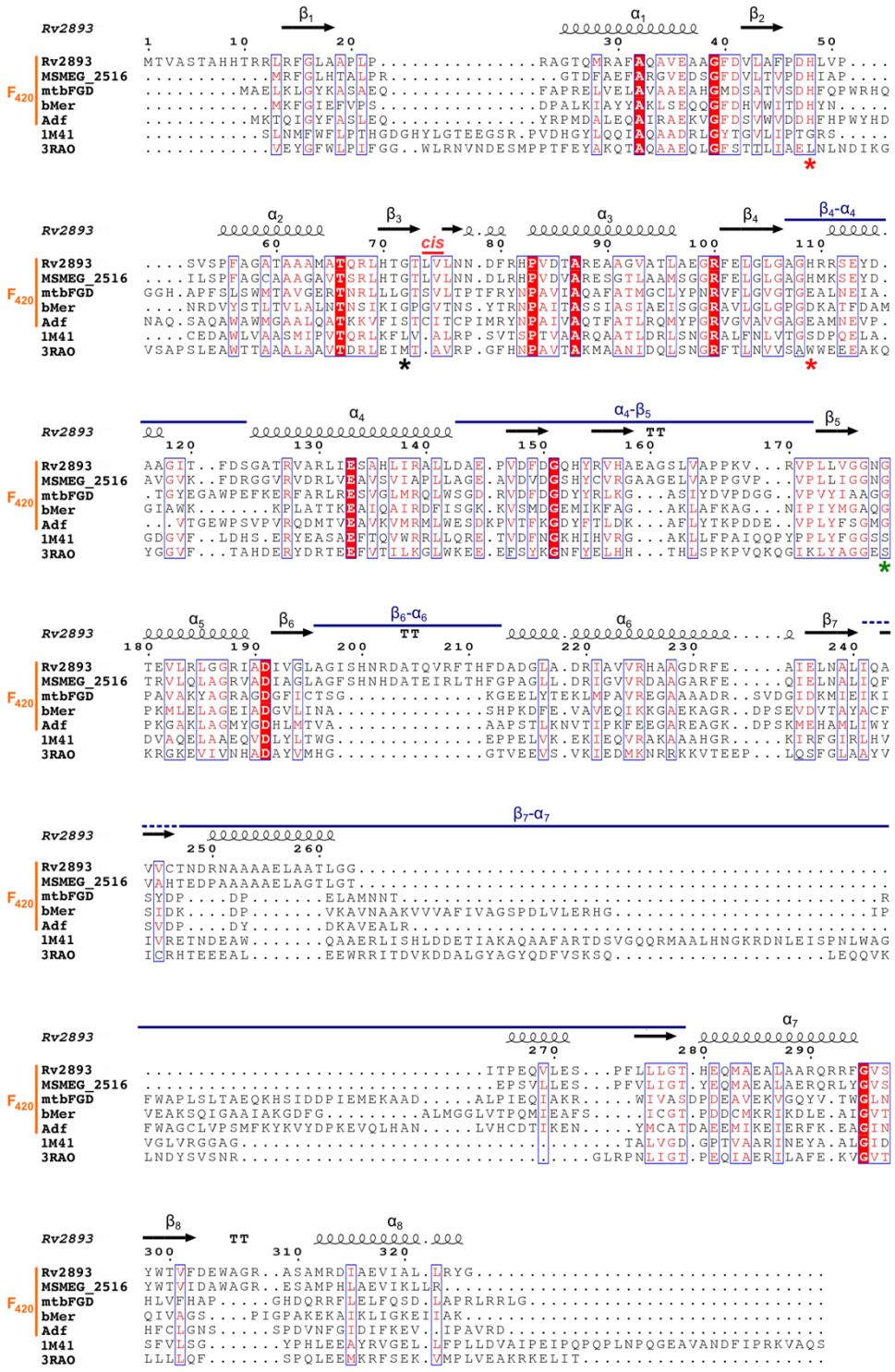


Figure 3.13. *Multiple sequence alignment of LLHT proteins.* Sequences are shown for *M. tuberculosis* Rv2893, *M. smegmatis* MSMEG_2516, and five other LLHTs (Table 3.6). F_{420}-dependent LLHTs are indicated. Secondary structure features for Rv2893 are shown above the sequences and labelled according to Figure 3.10. The non-prolyl *cis*-peptide bond is shown in red and labelled. The start of the β_7-α_7 loop in Rv2893 forms part of an extended β_7 strand in other F_{420}-dependent LLHTs (Figure 3.12) and is indicated by a dashed blue line. Putative active site residues, the Rangipo G72S SNP, and the conserved glycine for F_{420} binding, are indicated below the sequences with red, black and green asterisks, respectively. The alignment was generated in STRAP (Gille and Frömmel 2001) using Aligner3D and the figure drawn using ESPript (Robert and Gouet 2014).

3.4.1 β_6–α_6 loop

Superimposition and sequence alignment of Rv2893 with other F_{420}-LLHTs shows that the region connecting β_6 and α_6 of the TIM barrel – the β_6-α_6 loop, has a 13–14 aa insertion in Rv2893 (Figure 3.13, Figure 3.14A). This loop appears to be a flexible region of the protein and was only able to be fully modelled in three of four chains in the F_{420}-bound structures and for one pair of chains related by pseudo-translational symmetry in apo structures. The Rv2893 β_6-α_6 loop extends from the C-terminal end of the barrel bridging over the tail of F_{420} to contact the β_4-α_4 loop. Three conformational states were modelled for this loop; an 'open' confirmation observed in both F_{420} bound and apo monomers; a more closed state only observed in F_{420} bound monomers (closed-1); and a different closed state in monomers with Mn^{2+} and the unexpected ligand (closed-2) (Figure 3.14B). A fourth conformation also appears to be present in the apo structures, however it was not able to be fully modelled. In the open conformation, the region of the loop between Asp201 and Phe209 extends outwards from the core of the TIM barrel, whereas in the closed-1 form this region is bent ~90 º towards to the core of barrel to sit above His200. This loop is exposed to the substrate binding cleft and is associated with differing positions of His200, which may be important for substrate binding (Section 3.4.3).

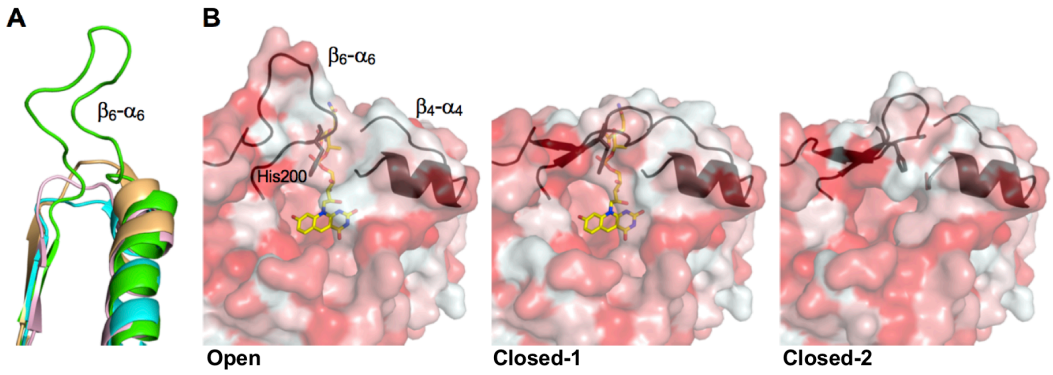


Figure 3.14. *The β_6-α_6 loop.* (A) Structural overlay of the β_6-α_6 loop from Rv2893 (green), mtbFGD (gold), bMer (cyan) and Adf (pink). (B) A surface representation of Rv2893 is shown coloured by hydrophobicity (red, hydrophobic; white hydrophilic) and the underlying β_6–α_6 loop shown as a black cartoon with His200 as sticks. The β_6–α_6 loop was modelled in three different conformations in Rv2893 termed 'open', 'closed-1' and 'closed-2'. The open confirmation is observed in both apo and F_{420} -bound monomers, whereas the closed-1 confirmation was only observed in F_{420} bound monomers, and closed-2 only in monomers with the unknown ligand bound.

3.4.2 F_{420} binding

The apo- and F_{420}-bound Rv2893 structures are highly similar with only minor conformational changes in the main chain (average RMSD 0.72 Å) (Appendix B.4). The most notable difference was observed in the β_6-α_6 loop, as discussed in the previous section. As in other F_{420}-LLHTs, F_{420} is bound at the C-terminal end of the TIM barrel with the isoalloxazine ring positioned innermost to the core of the barrel and the tail projecting outwards (Figure 3.15). The tail of F_{420} binds in a largely negatively charged cleft created between helix α_4, α_5, and the β_4-α_4 and β_5-α_5 loops (Figure 3.15A). The β_6-α_6 loop folds over the cleft but is not within bonding distance of F_{420}. The isoalloxazine ring binds with the *Re*-face adjacent to the non-prolyl *cis*-peptide that connects Leu74 and Val75, as in other F_{420}-LLHTs (Figure 3.15B) (Aufhammer *et al.* 2004; Aufhammer *et al.* 2005; Bashiri *et al.* 2008). Rv2893 has the conserved aspartate on the β_2–α_2 loop (Asp47) found in other F_{420}-LLHTs that is thought to stabilise the bulge of the *cis*-peptide bond (Aufhammer *et al.* 2004). However, unlike the other F_{420}-LLHTs the F_{420} isoalloxazine ring system lacks the pronounced butterfly bend in Rv2893 (Figure 3.15C). In Rv2893, the F_{420} isoalloxazine ring adopts a planar conformation with a bend in the C5 about the pyrimidine and phenol rings of only 7 º.

Rv2893 predominantly contacts F_{420} through hydrogen bonds involving main chain amide and carbonyl groups (Figure 3.15D). The pyrimidine ring of the isoalloxazine ring makes three hydrogen bonds to the peptide backbone, one to the side chain of Asn77, and one to an ordered water. The phenolic ring at the other end is held in a predominantly hydrophobic environment and makes only one weak hydrogen bond with the peptide backbone to a water molecule. The OH groups of the ribitol and phosphate moieties and the first glutamate residue of the tail make several hydrogen bonds to the backbone before the remainder of the tail the extends into the solvent. Upon F_{420} binding the Cα of Asn178 on the β_5–α_5 loop shifts ~2.4 Å away from the F_{420} tail and its side chain flips outwards to accommodate the phosphate group of F_{420}, which then hydrogen bonds the backbone amide of Asn178 and Gly179 (Figure 3.15E).

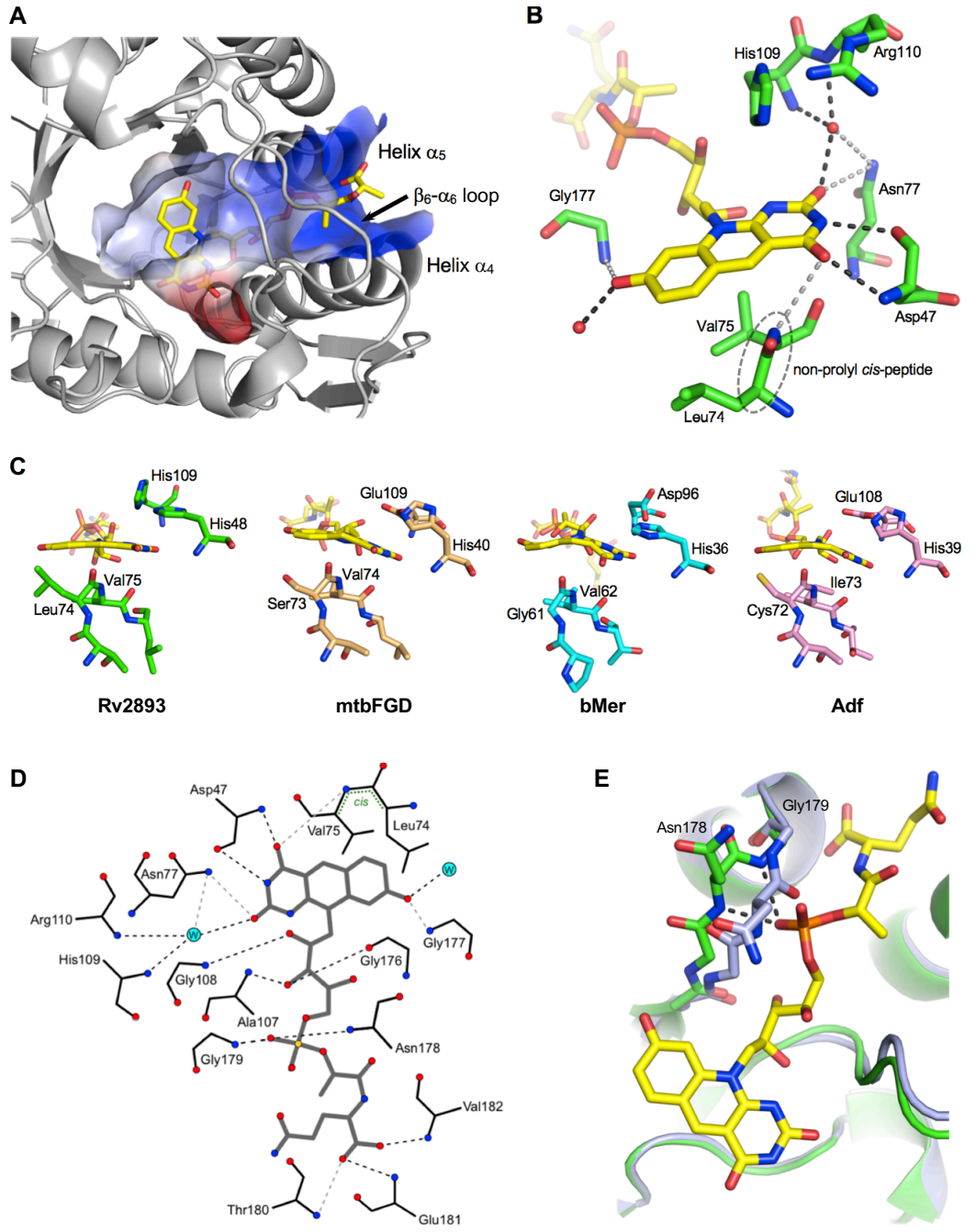


Figure 3.15. *F_{420} binding to Rv2893 and other LLHTs.* (A) Cartoon representation of Rv2893 with a surface charge representation of the F_{420} binding site (surface carve set to 5 Å). Local regions of positive, negative and neutral electrostatic potential are coloured blue, red and white, respectively. (B) The bulge created by the non-prolyl *cis*-peptide bond packs against the *Re*-face of the F_{420} isoalloxazine ring. Hydrogen bonding interactions of the ring are indicated by dark grey (strong bonds) and light grey (weak bond) dashed lines. (C) F_{420} binding in different LLHTs; *M. tuberculosis* Rv2893, *M. tuberculosis* FGD (3B4Y), *M. barkeri* Mer (1Z69) and *M. thermophilus* Adf (1RHC). F_{420} bound to Rv2893 lacks the pronounced butterfly bend apparent in the other structures. (D) Schematic showing hydrogen bonding contacts to F_{420} (bold, dark grey). Oxygen atoms are coloured red, nitrogen atoms blue, and phosphorous gold. Strong hydrogen bonds are shown as black dashed lines and weak bonds as light grey dashed lines. The non-prolyl *cis*-peptide bind between Leu74 and Val75 is indicated. (E) Conformational shift of Asn178 upon F_{420} binding to accommodate the phosphate group. The F_{420} bound conformation is shown in green and non-F_{420} bound in light blue.

3.4.3 Substrate binding region and active site

The substrate binding site is at the base of a deep solvent exposed cleft above the *Si*-face of the F_{420} isoalloxazine ring system and has a predominately hydrophobic character (Figure 3.16). The guanidine group of the Arg110 sidechain sits above the imidazole group of His48 and together with His48 and His109, and His200, form a hydrophilic region above the pyrimidine ring. In mtbFGD and Adf, hydride transfer is mediated by conserved histidine (His40 and His39, respectively) and glutamate residues (Glu109 and Glu108, respectively) (Aufhammer *et al.* 2004; Bashiri *et al.* 2008). Although His40 plays a crucial role in mtbFGD catalysis, it has recently been shown not act as a general base as previously proposed (Oyugi *et al.* 2016). Rv2893 retains the conserved histidine, His48, and this adopts the same confirmation at the *Si*–face of F_{420} as in other F_{420}-LLHTs and is presumably positioned to bond with the substrate and partake in catalysis. His109 occupies the site of the catalytic glutamate that serves as the general acid donating a proton to N1 of F_{420} (Oyugi *et al.* 2016). Depending of its protonation state, His109 could act as either a general base or acid in catalysis. In apo-monomers, the His109 sidechain adopts either the same conformation as in F_{420} bound monomers above the O4 atom of F_{420}, or, is rotated about 90 º in the N-terminal direction and hydrogen bonds with the peptide backbone. A third histidine in Rv2893, His200 on the β_6-α_6 loop, is also positioned in the substrate binding region with the imidazole ring ~5 Å above the F_{420} ribitol moiety in a slightly offset stacked confirmation with His109. In this conformation, the sidechains of His48, His109 and His200 all point into the substrate binding region above the *Si*-face of F_{420}. In Adf and mtbFGD an essential tryptophan residue on the β_2–α_2 loop stabilises the transition state (Aufhammer *et al.* 2004; Bashiri *et al.* 2008; Oyugi *et al.* 2016). Rv2893 lacks this residue and Arg110 on the β_4-α_4 occupies the region of the tryptophan sidechain, however the Arg110 sidechain was not able to be modelled in all chains suggesting it is highly labile.

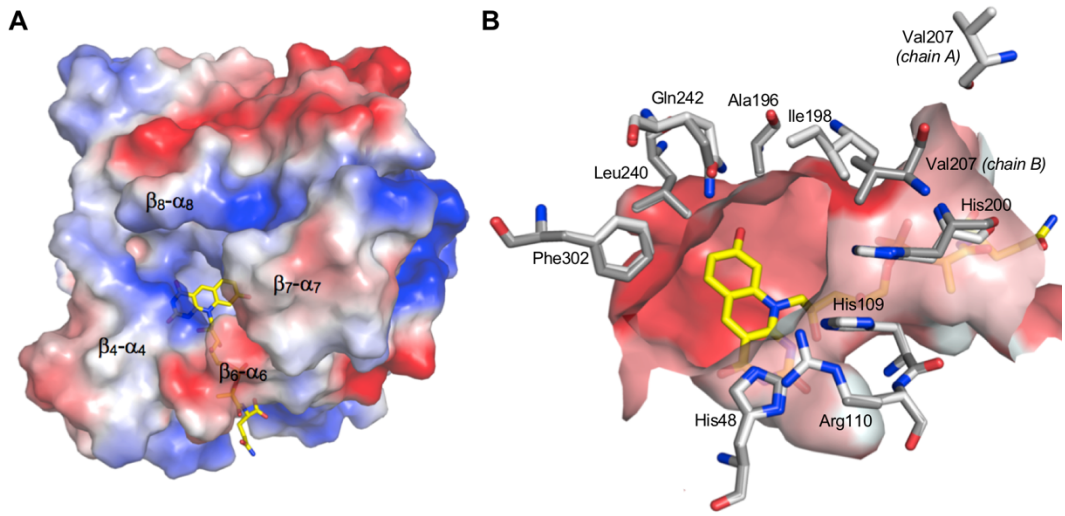


Figure 3.16. *Rv2893 substrate binding cleft*. (A) Surface view of Rv2893 showing F_{420} bound at the base of the cleft. Local regions of positive, negative and neutral electrostatic potential are coloured blue, red and white, respectively. The β_6-α_6 loop is in the closed-1 conformation. (B) Detailed view of the substrate binding region. Sidechain residues that make up the substrate binding region are shown as sticks and a surface representation shows protein surfaces within 6 Å of F_{420} coloured by hydrophobicity (red, hydrophobic; white hydrophilic). Residues for chain A (β_6-α_6 loop in the open conformation) are shown in light grey, and chain B (closed-1 conformation) in a darker grey. The closed-1 conformation brings Val207 into the substrate binding region.

The other end of the substrate binding region above the F_{420} phenol ring is comprised predominately of hydrophobic resides and shares similar characteristics with that of Adf. In Adf, Val193, Leu227 and Trp229 are positioned to interact with both the phenolic end of the F_{420} ring as well as the hydrocarbon end of the substrate. These residue characteristics are conserved in Rv2893, whereby Ala196, Leu240 and Phe302 occupy the corresponding positions. However, in Adf, the β_7-α_7 loop closes over the active site and interacts with β_2-α_2 and β_4-α_4 creating a substrate binding pocket that is completely buried inside the enzyme. These loops adopt a very similar conformation in *M. tuberculosis* FGD, although the mtbFGD substrate binding pocket comprises fewer hydrophobic and more polar residues. In Rv2893, the β_2-α_2 loop is shortened and the β_7-α_7 loop does not close over the core of the barrel, leaving a cleft above the F_{420} *Si*-face exposed to the bulk solvent. This cleft is also lined by predominantly hydrophobic residues and in the β_6-α_6 loop closed-1 conformation found only in F_{420} bound monomers (Section 3.4.1), the hydrophobic sidechain of Val207 also faces into the upper region of the substrate cleft above the F_{420} C5 atom, further contributing to the hydrophobic environment

(Figure 3.16B). Mer also lacks an expanded β_2-α_2 loop and the *Si*-face of F_{420} and substrate binding site are open and largely exposed to the bulk solvent.

Electron density that was neither protein, water, nor F_{420}, was observed in the substrate binding region in a single apo-monomer in both F_{420} bound structures (Figure 3.17). An Mn^{2+} cation was modelled into part of this density, however, despite attempts to identify the ligand using automated modelling of a ligand library in PHENIX and manually modelling numerous different ligands, its identity currently remains unknown. Binding of this ligand disrupts the arrangement of His48, His109 and His200 (Figure 3.17A) and the β_6-α_6 loop adopts the closed-2 conformation. The His48 sidechain is rotated ~25 º away from the F_{420} pyrimidine ring and the Mn^{2+} cation occupies the site between the imidazole rings of His48, His109 and the carboxylate of Glu113 on the β_4-α_4 loop. The imidazole ring of His109 is shifted into the region occupied by the O2 and N3 atoms of F_{420} and the other end of the ligand is positioned where the hydroxyl group of the F_{420} phenolic ring binds. This would prevent F_{420} binding, therefore the bound ligand is unlikely to be a substrate of Rv2893 but rather may represent a competitive inhibitor or regulator.

The 'head' of the ligand is positioned adjacent to the cation and the shape of the electron density in this region is highly suggestive of a phosphate group. Library ligands modelled by PHENIX frequently placed a phosphate group adjacent to the cation and the shape of the density is appropriate for either a phosphate or sulphate group. Modelling PO_4^- in provides a good fit to the density and provides the appropriate tetrahedral geometry for the coordination of Mn^{2+}, as well making hydrogen bonds to the His200 and Arg110 side chains (Figure 3.17B). The unmodeled density extends ~11 Å from the phosphorous atom into the core of TIM barrel. Both the central region and tail end of the ligand are planar and the two planes are offset by an approximately 120 º rotation. The shape of the density and modelling in various cyclic compounds suggests at these regions might be occupied by monocyclic rings.

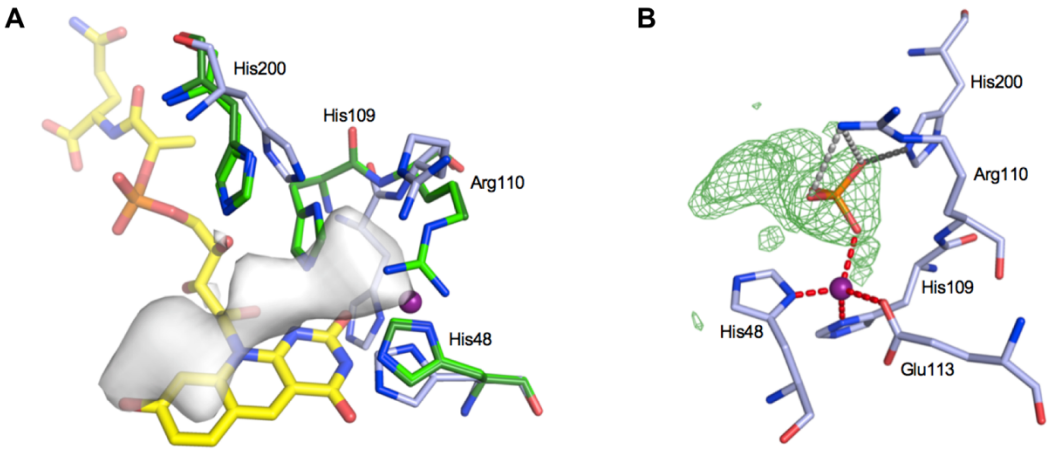


Figure 3.17. *The unidentified ligand bound in the Rv2893 active site*. (A) Arrangement of active site residues His48 and His109, and His200, in the substrate binding region and binding of the unexpected ligand (semi-transparent white blob). F_{420} is not found in monomers with the ligand bound (light blue sticks). F_{420}-bound monomers with the β_6-α_6 loop in the open and closed-1 conformations are shown as green and dark green sticks, respectively. Binding of the unexpected ligand disrupts the confirmation of the active site residues and the tail of the ligand occupies the region in which the hydroxyl of the F_{420} ring binds. (B) Phosphate modelled into the head end of the density for the unidentified ligand. Coordination of Mn^{2+} to His48, His109 and Glu113 is indicated by dashed red lines. Hydrogen bonds made by the phosphate are indicated with dark grey dashed lines for strong bonds and light grey dashed lines for weak bonds.

3.4.4 The Rangipo specific SNP encoding a G72S mutation

The Rangipo specific G72S mutation site is positioned at the base of the bulge created by the non-prolyl *cis*-peptide bond (Figure 3.19). No major global or local topology changes were observed between the wild type $Rv2893^{H37Rv}$ and Rangipo variant $Rv2893^{G72S}$ structures (average RMSD between apo monomers 0.44 Å).

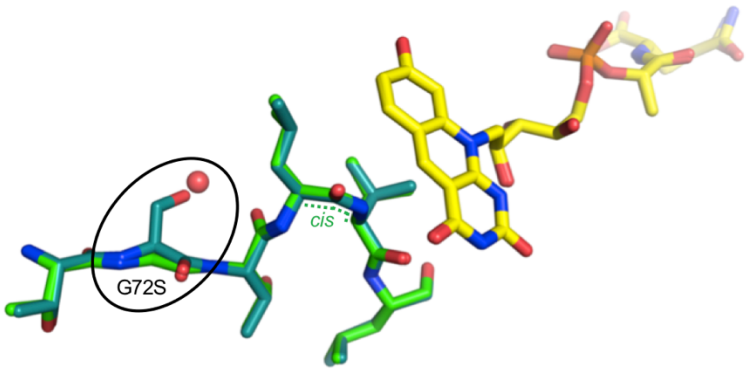


Figure 3.18. *Rangipo specific G72S SNP*. Superimposition of F_{420}-$Rv2893^{H37Rv}$ (green) and F_{420}-$Rv2893^{G72S}$ (dark teal) with the SNP site and non-prolyl cis-peptide bond indicated. In F_{420}-$Rv2893^{H37Rv}$, an ordered water molecule (red sphere) occupies a similar position as the Ser72 hydroxyl group. No other local or global rearrangements were observed.

The non-prolyl *cis*-peptide bond is conserved in other F_{420}-LLHTS and is important for F_{420} binding and catalysis as it creates a bulge that packs against the *Re*-face of F_{420}, helping induce a butterfly bend in the F_{420} isoalloxazine ring system increasing its reactivity (Bashiri *et al.* 2008). Although the mutation is not in a position to directly interact with F_{420} or the substrate, its location at the base of this bulge is a prime position for having a functional effect via interaction with the bulge. Examination of the region and adjacent β-sheets shows that the G72S mutation creates two additional hydrogen bonds between the β_3 and β_4 strands (Figure 3.19B) which may explain the increased thermal stability observed for the G72S variant compared to the wild-type (Section 3.3.3). In Rv2893^{G72S}, the OH group of Ser72 makes hydrogen bonds to the backbone carbonyl groups of Thr73 on β_3 and Glu102 on β_4, as well as with another ordered water that also hydrogen bonds to the carboxylate group of Glu102. In Rv2893^{H37Rv}, these intramolecular hydrogen bonds are lost and an ordered water sits between the two β-strands in the site occupied by the Ser72 hydroxyl and makes hydrogen bonds to the carbonyl groups of Thr73 and Glu102, and to the other ordered water.

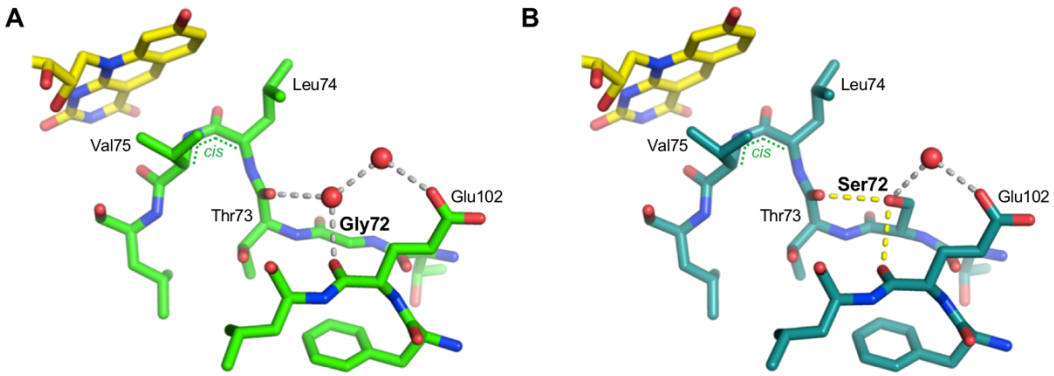


Figure 3.19. *Structural location of the Rangipo specific G72S SNP.* Arrangement of the residues surrounding the SNP position and non-prolyl *cis*-peptide bond in (A) Rv2893^{H37Rv} and (B) Rv2893^{G72S}. Ordered water molecules are shown as red spheres and F_{420} as yellow sticks. Intra-molecular hydrogen bonds are indicated by yellow dashed lines, and hydrogen bonds to water molecules by light grey dashed lines. Hydrogen bond distances were all within 2.7–3.1 Å and the non-prolyl *cis*-peptide bond is indicated.

3.5 Discussion

Members of the *M. tuberculosis* complex (MTBC) are genetically more diverse than traditionally assumed and increasing evidence suggests strain genetic background impacts virulence phenotypes and can have important clinical and epidemiological consequences (Section 1.4.1). Understanding the molecular mechanisms underlying this variation is an essential step towards understanding the complex relationship between bacterial genotype, virulence, and clinical outcomes. SNPs are one of the most common forms of genetic variation in the MTBC and have the potential to alter bacterial phenotype if they change the amino sequence of encoded proteins or alter transcriptional regulation.

The Rangipo strain has been the source of numerous tuberculosis outbreaks in New Zealand over the past 30 years and has been suggested to be highly virulent relative to other circulating strains (Section 1.1.3). Characterising the consequences of genetic variation unique to this strain may help understand factors contributing to its success. Analysis of whole genome SNP data from the New Zealand Rangipo *M. tuberculosis* cluster identified several Rangipo specific nsSNPs predicted to affect protein function (Section 2.2.2.2) and thereby may affect bacterial phenotype. The putative F_{420}-dependent oxidoreductase Rv2893 harbours a G72S mutation that I have identified as specific to the Rangipo strain (nsSNP) and is predicted to effect protein function *in silico* (Section 2.2.2.4). This SNP was selected for further investigation based on the biological roles of its predicted cofactor F_{420} in *M. tuberculosis*, as well as for practical considerations. F_{420} is an unusual low-redox potential cofactor that enhances the metabolic flexibility of mycobacteria and is important for survival in challenging environments (Gurumurthy *et al.* 2013; Hasan *et al.* 2010; Jirapanjawat *et al.* 2016; Purwantini and Mukhopadhyay 2009). However, the mechanisms by which F_{420} contributes to this and its roles in *M. tuberculosis* metabolism remain poorly understood.

Crystal structures of Rv2893 were solved at 2.12–2.31 Å for the *M. tuberculosis* reference strain H37Rv protein and the Rangipo G72S variant in the apo-form and with F_{420} bound, confirming this protein is indeed an F_{420}-dependent LLHT. Despite being expressed in *M. smegmatis* cells, which produce F_{420}, Rv2893 purified without F_{420} bound and it had to be soaked into Rv2893 crystals to obtain the liganded structure. The same scenario was reported for mtbFGD expressed in

M. smegmatis and its F_{420}-bound structure was also obtained by soaking crystals in F_{420} (Bashiri *et al.* 2008). Conversely, *M. tuberculosis* F_{420}-reducing hydroxymycolic acid reductase (fHMAD, Rv0132c) (Purwantini and Mukhopadhyay 2013) (also an LLHT family protein), has also been expressed in *M. smegmatis* and was found to co-purify with F_{420} (Bashiri *et al.* 2012). These are the only other published *M. tuberculosis* F_{420}-dependent LLHTs that have been purified and characterised, although the F_{420}-dependency of an additional mycobacterial LLHT – phthiodiolone ketoreductase (fPKR) (Rv2951c), has also been experimentally validated (Purwantini *et al.* 2016). Rv2893 therefore increases the number of experientially confirmed F_{420}-binding LLHTs in *M. tuberculosis* to four.

There are only three other F_{420}-dependent LLHT structures complexed with F_{420} deposited in the PDB; mtbFGD (3B4Y), Adf (1RHC) and bMer (1Z69), and a further three without F_{420} bound; rhFGD (5LXE), kMer (1EZW) and tMer (1F07). The overall binding mode of F_{420} in Rv2893 is the same as in mtbFGD, bMer and Adf with the F_{420} isoalloxazine ring bound innermost and its *Si*-face exposed to the substrate binding region. However, in Rv2893 the F_{420} isoalloxazine ring lacks the pronounced butterfly bend apparent in the other structures, despite conservation of the non-prolyl *cis*-peptide bond. Non-prolyl *cis*-peptide bonds are energetically unfavourable features rarely found in protein structures and tend to occur in important regions, such as next to an active site (Jabs *et al.* 1999). In F_{420}-LLHTs the non-prolyl *cis*-peptide bond is important for F_{420} binding and catalysis as it creates a bulge that packs against the *Re*-face of the isoalloxazine ring, inducing the formation and stabilisation of a butterfly bend confirmation about the C5 atom. This bent conformation is energetically unfavourable for oxidised F_{420} and is important for catalysis as in this conformation the C5 carbon adopts a more tetrahedral sp^3-like geometry, increasing its reactivity (Aufhammer *et al.* 2004; Bashiri *et al.* 2008). This pronounced bend was not observed in Rv2893, and it is speculated that substrate binding may be required to induce this more reactive confirmation in Rv2893. In both Adf and bMer, protein residues pack against the phenol ring of F_{420} inducing the bent conformation, and in FGD1 rather than protein residues, the inhibitor citrate packs against F_{420} ring and promotes the bend (Bashiri *et al.* 2008). Rv2893 lacks packing above the phenol ring either from substrate or protein

residues and substrate binding may therefore be required to promote the butterfly bend and assist catalysis.

The Rangipo G72S mutation is located adjacent to the non-prolyl *cis*-peptide and creates two additional hydrogen bonds between the β_3 and β_4 strands. Melting temperature analysis by DSF indicated that the Rangipo Rv2893^{G72S} variant has higher thermal stability than the wild-type Rv2893^{H37Rv} protein, which may be explained by these additional intra-molecular bonds. Although there is no simple relationship between changes in protein stability and protein functional changes, mutations that change stability frequently affect protein function (Bromberg and Rost 2009). Changes in stability alter the ratio of folded to unfolded molecules; high stability results in more folded molecules, whereas low stability results in a higher proportion of unfolded molecules and increases the propensity for aggregation. Changes in stability can have important phenotypic consequences, in humans for example, the majority of disease associated single nucleotide mutations affect protein stability, although typically these are destabilising mutations (Wang and Moult 2001). While the effects of destabilising mutations causing protein aggregation and decreasing the pool of active molecules are readily appreciated, changes that increase stability can also have marked effects. As well as increasing the ratio of folded to unfolded molecules, highly thermostable proteins are also more resistant to proteolysis (Daniel *et al.* 1982). Most notably, increasing stability can reduce flexibility, a critical feature of protein catalysis. There is a fine balance between protein stability and flexibility, whereby proteins must be rigid enough to retain confirmations required for function, but also still labile enough to allow structural changes required for substrate binding and catalysis. Thus, increased stability can adversely affect activity if it inhibits required conformational changes. No notable changes were apparent in F_{420} binding between Rv2893^{H37Rv} and Rv2893^{G72S} in the crystal structures, however, in biological reality proteins are not rigid molecules as portrayed in structural models. Thus, considering the location of the G72S mutation at the base of the *cis*-peptide bulge and the important role this feature plays in F_{420} binding and catalysis, concomitant with its effects on overall stability, it is speculated that this mutation will likely influence F_{420} binding affinity or catalysis in solution. Further work needs to be undertaken to characterize the binding kinetics of F_{420} to Rv2893^{H37Rv} and Rv2893^{G72S} to determine this.

Whether additional stability imposed by the G72S SNP impairs or enhances Rv2893 enzyme activity, its consequences on bacterial phenotype are likely to be largely dictated by its functional role in *M. tuberculosis*. Rv2893 retains the conserved catalytic histidine (His48) and another histidine (His109), occupies the site of catalytic acid found in other F_{420}-LLHTs. Depending on its protonation state, His109 could act as either a general base or acid in acid-base catalysis and Rv2893 therefore retains the necessary active site chemistry that would enable it to catalyse hydride transfer. It remains unclear whether Rv2893 is an F_{420}-reducing dehydrogenase or an $F_{420}H_2$-reductase. FGD is the main source of $F_{420}H_2$ in *M. tuberculosis*. This apparent from the phenotypic effects produced by disrupting the gene encoding FGD, which reflect those produced by eliminating F_{420} biosynthesis (Jirapanjawat *et al.* 2016; Purwantini and Mukhopadhyay 2009). fHMAD is also an LLHT family F_{420}-reducing dehydrogenase (Purwantini and Mukhopadhyay 2013). The majority of other F_{420}-dependent LLHTs however are expected to be reductases (Greening *et al.* 2016). It therefore seems more likely Rv2893 is a $F_{420}H_2$-reductase than a F_{420}-reducing dehydrogenase.

Rv2893 shares structural homology with other microbial F_{420}-LLHTs; Mer, Adf and FGD, as well as non-F_{420}-dependent LLHTs. These act on a diverse range of different substrates and have a broad range of roles in different organisms. FGD enzymes oxidise glucose-6-phosphate to 6-phospho-gluconolactone producing $F_{420}H_2$ which is then used by F_{420}-dependent reductases. Mer is involved in CO_2 reduction in methanogenic archaea and catalyses the reduction of N^5,N^{10}-methylene-tetrahydromethanopterin to N^5-methyl-tetrahydromethanopterin, with concomitant oxidation of $F_{420}H_2$ to F_{420}. Adf is a secondary alcohol dehydrogenase also found in methanogenic archaea and enables isopropanol to be used a hydrogen donor for CO_2 reduction, reducing F_{420} in the process (Widdel and Wolfe 1989). Rv2893 is well suited to bind a hydrophobic substrate and shares notable similarities with Adf in terms of the hydrophobic characteristics of the residues comprising the substrate binding site. However, unlike Adf, the Rv2893 substrate binding region has a more open character which would allow for binding of a larger substrate. The Mer substrate, methylene-tetrahydromethanopterin, is also much bulkier than those of FGD and Adf, and accordingly Mer has a relatively open substrate binding cleft rather than a tight substrate binding pocket.

Interestingly, F_{420}-bound structures were also found to be complexed with a currently unidentified ligand. This ligand represents a competitive inhibitor or regulator as the structural changes it induces would be inhibitory for F_{420} binding and catalysis. The head of the ligand is suggested to be a phosphate group and the remainder appears to comprised of one or more monocyclic rings, likely with a hydrophobic nature. As it was only bound in monomers unoccupied by F_{420} its binding is not dependent on F_{420} binding, so it is curious that it was not bound in the apo structures. Its presence could potentially be the result of an impurity in the F_{420} purification introduced during soaks. Alternatively, both apo, but not F_{420} bound structures, were solved from crystallised protein that had been purified with DDM in lysis and IMAC buffers. This detergent may have stripped away the ligand early on the purification. Identification of this ligand will likely provide additional clues as to the function of Rv2893 and further work needs to be undertaken to determine its identity.

3.5.1 Conclusions

The structure of the *M. tuberculosis* putative F_{420}-dependent oxidoreductase Rv2893 has been solved showing it shares the same overall TIM barrel fold as other bacterial LLHTs and validating it as an F_{420}-binding protein. Only three other F_{420}-LLHT structures complexed with cofactor F_{420} are currently available and this structure represents just the second F_{420}-LLHT from *M. tuberculosis.* The Rv2893 active site region and F_{420} binding mode share conserved features found in other F_{420}-LLHTs important for F_{420} binding and enzyme catalysis. However, in Rv2893 F_{420} lacks the pronounced butterfly bend and substrate binding may be required to induce this more reactive confirmation. The Rangipo specific G72S mutation is located at the base of the bulge created by the non-prolyl *cis*-bond important for F_{420} binding and catalysis. This SNP creates additional intramolecular hydrogen bonds and increases the thermal stability of Rv2893, which may have important consequences for F_{420} binding and/or enzyme activity. The biological consequences of this SNP will largely depend on the function of this protein in *M. tuberculosis* and further work will seek to elucidate its role. This will provide new insights regarding the role of F_{420} in *M. tuberculosis* and enable further investigation into the functional consequences of the Rangipo G72S SNP.

Chapter Four

Evolutionary analysis of the Rangipo and Otara tuberculosis clusters

4.1 Introduction

Modern tuberculosis strain diversity in New Zealand Europeans, Māori and Pacific People is dominated by L4/Euro-American strains (Yen *et al.* 2013). Although the annual incidence rate of TB notifications in New Zealand is comparatively low at around 6.5 cases per 100,000 population, the disease disproportionally affects Māori and Pacific People (ESR 2018). Around three-quarters of *Mycobacterium tuberculosis* isolates from Māori and Pacific People have non-unique MIRU-VNTR molecular typing patterns and this is typically interpreted to indicate recent transmission of strains. The two largest *M. tuberculosis* clusters in New Zealand – the Rangipo and Southern Cross clusters, are most prevalent in Māori and Pacific People, respectively. The Otara cluster is another large cluster predominantly found in Pacific People. The Rangipo cluster is of particular public health concern due to it being a source of several outbreaks over the last thirty years and it has also been suggested to be highly virulent (Colangeli *et al.* 2014; De Zoysa *et al.* 2001; McElnay *et al.* 2004).

One of the main objectives of this chapter was to trace the historical origins and dispersal of the Rangipo and Otara clusters. Both of these clusters belong to the L4.4 sublineage, which is the most prevalent L4/Euro-American sublineage in New Zealand born tuberculosis cases (43% of L4 cases) (Stucki *et al.* 2016b). Oceania was the last major global region to be reached by Europeans and it is generally assumed that tuberculosis was absent from New Zealand prior to European arrival (Durie 1998). When considering tuberculosis in New Zealand it is important to

understand the migratory and colonial history of New Zealand and the wider South Pacific.

4.1.1 Historical and demographic context

Oceania is geographic region in the Pacific Ocean comprised of Australia, Melanesia, Micronesia and Polynesia. The Polynesian sub-region is made up of over one thousand islands scattered across the central and southern Pacific Ocean, extending from Hawaii to the north to New Zealand and Easter Island to the south and east (Figure 4.1). New Zealand is home to both the indigenous Māori population and the largest diaspora communities of indigenous Polynesian people (Spickard *et al.* 2002), offering a unique setting for the investigation of *M. tuberculosis* dispersal and transmission in the South Pacific.

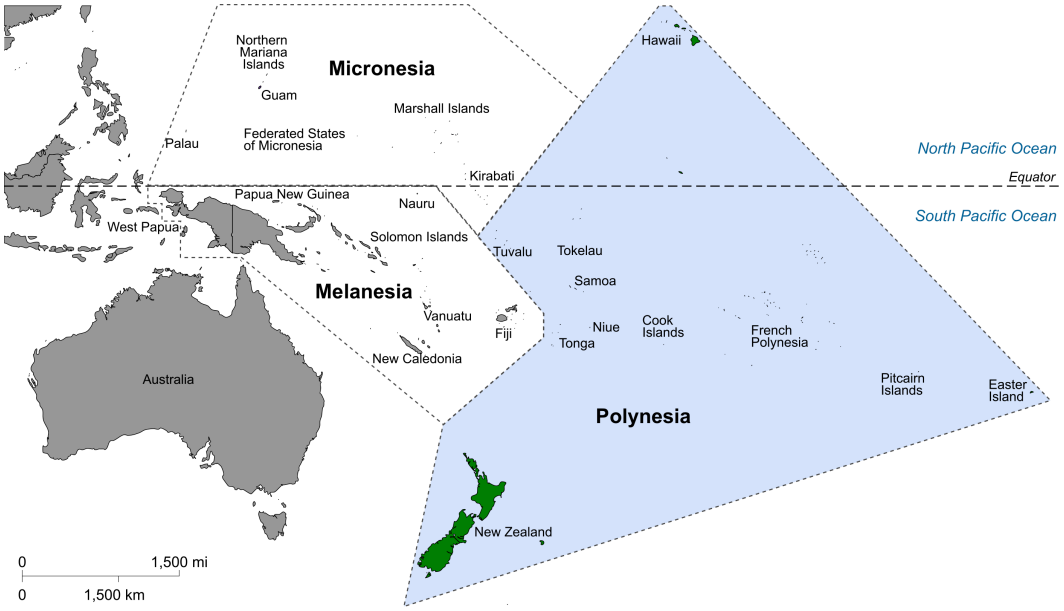


Figure 4.1. *Map of Oceania highlighting the Polynesian sub-region (blue).* (Modified from https://commons.wikimedia.org/wiki/File:Oceania_UN_Geoscheme_Regions_with_Zones_and_ISO3166_labels.svg)

Māori make up 15% of the New Zealand population and Pacific People account for 7% (StatsNZ 2013). The majority of New Zealand's Pacific island population identify with ethnic groups from Polynesia (excluding Hawaii), and Fiji is the only notable contributor from outside this region (4.9% of Pacific People identify as Fijian). The three major Pacific ethnic groups represented in New Zealand are Samoan, Cook Island Māori and Tongan, and 90% of Pacific People in New Zealand identify with one or more of these groups (StatsNZ 2013).

4.1.1.1 Early European presence in Polynesia

The islands of Polynesia have been inhabited by humans for around three thousand years, whereas Europeans have only been part of the history of this region for approximately the last five hundred years. The early-1500s to late-1700s were an era of great European exploration in the South Pacific, dominated by the Spanish in the 1500s, the Dutch in the 1600s, and then the British and French from the mid-1700s. These early voyages commonly involved contact and trade with local communities, however, early contact between Europeans and indigenous Pacific peoples was both irregular and brief. The great European exploration and scientific voyages ceased from the late 1700s and the European presence in the Pacific began to be dominated by whalers, sealers and other traders. This led to the establishment of more regular and intensified contact between Pacific Islanders and Europeans and had much greater lasting impacts on Pacific societies than the previous passing visits of European explorers. The whaling industry had a particularly profound effect on indigenous Pacific Island populations and was a major driving force for European colonisation. From 1840, European colonies and protectorates were established, however with the exception of New Zealand (Section 4.1.1.2), the islands of Polynesia did not receive mass influxes of European migrants. The increasing volume of European ships passing through the Southern Pacific also promoted the growth of trading settlements and larger commerce hubs. Pacific societies changed significantly as people relocated to harbours for work and trade, and local economies shifted to provide goods for trade. (Campbell 2011; Fischer 2013).

4.1.1.2 European discovery and early settlement of New Zealand

The Polynesian ancestors of the Māori arrived in New Zealand around 1230–1280 AD during an era of widespread Polynesian voyaging (King 2004; Wilmshurst *et al.* 2011). The era of long-distance Polynesian voyaging rapidly declined around 1450 AD (Rolett 2002) and Māori were isolated from island Polynesia and the rest of the world for several hundred years until the arrival of Europeans. The first European explorer to reach New Zealand was the Dutch explorer Abel Tasman in 1642, however it was not until 1769 that the first European set foot in New Zealand, Englishman Captain James Cook. Cook met with Māori on numerous occasions and the consequences of these encounters have been described as "far-reaching, although they did not change the cultural pattern or the quality of day-to-day Māori life" (King 2004). The French became interested in the Pacific around the same time as the English. French explorers were the only other Europeans who played significant role in mapping the New Zealand coastline, and they also interacted with local Māori during visits to New Zealand (Wilson 2016). The first Europeans to live in New Zealand were predominantly sealers, whalers, ship deserters or convicts, many of whom were of English or Irish descent (Phillips 2015). European migration initially occurred at low levels, just over 300 Europeans had settled in New Zealand by 1830 and this increased to around 2,000 by 1840 (King 2004). From 1840 the rate of immigration increased significantly and by the end of the 19th century the European population numbered nearly 800,000 (Orange 2012) (Figure 4.2). The majority of Europeans who moved to New Zealand were from Britain and Ireland, with only minor contributions from other countries including a small number of French who settled in Akaroa (fewer than 100), and 281 Germans who settled near Nelson (Phillips 2015).

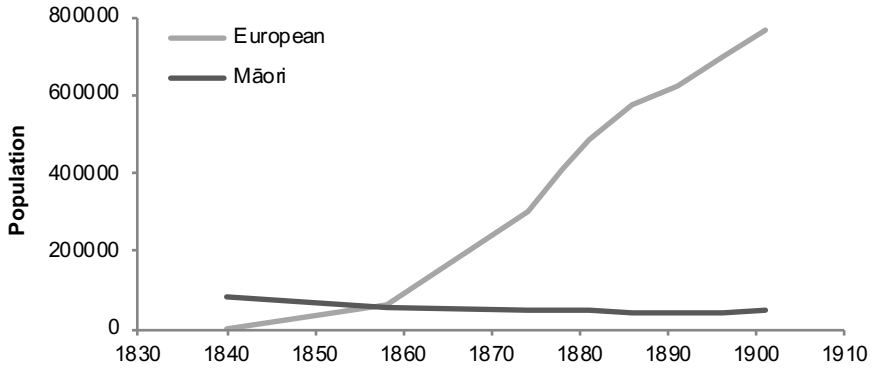


Figure 4.2. *Māori and European population in New Zealand, 1840–1901.* Source: Orange (2012).

4.1.1.3 New Zealand and the Pacific Islands

From the early 1800s only very small numbers of Pacific people began to settle in New Zealand and by 1916 when the first census data counting Pacific ethnicities is available there were only 151 Pacific Island Polynesians in New Zealand (StatsNZ 1916). The Pacific population began to slowly increase in the first half of the 20th century until a surge in migration in the 1950s to early 1970s with most migration coming from Samoa, the Cook Islands, Tonga, Niue, Tokelau and Fiji (Dunsford *et al.* 2011). Between 1951 and 1976 the Pacific Island Polynesian population increased from 3,624 to 61,354 (StatsNZ 1996) (Figure 4.3) and there are now approximately three hundred thousand Pacific people in New Zealand, comprising seven percent of the population (StatsNZ 2013).

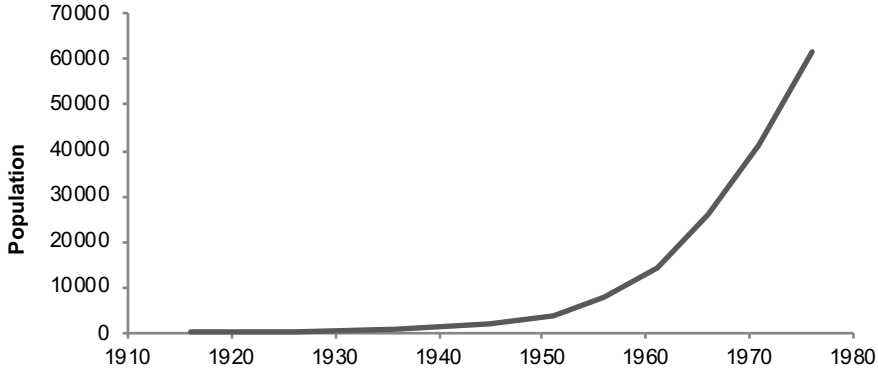


Figure 4.3. *Pacific Island Polynesian population in New Zealand, 1916–1976.* Source: (StatsNZ 1996).

4.1.1.4 A brief history of tuberculosis in New Zealand

Tuberculosis is generally considered to have been absent from New Zealand prior to European arrival (Durie 1998). The disease may have been introduced as early as the arrival of English captain James Cook in 1769, as crew suffering from tuberculosis were present on the Endeavour (Gluckman 1976). Māori deaths from tuberculosis following voyages to England have been recorded as early as 1807 and by the 1830s high tuberculosis rates amongst Māori were becoming apparent (Gluckman 1976). In the first two decades of the 20th century public health efforts and discussion around tuberculosis in New Zealand focused on Europeans and the tuberculosis problem in Māori was largely ignored (Dow 1999; Durie 1998). By the 1920s, the threat of tuberculosis to Māori was well known, however, the full extent of the problem was not acknowledged until the 1930s (Dow 1999). Although European tuberculosis morality rates have been recorded since the 1870s, the first reliable records for Māori mortality rates were not collected until the 1930s and show rates as high as 407.8 per 100,000 population (MacLean 1964). Following intensified public health efforts in the 1940–1950s, there was a marked decline in Māori tuberculosis notification rates (Figure 4.4) and by the 1970s tuberculosis was considered to have been bought under control (Durie 1998). Māori rates however, were still around five times higher than in Europeans, a trend that continues to persist today.

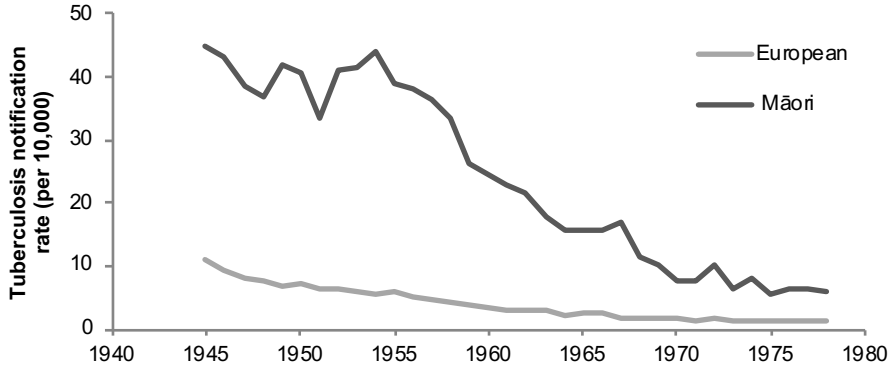


Figure 4.4. *New Zealand tuberculosis notification rates for European and Māori, 1945–1978.* Source: Appendices to the Journals of the House of Representatives (AJHR), H-31, 1951-97 in Dunsford (2008).

In 2015, the European and Other ethnic group had a tuberculosis notification rate of 0.6/100,000, whereas the rate in Māori was 3.2/100,000, and in Pacific People 20.2/100,000 (ESR 2018). Furthermore, around three-quarters of *M. tuberculosis* isolates from Māori and Pacific People have non-unique molecular types (i.e. are clustered) compared to approximately one-third in the European or Other group (ESR 2018). Molecular clustering is typically interpreted to indicate recent transmission but can also represent the reactivation of latent infection of endemic strains (Nguyen *et al.* 2003).

4.1.2 Phylodynamics and molecular dating

Phylodynamic approaches combining molecular-dating with spatial and epidemiological data offer a powerful way to study pathogen evolution in space and time. Recent phylodynamic studies of *M. tuberculosis* have provided valuable new insights into the evolution of drug resistance (Eldholm *et al.* 2015; Eldholm *et al.* 2016), the emergence and spread of outbreaks (Bjorn-Mortensen *et al.* 2016; Roetzer *et al.* 2013), and historic events associated with *M. tuberculosis* dispersal (Eldholm *et al.* 2016).

Molecular dating methods estimate rates of genomic evolution in units of time, enabling absolute timescales to be placed on phylogenetic trees and evolutionary events such as lineage divergence or the appearance of mutations to be dated. Converting rates of sequence evolution and divergence events from relative to absolute time requires calibration of the molecular clock. Clock calibrations can be performed using information from the fossil record, geological or historical events, or from sample tip-dates. Tip-based calibration requires that samples are collected at different time points and that sequence data captures measurable levels of genetic change with time, providing a clock like signal that is used to calibrate the tree. Sequences that have these properties are termed ‘measurably evolving’ or ‘heterochronous’ (Drummond *et al.* 2003). The ability to capture sufficient temporal signal in the data is dependent on the mutation rate, the number of variable sites in the sequences, and the temporal width of the sampling interval (Rieux and Balloux 2016). Whole genome sequencing of large datasets of bacterial pathogens, such as *M. tuberculosis*, shows the accumulation of genetic variation over epidemiological timescales, pushing *M. tuberculosis* into the realm of measurably

evolving populations and allowing tip-based calibration for fine-scaled molecular dating studies (Biek *et al.* 2015). The increased accessibility of whole genome sequencing (WGS) has also led to the sequencing of thousands of MTBC strains from throughout the globe and the deposition of this data into public repositories provides a rich archive of genomic data for studies of MTBC diversity and dispersal.

4.1.3 Objectives

Our collaborators Greg Cook and Htin Aung at the University of Otago, N.Z., have sequenced the genomes of thirteen Rangipo, seven Otara and five Southern Cross isolates on the Illumina MiSeq platform. This chapter presents genomic analyses I have undertaken using these data. The primary objectives of these analyses were; (1) to characterise the phylogenetic relationships and genetic diversity within and between these clusters, and (2) to trace the evolutionary history and identify the historical origin of the Rangipo strain.

4.2 Methods

4.2.1 Whole genome sequencing of New Zealand *M. tuberculosis* isolates

4.2.1.1 Clinical M. tuberculosis *isolates*

Clinical New Zealand *M. tuberculosis* isolates or genomic DNA (gDNA) for whole genome sequencing were kindly provided by the Waikato Hospital, Hamilton, New Zealand, and LabPLUS, Auckland, New Zealand.

Thirteen isolates in the Waikato Hospital archive that had been identified as Rangipo strains on the basis of contact tracing and/or IS*6110* molecular typing were available for WGS. Collection dates spanned an 18-year period (1991–2009). Four isolates; O, F, R/C2 and A, have previously been sequenced on the ABI SOLiD platform (Colangeli *et al.* 2014), and were re-sequenced under the IDs NZLK, NZLL, NZLM, NZLN, respectively. NZLK, NZLL and NZLN have previously been sent to LabPLUS for MIRU-VNTR 24-loci typing and share the ‘primary’

Rangipo MIRU-VNTR 24-loci profile (233325153324-341444223362) (Ruthe 2015).

Seven Otara and five Southern Cross isolates were selected for WGS by collaborators at LabPLUS. These were chosen to encompass a range of collection dates and variant MIRU-VNTR 24-loci typing patterns among Otara cluster isolates. All Southern Cross isolates had identical MIRU-VNTR 24-locus profiles (123326153326-343224123253). Three Otara isolates (NZL01, NZL10, NZL13) shared the 'primary' Otara MIRU-VNTR 24-locus profile and the remaining four isolates represented three Otara variants (Appendix C.1). Collection dates of the Southern Cross strain isolates spanned a 13-year period (2002–2015) and Otara strain isolates spanned a seven-year period (2006–2013).

4.2.1.2 Whole genome sequencing

Rangipo isolates provided by the Waikato Hospital were sent to collaborators Greg Cook and Htin Aung at the University of Otago for culture and gDNA extraction as previously described (Aung *et al.* 2016). gDNA extracted from Otara and Southern Cross isolates was provided by collaborators Sally Roberts and James Bower at LabPLUS for sequencing. gDNA was sequenced at New Zealand Genomics Limited (Otago, New Zealand) using paired-end 250-bp reads on an Illumina MiSeq using the Nextera^{TM} XT DNA Kit (Illumina Inc., Hayward, CA). For all isolates, raw paired-end sequencing data was supplied in fastq format for analysis.

4.2.2 Lineage assignment

Fastq files were analysed in the program KvarQ (Steiner *et al.* 2014) using the 'coll14' test suite to determine MTBC lineage. This test suite contains the full set of MTBC barcode/marker single nucleotide polymorphisms (SNPs) described in Coll *et al.* (2014). The 'resistance' test suite was used to screen isolates for common drug resistance mutations. This test suite contains common drug resistance-associated SNPs and checks short gene regions known to harbour different drug resistance-associated mutations (Steiner *et al.* 2014).

4.2.3 Global *M. tuberculosis* L4.4 dataset

4.2.3.1 Additional Rangipo M. tuberculosis *genomes*

An additional nine Rangipo genomes from a recently published study (Gautam *et al.* 2017) were kindly provided by Ronan O'Toole and Sanjay Gautam (University of Tasmania, Australia) for inclusion in the analyses. These isolates came from multiple geographical locations and were collected over a two-year period (2010–2011). Eight isolates shared identical Rangipo strain MIRU-VNTR 24-loci profiles (233325153324 341444223362), and the ninth isolate (strain 356) differed at one locus (Gautam *et al.* 2017). The code 'NZLR' was added to the start of the published IDs of each of these isolates.

4.2.3.2 Canadian DS6^{Quebec} lineage M. tuberculosis *genomes*

Unpublished WGS data for 25 Canadian clinical isolates belonging to the DS6^{Quebec} lineage were kindly provided by Caitlin Pepperell (University of Madison-Wisconsin, Wisconsin, U.S.A.). These were selected to encompass a broad range of collection dates and molecular typing patterns. Libraries were prepared as previously described (Eldholm *et al.* 2016) and sequenced on the Illumina MiSeq platform.

4.2.3.3 Publicly available L4.4 genomes

A list of 401 L4.4 genomes in publicly available databases was kindly provided by Sebastien Gagneux and Daniela Brites (Swiss Tropical and Public Health Institute, Basel, Switzerland). This list had been generated by screening over 12,000 genomes using KvarQ (Steiner *et al.* 2014) and IDs of genomes identified as L4.4 extracted. From the provided list, I excluded genomes that were identified as having low (<10X) or mixed coverage. Metadata for the remaining genomes were obtained from the NCBI Short Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) and the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena), and relevant publications in which genomes were first described. Genomes were excluded if country data was not available. If multiple genomes were available for an isolate only the first was included. Literature searches together with KvarQ screening of

fastq files was performed to identify additional recent L4.4 isolates not deposited when the original list was compiled.

Publicly available genomes included in the final L4.4 dataset (Section 4.2.3.4) included 188 genomes from 17 countries from recently published studies (Bjorn-Mortensen *et al.* 2016; Bryant *et al.* 2013a; Bryant *et al.* 2013b; Casali *et al.* 2014; Casali *et al.* 2012; Clark *et al.* 2013; Guerra-Assunção *et al.* 2015; Guerra-Assunção *et al.* 2014; Holt *et al.* 2018; Walker *et al.* 2013a; Zhang *et al.* 2013) and Broad Institute initiatives (broadinstitute.org) (Appendix C.4). Genomes were downloaded from the NCBI Short Read Archive (Leinonen *et al.* 2011) using fastq-dump (available in the NCBI SRA Toolkit).

4.2.3.4 Datasets for phylogenetic analyses

Two datasets were used for phylogenetic analyses: a global L4.4 dataset for maximum likelihood phylogenetic analysis (dataset 1, n = 236); and a subset comprised of all high quality L4.4.1.1/S isolates with known year of isolation for Bayesian phylogenetic analysis and molecular dating (dataset 2, n = 117) (Appendix C.4). Genomes were only included if they met the quality criteria described in Section 4.2.4. Datasets included New Zealand Rangipo and Otara cluster isolates sequenced on the Illumina platform as part of this project (Section 4.2.1), previously published Rangipo isolates (Section 4.2.3.1), Canadian DS6^{Quebec} lineage isolates (Section 4.2.3.2) and other global L4.4 genomes (Section 4.2.3.2).

4.2.4 Reference guided assembly and variant calling

FASTQ files were processed through the Reference Guided Assembly Pepperell Lab Pipeline (available at https://github.com/pepperell-lab/RGAPepPipe). Raw reads were trimmed using a quality threshold of 15 and reads less than 20 bp long were discarded using TrimGalore!, a wrapper script around Cutadapt (Martin 2011) and FASTQC (Andrews 2010). Trimmed reads were mapped to the *M. tuberculosis* H37Rv reference genome (NC_000962.3) (Cole *et al.* 1998b) using BWA MEM (Li 2013). Duplicates were removed using Picard tools (http://broadinstitute.github.io/picard) and local realignment around indels was performed using GATK (DePristo *et al.* 2011). Variants were called with Pilon

(Walker *et al.* 2014) using a minimum depth threshold of 10, base quality of 20 and mapping quality of 40. VCF files generated by Pilon were converted to FASTA format using Pepperell lab in house scripts that treat ambiguous calls and deletions as missing data ('?' and '-').

Bases in or within 10 bp of repetitive regions, including genes annotated as PE/PPE/PGRS, REP13E12, transposable and phage elements were removed from FASTA sequences prior to alignment as these are prone to mapping errors. Geneious R8 (Biomatters, N.Z.) was used to search H37Rv reference annotations (NC_000962.3) to identify repetitive regions and these were curated into a bed file. Repetitive regions were masked in FASTA files using bedtools maskfasta and then removed.

Mapping quality was assessed using Qualimap (Garcia-Alcalde *et al.* 2012). Genomes were excluded if the depth of coverage was <25X or if <75% of reads mapped to the reference genome. Genomes with missing data at >10% of sites were excluded from further analyses.

4.2.5 Phylogenetic inference

4.2.5.1 Nucleotide Alignments

Variant sites were extracted from concatenated whole genome alignments using SNP-sites (Page *et al.* 2016). Only genomes with missing data at <10% of sites and sites where at least 90% of isolates had high quality base calls were included in phylogenetic and molecular dating analyses. The alignment-thin tool in the program BALi-Phy (Suchard and Redelings 2006) was used to remove positions from alignments with <90% high confidence base calls.

4.2.5.2 Selection of nucleotide substitution model

jModelTest2 (Darriba *et al.* 2012; Guindon and Gascuel 2003) was used to perform statistical selection of the best fitting nucleotide substitution model for phylogenetic analyses. SNP alignments were input into jModelTest2 (Darriba *et al.* 2012) and options were set to include models with different rate categories (x4) and unequal base frequencies. Models with invariant sites were not considered as by default SNP

alignments lack invariant sites. Model selection was based on the Bayesian information criterion.

4.2.5.3 Maximum likelihood phylogenetic analysis

Maximum likelihood phylogenetic trees were inferred from SNP alignments in PhyML (Guindon *et al.* 2010; Guindon and Gascuel 2003) using x1000 bootstrap replicates and the general time reversible (GTR) substitution model as this was the best fitting model jModelTest2 (Section 4.2.5.2). Trees were visualised in ggplot (Yu *et al.* 2017) and Figtree (tree.bio.ed.ac.uk/software/figtree/). Mesquite was used to infer SNP distances along branches by parsimony reconstruction of ancestral states using maximum likelihood trees and SNP alignments as input.

4.2.6 Diversity and clustering analyses

The R-package PopGenome (Pfeifer *et al.* 2014) was used to calculate fixation indices (F_{ST}), and genetic diversity (π, average nucleotide pairwise difference between sequences) from SNP alignments. Pairwise SNP distances were calculated using the poppr::bitwise.dist package in R using the 'missing_match = T' option to count sites with missing data as matching (Kamvar *et al.* 2014). WGS-defined clusters were identified based on SNP-distance using the criteria of Walker *et al.* (2013a), whereby recent transmission was ruled out if isolates were separated by >12 SNPs.

4.2.7 Rangipo SNPs

High quality SNPs (coverage depth ≥10, base quality ≥20, minimum mapping ≥40, and excluding SNPs in repetitive regions) were extracted from VCF files and the vcf-isec tool in the VCFtools program package (Danecek *et al.* 2011) was used to identify Rangipo specific SNPs. SNPs shared by all Rangipo genomes were first identified and then compared against all other individual non-Rangipo genomes in the global L4.4 dataset (n = 220, Section 4.3.3.1) to identify SNPs unique to the Rangipo cluster. SNPs were annotated with snpEff v4.3 using the *M. tuberculosis* H37Rv database (Cingolani *et al.* 2012).

STRING network analysis was performed on proteins harbouring non-synonymous SNPs (nsSNPs) shared by all Rangipo strain isolates and Rangipo specific nsSNPs (RS-nsSNPs). Network analysis and clustering using the Markov Cluster Algorithm was performed using default settings to identify clusters of functionally connected proteins. Clusters containing ≥10 proteins were then analysed independently to identify functional enrichments by gene ontology (GO) biological processes.

4.2.8 Bayesian evolutionary analysis

4.2.8.1 Assessment of temporal signal in data

Prior to BEAST2 analyses, datasets were assessed to determine if there was sufficient temporal signal in the data for accurate molecular dating. Maximum likelihood trees for all L4.4.1.1/S isolates with dates (dataset 2, n = 117) and for the DS6Q clade containing the New Zealand isolates (n = 47) were constructed in PhyML as in Section 4.2.5.3. Tempest (Rambaut *et al.* 2016) was used to determine the root-to-tip distance for regression analysis against tip date and regression analysis was performed in R. As further assessment of the strength of the temporal signal in the data, date randomisation was performed on the L4.4.1.1/S dataset used for molecular dating. Sampling dates of isolates were randomly shuffled 20 times and the randomised dates were then substituted for the real dates in XML files. Date randomised and original datasets were analysed in BEAST2 as described in Section 4.2.8.2 using a strict clock and constant coalescent demographic model.

4.2.8.2 BEAST analyses

BEAST (Bayesian Evolutionary Analysis by Sampling Trees) is a software package for Bayesian phylogenetic analysis of molecular sequence data (Bouckaert *et al.* 2014). It is focused towards inferring rooted, time-scaled phylogenies and provides a platform for parameter estimation and hypothesis testing of evolutionary models. Clock calibrations can be incorporated as priors on node-ages or using a tip-dating approach. BEAST uses Markov chain Monte Carlo (MCMC) sampling to estimate the posterior distribution (the probability distribution over parameter state space). Evolutionary parameters are sampled from their target distributions and the output provides an approximation of the posterior distribution from which statistics such

as the median and the 95% highest posterior density (HPD) interval are reported. The *x*% HPD is the smallest interval that encompasses *x*% of the probability distribution and is analogous to confidence intervals reported in classical statistics.

Mutation rates and divergence times were estimated for the L4.4.1.1/S sublineage using MCMC sampling in BEAST2 (v2.4) (Bouckaert *et al.* 2014) with the BEAGLE library v2.1 (Ayres *et al.* 2012) to facilitate rapid likelihood calculations. XML input files were generated in BEAUTi using an input alignment of 3161 variant sites from 117 L4.4.1.1/S genomes with known year of isolation (dataset 2). Only nucleotide positions where at least 90% of isolates had high quality data where included in molecular dating analyses (3,949,977 sites total). XML-input files were manually modified to specify the number of invariant sites, as calculated by scaling the number of non-SNP sites by the frequency of each base in the alignment.

BEAST2 analyses were performed using the GTR substitution model with the strict and relaxed (uncorrelated log-normal distribution (UCLD)) clocks, and coalescent constant, exponential, or Bayesian skyline plot (BSP) demographic models. The performance of various clock and demographic models was assessed using path sampling (Section 4.2.8.3). Trees were calibrated using tip dates covering a 27-year period (1987 to 2013). Two monophyletic taxon sets were created to ensure the root was correctly placed as determined from the maximum likelihood tree. Uniform prior distributions were defined for the substitution rate (1×10^{-10}–1×10^{-6} s/s/y). A uniform prior distribution was defined for population size in the constant and exponential demographic models (upper bound of 1×10^{10}), and the Jeffrey's (1/X) prior was unselected for the skyline model. The default Jeffrey's (1/X) prior was not used for this parameter as it is an improper prior and therefore an unsuitable prior for path sampling as this requires sampling from the prior distribution (Baele *et al.* 2013). Default priors were used for the remaining parameters.

To estimate posterior distributions for each model, three independent MCMC chains were run for 150–350 million states, sampling every 10,000 states. The first 10% of states were discarded as burn-in and chains were assessed for convergence and sufficient mixing as indicated by effective sample sizes >200 for all parameters. The three chains less burn-in were then combined in LogCombiner and parameter estimation was based on the combined chain. Median parameter estimates are reported unless otherwise specified.

The maximum clade credibility tree was estimated from the combined tree samples in TreeAnnotator using median node heights. Trees were visualised and edited in Figtree (tree.bio.ed.ac.uk/software/figtree/) and ggplot (Yu *et al.* 2017).

For all models, MCMC sampling was also used to sample from the prior in the absence of data to check there were no unexpected interactions between priors, and that the posterior estimate is not simply an artefact reflecting the prior. The effect of the prior selection on parameter estimation was examined by defining different upper bounds on the uniform distribution for the effective population size (1×10^8 and 1×10^6) and using the default 1/X prior, in a strict clock and constant demographic model with all other parameter settings kept constant.

4.2.8.3 Model selection

In order to assess the performance of the various clock and demographic models used for molecular dating, marginal likelihood estimation (MLE) was performed using path sampling (Lartillot and Philippe 2006). Path sampling runs were set up and analysed using the BEAST Model Selection package (v1.3.4) and performed in BEAST2 (v2.5). For each model, 100 path steps were specified using the proportions of a $\beta(0.3, 1.0)$ distribution to space out the steps. The pre-burn-in was specified as 10% of the total number of steps used in the standard MCMC run and for each step the default burn-in of 50% was used. Models were ranked based on average log MLE for two replicate runs. The log Bayes factor were calculated relative to the top ranked model. Log Bayes factors were interpreted according to the guidelines of Kass and Raftery (1995).

4.3 Results

4.3.1 New Zealand *M. tuberculosis* cluster isolates

4.3.1.1 Illumina whole genome sequencing

Thirteen Rangipo, seven Otara and five Southern Cross New Zealand clinical *M. tuberculosis* isolates were sequenced on the Illumina MiSeq platform (Section 4.2.1) and I was provided with raw data in fastq format for analysis as part of this

thesis. The mean raw read depth was 158X (34–258X) and median read length was 250 bp (32–251 bp).

Trimmed reads were mapped to the *M. tuberculosis* H37Rv reference genome as described in Section 4.2.4. The mean depth of coverage across the reference genome was 105X (30–226X) and on average 98.8% of the genome had a depth of coverage of at least 10X (95.0–99.4%) (Table 4.1, Appendix C.2).

Table 4.1. *Summary of Illumina WGS data for New Zealand* M. tuberculosis *cluster isolates.*

Cluster	n	Year	Mean raw read depth	Mean coverage depth[1]
Rangipo	13	1991–2009	64X (34–107)	50X (30–99)
Southern Cross	5	2002–2015	234X (206–258)	190X (168–226)
Otara	7	2003–2013	172X (77–237)	146X (72–207)
All genomes	25	1991–2015	128X (34–258)	105X (30–226)

[1] Average mapped coverage across the H37Rv reference genome (NC_000962.3).

4.3.1.2 Lineage assignment

Lineage classification was performed using KvarQ (Steiner *et al.* 2014). This classified the Rangipo and Otara clusters as belonging the L4.4.1.1/S sublineage of the L4/Euro-American lineage and Southern Cross as L4.3.3/LAM, consistent with the classification in Section 2.2.3. No drug resistance mutations were identified.

4.3.1.3 Phylogeny, clustering and genetic diversity of the New Zealand isolates

Twenty-four of our New Zealand *M. tuberculosis* genomes and an additional four Rangipo strain *M. tuberculosis* genomes available from a previous study (Gautam *et al.* 2017) met the quality criteria for inclusion in downstream phylogenetic analyses. A maximum likelihood tree of these 28 genomes was inferred using a 1735 bp SNP alignment as described in Section 4.2.5.3 and included an L2/Beijing strain isolate as an outgroup to root the tree (Figure 4.5). This SNP-based phylogeny shows that the Rangipo, Otara and Southern Cross MIRU-VNTR-defined clusters form three well-differentiated monophyletic clades with differing phylogenetic structures. The Rangipo and Otara clusters were separated from one another by 108 SNPs and from the Southern Cross cluster by 554 and 524 SNPs, respectively.

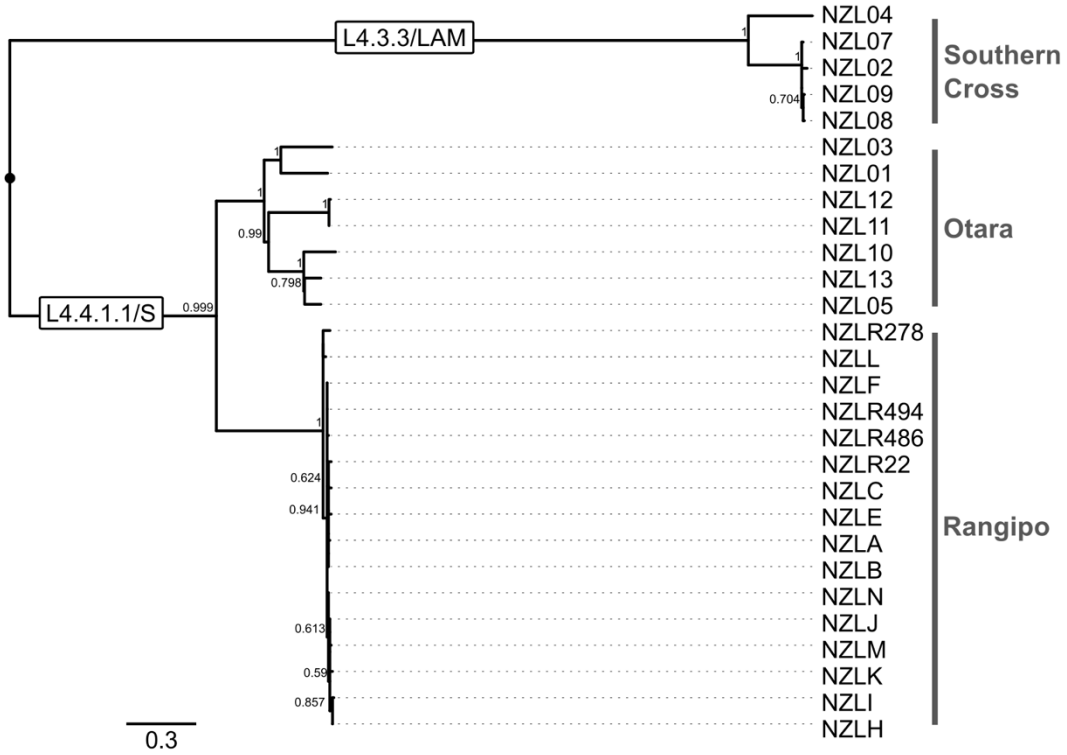


Figure 4.5. *Phylogeny of New Zealand* M. tuberculosis *cluster isolates.* Maximum likelihood phylogeny of 28 clinical New Zealand *M. tuberculosis* genomes belonging to the Rangipo, Otara and Southern Cross clusters. The most recent common ancestor is shown as a black circle (rooted to L2/Beijing, not shown). Bootstrap values are shown for nodes with >50% support.

WGS offers higher resolution to delineate recent transmission events and define clusters of infection compared with MIRU-VNTR (Bryant *et al.* 2013b; Roetzer *et al.* 2013; Walker *et al.* 2013a). A threshold of five or fewer SNPs has been suggested to define strains involved in recent transmission events and more than 12 SNPs to rule out recent transmission (Walker *et al.* 2013a). Using these thresholds, SNP distances were used to identify WGS-defined clusters within each of the MIRU-VNTR clusters (Figure 4.6). As phylogenies were constructed using only high-quality sites with >90% confident base calls (which could potentially decrease the number of SNPs between isolates leading to an over estimation of clustering), pairwise differences calculated from all SNP positions were also examined (Table 4.3) and all isolates still fell into the same WGS-defined clusters.

Table 4.2. *Genetic variation in New Zealand* M. tuberculosis *cluster isolates.*

Cluster	n	Total SNPs[1]	Pairwise SNPs[2]	π
Rangipo	16	729	0–12 (4)	3.3
Southern Cross	5	812	2–92 (6)	28.2
Otara	7	886	2–107 (84)	66.3

[1] SNPs relative to H37Rv.
[2] Median SNP distance is shown in brackets.

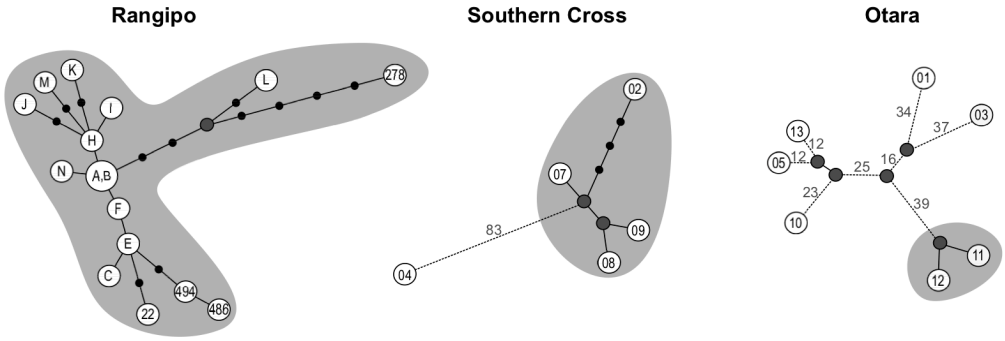


Figure 4.6. *WGS-based clustering of New Zealand* M. tuberculosis *isolates*. Maximum likelihood SNP phylogenies of MIRU-VNTR defined clusters showing WGS-defined clusters in grey. Individual isolates are indicated by numbers within white circles (prefixes NZL and NZLR removed from IDs) and unsampled internal nodes are shown in dark grey. Smaller black circles indicate nodes separated by more than one SNP, with each circle representing one additional SNP difference. Distances between nodes outside of clusters are shown with dashed lines and numbers represent SNP differences between nodes. Not to scale.

Pairwise distances and nucleotide diversity (π) were calculated as measures of genetic variation within each MIRU-VNTR defined cluster (Table 4.2, Table 4.3). Rangipo isolates harboured the least genetic diversity of the three clusters ($\pi = 3.3$) with a maximum of 12 SNPs between any two isolates (0–12 SNPs; median, 4). Accordingly, all Rangipo isolates belong to a single WGS-defined cluster (Figure 4.6) and the phylogeny is characterised by short terminal branches (0–5 SNPs, median 1 SNP), indicative of recent clonal expansion and temporally short transmission chains. A total of 729 SNPs relative to H37Rv were identified in the 16 Rangipo strain genomes. This is comparable to previous reports of 747 and 727 SNPs identified among ten and nine Rangipo strains, respectively (Colangeli *et al.* 2014; Gautam *et al.* 2017).

The Otara phylogeny is characterised by long terminal branches (0–37 SNPs; median, 12 SNPs) and Otara isolates harboured the greatest genetic diversity of the three MIRU-VNTR clusters (π = 66.3) (2–107 pairwise SNPs; median, 94) indicating that this cluster represents an endemic strain in Pacific people. Unlike the Rangipo and Southern Cross isolates which shared identical MIRU-VNTR typing patterns (except for one Rangipo isolate), the Otara isolates represent four variant typing profiles, distinguishable by a single locus difference (Appendix C.1). Two isolates sharing the same variant MIRU-VNTR profile (NZL11 and NZ12) were separated by only two SNPs indicating recent transmission. All remaining pairwise comparisons differed by 28–107 SNPs and the three isolates sharing the 'primary' Otara MIRU-VNTR type (NZL01, NZL10 and NZL13) were not monophyletic and differed by 40–103 SNPs. This indicates that the higher levels of SNP diversity observed for this cluster are not simply the result of the inclusion of isolates with more variant MIRU-VNTR profiles.

The five Southern Cross isolates differed by 2–92 SNPs (median, 6 SNPs), although large SNP differences were attributable to just a single isolate, NZL04. The phylogeny shows NZL02, NZL07, NZL08 and NZL09 are closely related differing by only 2–6 SNPs and belong to a WGS-defined cluster, indicating relatively recent transmission. This strain caused a large tuberculosis outbreak in 1999 at an Auckland church attended by Pacific peoples, resulting in 27 cases of active TB and a further 57 cases of latent TB (Hill and Calder 2000). This cluster of isolates likely traces back to that outbreak. NZL04 however was separated by 89–92 SNPs from other Southern Cross isolates and does not belong to a WGS-defined cluster, despite sharing matching MIRU-VNTR profiles with the other Southern Cross isolates. This hints that the Southern Cross cluster might belong to a larger endemic Pacific strain, as for the Otara cluster, however more data is needed to investigate this.

Table 4.3. *Pairwise SNP distances*. Pairwise distance matrixes showing SNP variation between isolates within each of the three New Zealand M. tuberculosis clusters.

Rangipo

	NZLA	NZLB	NZLC	NZLE	NZLF	NZLH	NZLI	NZLJ	NZLK	NZLL	NZLM	NZLN	NZLR22	NZLR278	NZLR486	NZLR494
NZLA																
NZLB	0															
NZLC	3	3														
NZLE	2	2	1													
NZLF	1	1	2	1												
NZLH	1	1	4	3	2											
NZLI	2	3	5	4	4	1										
NZLJ	3	3	6	5	4	2	3									
NZLK	3	3	6	5	5	2	3	4								
NZLL	5	5	7	6	5	6	6	8	8							
NZLM	3	3	6	5	4	2	3	4	4	8						
NZLN	1	1	4	3	2	2	3	4	4	6	4					
NZLR22	4	5	3	2	4	5	6	7	7	8	7	5				
NZLR278	8	9	11	10	9	9	10	11	11	7	11	9	12			
NZLR486	4	5	4	3	4	5	5	7	7	8	7	5	5	12		
NZLR494	4	5	3	2	4	5	5	7	7	8	7	5	4	12	1	

Southern Cross

	NZL02	NZL04	NZL07	NZL08	NZL09
NZL02					
NZL04	92				
NZL07	5	89			
NZL08	6	90	3		
NZL09	6	90	3	2	

Otara

	NZL01	NZL03	NZL05	NZL10	NZL11	NZL12	NZL13
NZL01							
NZL03	78						
NZL05	93	97					
NZL10	103	107	37				
NZL11	98	103	86	96			
NZL12	96	101	84	94	2		
NZL13	95	99	28	40	88	86	

4.3.1.4 Phylogenetic analysis of additional Rangipo strain genomes

Six of the 22 available Rangipo genomes failed to meet the strict quality criteria for inclusion in downstream phylogenetic analyses. This was predominantly due to low coverage (<25X) and/or a high proportion of missing base calls (>10%). To examine the phylogenetic relationship between these isolates and the included genomes, a phylogenetic analysis using more relaxed quality thresholds was performed. Genomes were re-processed as in Section 4.2.4 with the exception that variants were called using a minimum depth threshold of five instead of ten. One genome still had a high number of missing base calls and was excluded (NZLR547, 31% missing sites). The remaining 21 genomes had <12% of missing sites and were all used for this analysis. Isolate collection dates covered a 21-year period (1991–2011) and came from a range of geographical locations in New Zealand.

The 21 Rangipo genomes were found to differ by 38 SNPs and a total of 742 SNPs relative to H37Rv were identified. Individual isolates were separated by 0–14 SNPs. A maximum likelihood phylogeny was constructed from a 710 bp alignment including H37Rv to root the tree as in Section 4.2.5.3 (Figure 4.7A). An unrooted maximum likelihood phylogeny was also constructed using a 38 bp alignment including all variant bases between the 21 Rangipo genomes (Figure 4.7B). The excluded isolates are dispersed throughout the phylogeny and none are more deeply rooted than those retained, indicating that the retained isolates provide a good representative sample of this cluster.

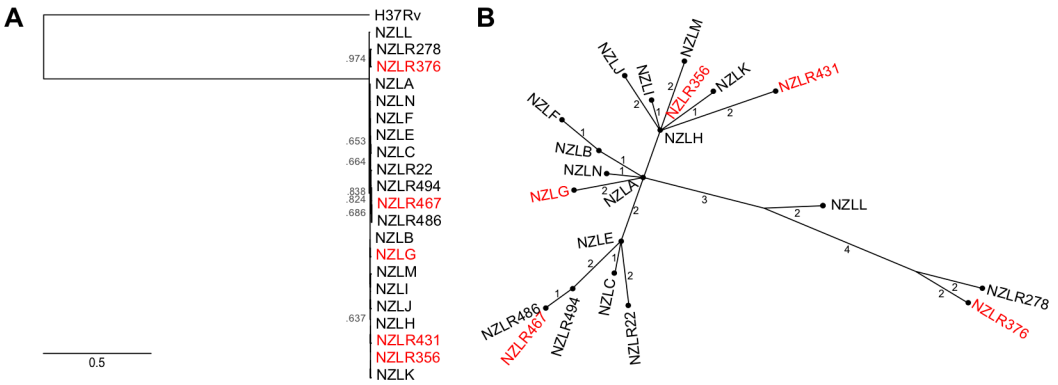


Figure 4.7. *Phylogeny of 21* M. tuberculosis *Rangipo strain genomes.* Genomes excluded from downstream phylogenetic analyses are shown in red. (A) Maximum likelihood phylogeny inferred from a 710 bp alignment rooted to H37Rv. (B) Unrooted maximum likelihood phylogeny inferred from a 38 bp alignment including all variant positions between Rangipo genomes. Distances between nodes are in number of SNPs.

4.3.2 Rangipo strain SNPs

4.3.2.1 Common Rangipo SNPs

A total of 513 high quality SNPs were identified in all 16 high quality Rangipo genomes (Appendix C.3). 438 of these were found in protein coding regions, 272 of which were nsSNPs (62.1%), including four stop codon gains and one stop codon loss. 254 different genes harboured common Rangipo SNPs that altered protein coding sequence. Functional connections among the affected proteins were investigated using the STRING database (Szklarczyk *et al.* 2014). STRING identifies known and predicted protein-protein associations including direct (physical) interactions, as well as indirect (functional) interactions. The STRING network identified 424 functional interactions and an average of 3.3 interactions per protein (Table 4.4). No significant GO functional enrichments were detected. Cluster analysis identified 25 clusters of three or more functionally connected proteins three of which comprised ≥10 proteins. The largest was comprised of 13 proteins, including eight polyketide synthases (Figure 4.8). GO functional enrichment analysis showed this cluster was enriched for proteins in lipid biosynthetic process, cellular lipid metabolic process and DIM/DIP cell wall layer assembly pathways (Table 4.5). The GO term "DIM/DIP cell wall layer assembly," encompasses genes involved in the synthesis and export of surface glycolipid phthiocerol dimycocerosates (PDIM). The third largest cluster was enriched for GO pathways relating to DNA replication and repair and no significant enrichments were identified in the second largest cluster.

Table 4.4. *STRING network analysis of proteins harbouring common Rangipo strain nsSNPs*. STRING network statistics for 254 proteins containing nsSNPs shared by all Rangipo genomes and for clusters of ≥10 proteins within the full network.

	All	Cluster 1	Cluster 2	Cluster 3
Number of nodes	254	13	11	10
Number of edges	424	31	24	17
Average node degree[1]	3.3	4.8	4.4	3.4
Average local clustering coefficient[2]	0.345	0.404	0.721	0.796

[1] Average number of interactions for protein in the network
[2] How connected the nodes in the network are, high values = highly connected.

Table 4.5. *Functional enrichment analysis for clusters of ≥10 proteins harbouring common Rangipo strain nsSNPs.* Enrichment of GO terms in the first and third largest clusters of proteins harbouring nsSNPs common to all Rangipo genomes are shown. The second largest cluster did not have any function enrichments.

Cluster	Nodes	Pathway description	Genes	False discovery rate
1	13	lipid biosynthetic process	6	0.00331
		cellular lipid metabolic process	6	0.00331
		DIM/DIP cell wall layer assembly	3	0.0102
3	10	DNA replication	5	1.3e-05
		DNA ligation involved in DNA repair	2	0.00538
		DNA metabolic process	4	0.012
		cellular macromolecule biosynthetic process	6	0.012
		cellular response to stress	4	0.0144
		DNA repair	3	0.0366

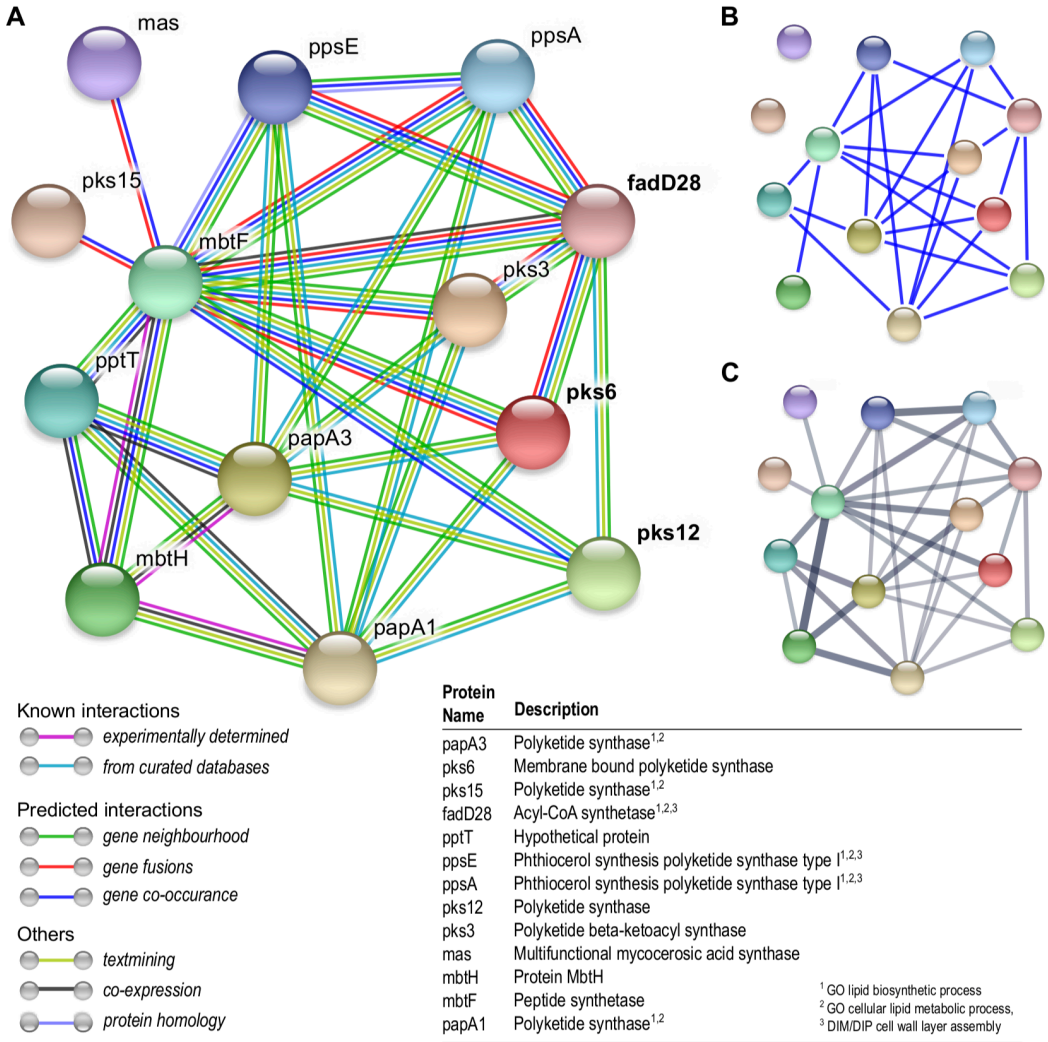


Figure 4.8. *STRING protein-protein associations among the largest cluster of proteins harbouring common Rangipo SNPs.* SNPs identified as specific to the Rangipo strain are shown in bold. (A) Network showing all types of interaction evidence among proteins. (B) predicted binding interactions. (C) strength of data supporting interactions, the thicker and darker the line the higher the confidence in the association.

4.3.2.2 Rangipo specific SNPs

The 513 common Rangipo SNPs were compared against 220 non-Rangipo L4.4 sublineage *M. tuberculosis* genomes in the global L4.4 dataset compiled for phylogenetic analyses (Section 4.3.3.1). 53 SNPs were found to be specific to the Rangipo strain (RS-SNPs), 29 of which were nsSNPs (54.7%) (Table 4.6, Appendix C.3). The positions of all 29 RS-nsSNPs and the Rv1821/*secA2* diagnostic marker synonymous SNP (sSNP) were examined in the six Rangipo genomes excluded from phylogenetic analyses (Section 4.3.1.4) and all carried the Rangipo variant base. All 14 Rangipo specific nsSNPs and six of eight sSNPs identified based on SOLiD WGS data in Section 2.2.2 were confirmed as RS-SNPs in this analysis; including the Rv2893 G72S SNP further investigated in Chapter Three and the Rv1821/*secA2* diagnostic maker SNP (Section 2.2.5).

In silico functional effect prediction was performed for the additional 15 RS-nsSNPs identified in this analysis the using the program SNAP2 (Section 2.1.3.1). Three of these were predicted have functional effects, two of which are found in known virulence factors – Rv0022c/*whiB5* and Rv2048c/*pks12*. Overall, 11 of the 29 RS-nsSNPs (37.9%) were predicted to have functional effects, including four in known virulence factors: Rv0405/*pks6* S1236L, Rv1161/*narG* Y802H, Rv0022c/*whiB5* R43H and Rv2048c/*pks12* R669P (Table 4.6).

Functional category assignment of genes harbouring the 29 RS-nsSNPs shows genes involved in the lipid metabolism category accounted for 17.2% of RS-nsSNPs (5/29), a nearly 3-fold enrichment relative to the overall proportion in the *M. tuberculosis* genome (6%) (Camus *et al.* 2002) (Table 4.7). This is similar to the analysis of the 14 RS-nsSNPs in Section 2.2.2.1, in which a 3.6-fold enrichment of genes involved in lipid metabolism was observed (3/14, 21.4% of RS-nsSNPs). STRING network analysis was performed on all 29 proteins carrying RS-nsSNPs. This identified 16 predicted interactions and an average of 1.1 interactions per protein (Figure 4.9). No significant GO functional enrichments were detected. Cluster analysis identified six different clusters, three comprised of three proteins and the remaining three of two proteins. The three proteins in one of these clusters (Pks12, FadD28 and Pks6) were also found in the largest cluster in the full network of 254 proteins (Figure 4.8).

Table 4.6. *Rangipo specific nsSNPs identified from Illumina WGS data*. Functional effect prediction output from SNAP2 is shown and virulence factors are indicated in bold type.

SNP	Locus/ Gene ID	aa change	Product	SNAP2 Effect	SNAP2 Score	Predicted Accuracy
Rv3253CG	Rv0002/*dnaN*	P401R	DNA polymerase III (beta chain) DnaN	N	-26	61%
Rv27315CT	**Rv0022c/*whiB5***	R43H	Probable transcriptional regulatory protein WhiB-like WhiB5	Y	59	75%
Rv45753TC	Rv0041/*leuS*	V731A	Probable leucyl-tRNA synthetase LeuS	N	-84	93%
Rv129668GA	Rv0107c/*ctpI*	P292S	Probable cation-transporter ATPase I CtpI	N	-61	82%
Rv489437CT	**Rv0405/*pks6***	S1236L	Membrane bound polyketide synthase	Y	46	71%
Rv550620GT	Rv0458	D316Y	Probable aldehyde dehydrogenase	Y	14	59%
Rv616877CT	Rv0526	A11V	Possible thioredoxin protein	N	-18	57%
Rv1032721CT	Rv0926c	A356T	Conserved hypothetical protein	N	-96	97%
Rv1289731TC	**Rv1161/*narG***	Y802H	Respiratory nitrate reductase (alpha chain) NarG	Y	80	91%
Rv1836099TG	Rv1631/*coaE*	Y363D	Probable dephospho-CoA kinase CoaE	Y	87	91%
Rv2304891CG	**Rv2048c/*pks12***	R699P	Polyketide synthase Pks12	Y	24	63%
Rv2402434GA	Rv2142c/*parE2*	A26V	Possible toxin ParE2	N	-25	61%
Rv2807374GC	Rv2492	G33R	hypothetical protein	Y	79	85%
Rv3202633GA	Rv2893	G72S	Possible oxidoreductase	Y	28	63%
Rv3219432CG	**Rv2912c**	G144R	Probable transcriptional regulatory protein (probably TetR-family)	N	-55	78%
Rv3283879TC	**Rv2941/*fadD28***	V182A	Fatty-acid-AMP ligase FadD28	N	-77	87%
Rv3366098GA	Rv3007c	P118L	Possible oxidoreductase	Y	59	75%
Rv3561770AG	Rv3193c	L468S	Probable conserved transmembrane protein	N	-17	57%
Rv3654060GA	Rv3272	A205T	Conserved hypothetical protein	Y	68	80%
Rv3858800GA	Rv3439c	A288V	Conserved hypothetical alanine and proline rich protein	N	-54	78%
Rv3895925GA	Rv3479	A36T	Possible transmembrane protein	N	-80	87%
Rv3925115GT	Rv3506/*fadD17*	A76S	Fatty-acid-CoA synthetase FadD17	N	-52	78%
Rv3980075CA	**Rv3540c/*ltp2***	E195D	Probable lipid transfer protein or keto acyl-CoA thiolase Ltp2	N	-67	82%
Rv3985279CG	**Rv3545c/*cyp125***	G56A	Probable cytochrome P450 125 Cyp125	N	-91	97%
Rv4085870GA	Rv3646c/ *topA*	T463M	DNA topoisomerase I TopA	N	-5	53%
Rv4150655GA	Rv3707c	A129V	Conserved hypothetical protein	N	-62	82%
Rv4163025GC	Rv3719	Q240H	Conserved protein	N	-89	93%
Rv4177114CA	Rv3728	P748T	Probable conserved two-domain membrane protein	N	-46	72%
Rv4377908CT	Rv3894c/*eccC*$_2$	G849S	ESX-2 type VII secretion system protein EccC2	Y	20	63%

Table 4.7. *Gene functional categories of Rangipo specific nsSNPs*. The percentage of genes in H37Rv as reported in Camus *et al.* (2002) in each category is also shown.

Gene Functional Category	nsSNPs	H37Rv
Cell wall and cell processes	5 (17.2%)	18%
Conserved hypotheticals	6 (20.7%)	26%
Information pathways	3 (10.3%)	6%
Intermediary metabolism and respiration	7 (24.1%)	22%
Lipid metabolism	5 (17.2%)	6%
Regulatory proteins	2 (6.9%)	5%
Virulence, detoxification, adaptation	1 (3.4%)	2%
Total	29	

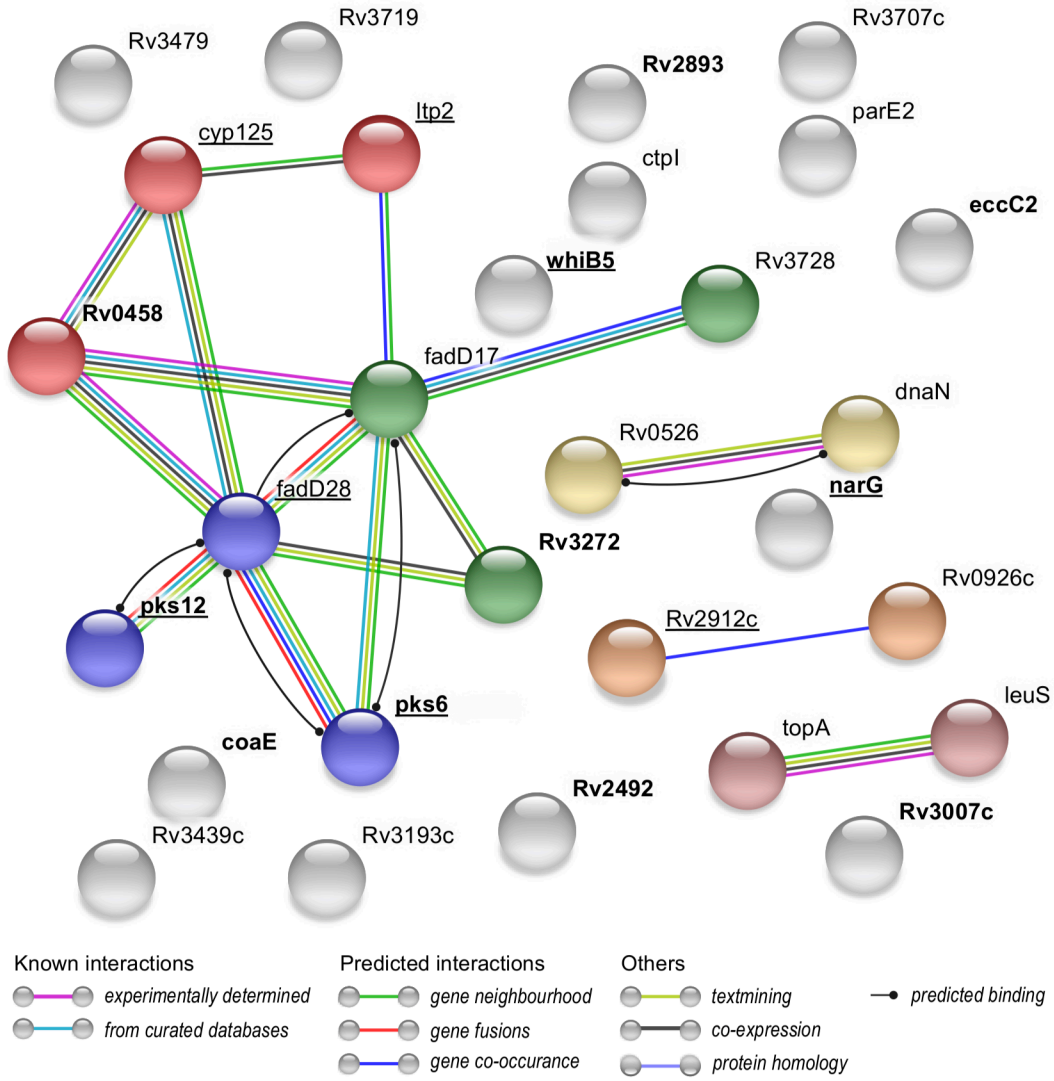


Figure 4.9. *STRING protein-protein associations among proteins harbouring Rangipo specific nsSNPs.* Proteins harbouring SNPs with predicted functional effects are shown in bold type and known virulence factors are underlined. Nodes are coloured by clusters as identified by STRING.

4.3.3 Global phylogeny of L4.4

To identify possible historical sources of the Rangipo and Otara clusters, a large genomic dataset comprising the New Zealand L4.4.1.1/S genomes and publicly available L4.4 genomes from around the world was compiled and used to infer a global L4.4 phylogeny. This was also used to determine the phylogenetic placement of the DS6^{Quebec} deletion and provide new insights into the global population structure and phylogeographical distribution of L4.4.

4.3.3.1 L4.4 dataset

The final dataset used to construct the L4.4 global phylogeny was comprised of 236 isolates (Appendix C.4, C.5). This included 23 high quality New Zealand Rangipo and Otara strain genomes (Section 4.3.1.3) and 213 global L4.4 genomes from 19 different countries representing all five global regions (Figure 4.10). Collection dates were available for 72.4% of isolates (171/236) and spanned a 27-year period from 1987 to 2013. L4.4 is reported to occur at high proportions in parts of Africa and Asia (Stucki *et al.* 2016b). Reflective of this, Africa was the most highly represented region accounting for 41.4% of all isolates (97/236) followed by Asia with 22.9% of isolates coming from this region (54/236). The New Zealand strains accounted for 9.7% of all L4.4 isolates (23/236).

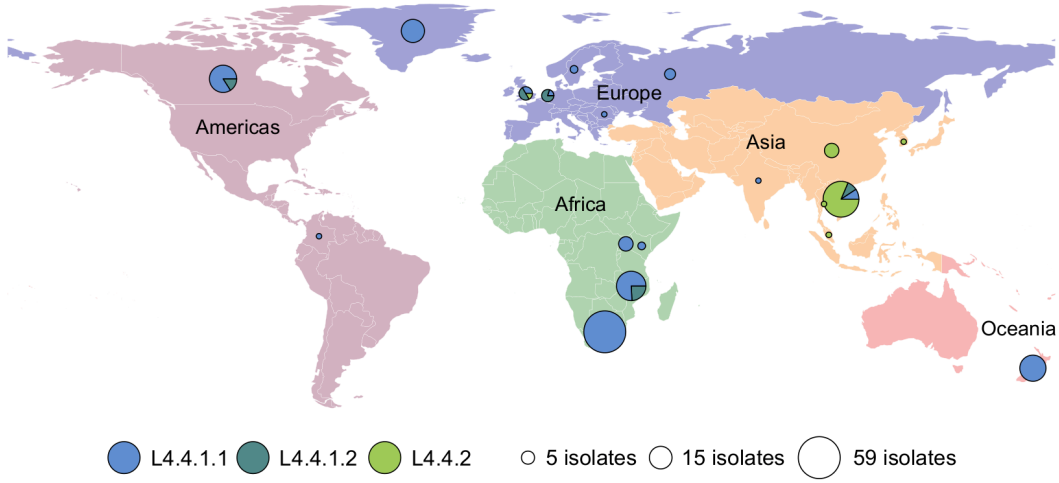


Figure 4.10. *Global distribution of 236 L4.4 sublineage genomes included in this study (dataset 1, n = 236).*

4.3.3.2 Maximum likelihood phylogeny of L4.4

A whole genome maximum likelihood phylogeny was constructed from a 9024 bp SNP alignment of 236 global L4.4 isolates and H37Rv to root the tree (Figure 4.11, Figure 4.12). This phylogeny splits L4.4 into three sublineages, L4.4.1.1/S, L4.4.1.2 and L4.4.2, consistent with the classification system of Coll *et al.* (2014). Pairwise fixation indices (F_{ST}) were calculated to estimate population separation between sublineages. F_{ST} values provide a measure of the genetic differentiation between populations and range from zero (no separation) to one (complete separation). F_{ST} values between three L4.4 sublineages were between 0.50 and 0.57, indicating that the sublineages are well differentiated.

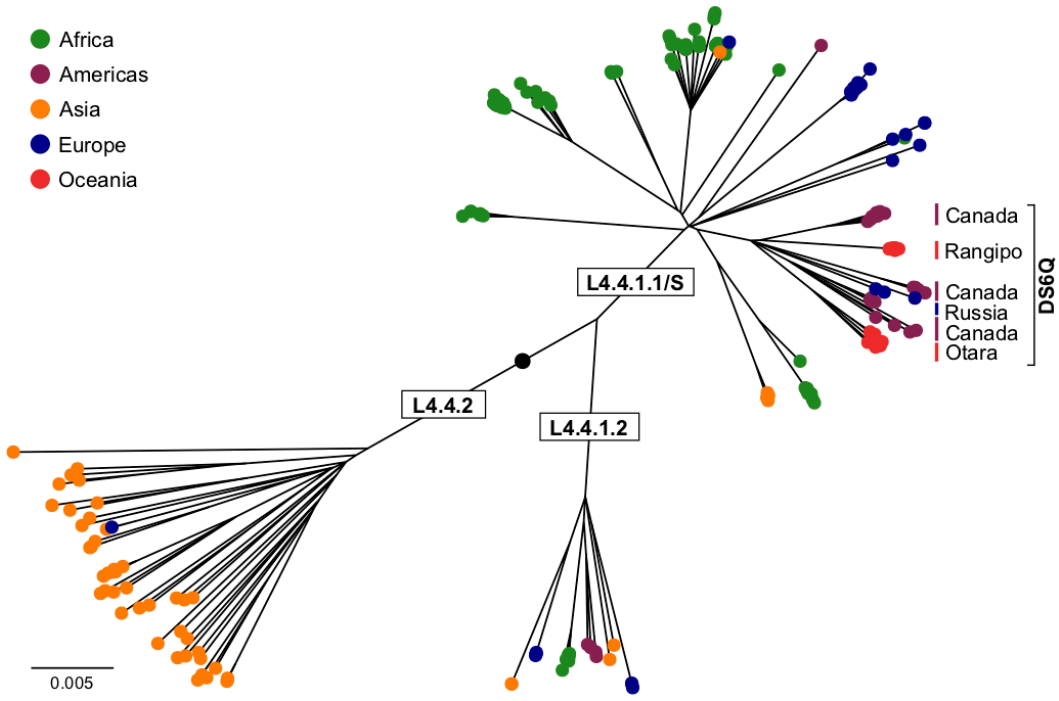


Figure 4.11. *Global phylogeny of the L4.4 sublineage*. Maximum-likelihood phylogeny of 236 global *M. tuberculosis* L4.4 sublineage genomes. Tips are coloured by global region and sublineages are labelled according to the nomenclature of Coll *et al.* (2014). A black circle indicates the node position of the most recent common ancestor of L4.4 (rooted to H37Rv, not shown). The DS6Q clade, the New Zealand Rangipo and Otara clusters, and country of other clades within the DS6Q are labelled.

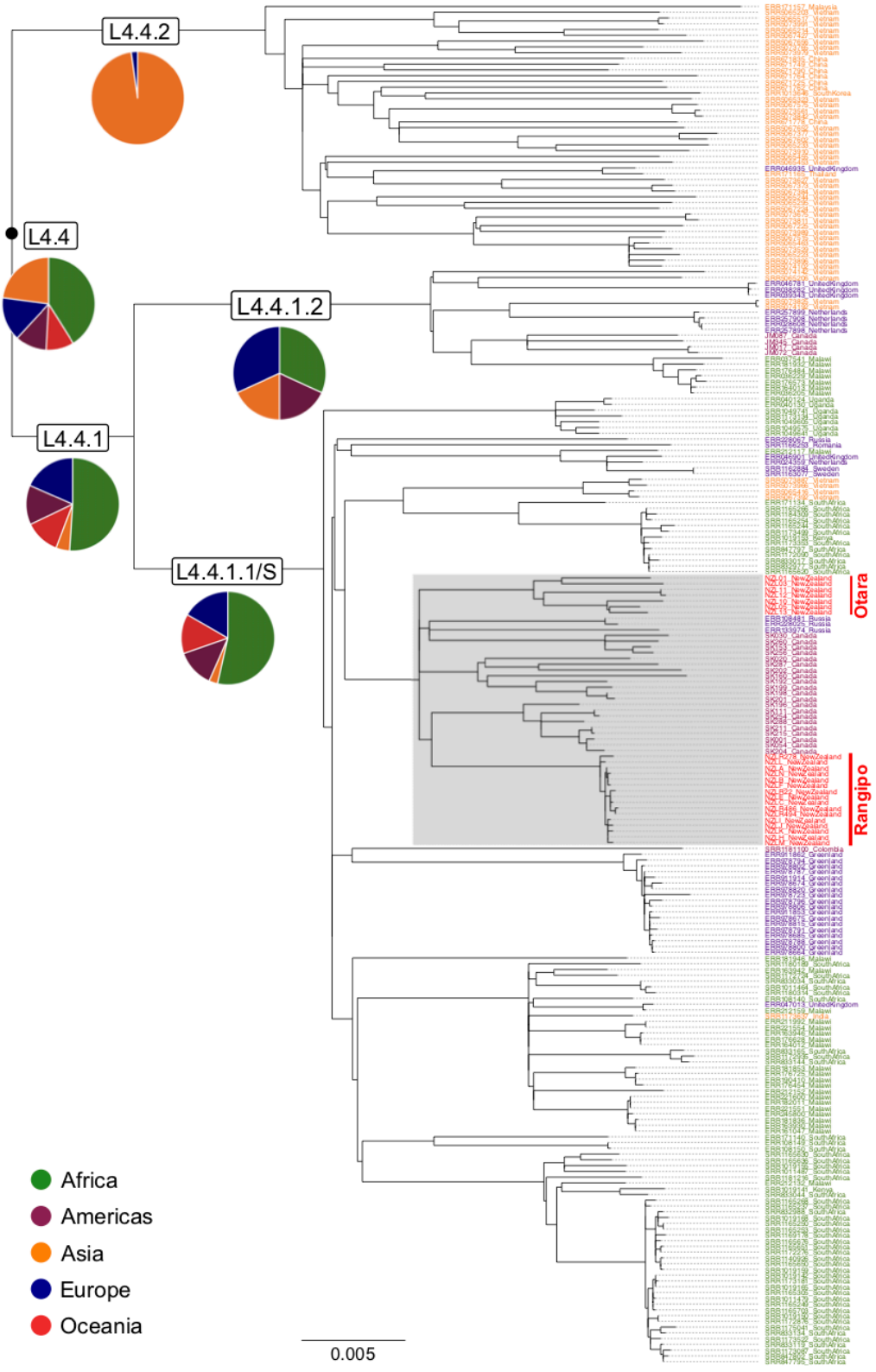


Figure 4.12. *Global phylogeny of the L4.4 sublineage and regional distribution of sublineages*. Maximum-likelihood phylogeny of 236 global *M. tuberculosis* L4.4 genomes. Lineages are labelled according to the nomenclature of Coll *et al.* (2014). Tips labels are coloured by global region and pie charts show the regional distribution of isolates within sublineages. A black circle indicates node position of the most recent common ancestor of L4.4 (rooted to H37Rv, not shown). A grey box highlights the DS6Q.

The L4.4 phylogeny revealed differing global distributions and population structures of the three L4.4 sublineages, which presumably reflects different demographic histories of these sublineages. L4.4.2 was essentially restricted to Eastern Asia (45/46 isolates, 97.8%) indicating *in-situ* expansion of this sublineage. Conversely, L4.4.1.1/S and L4.4.1.2 have broader global distributions and show limited phylogeographical structure, indicative of high rates of migration and efficient dispersal.

The New Zealand isolates accounted for 13.8% of L4.4.1.1/S sublineage genomes (23/168). The Rangipo and Otara clusters form two separate monophyletic clades within L4.4.1.1/S and are part of a larger clade consisting of 47 genomes from New Zealand, Canada and Russia (23, 21 and 3 genomes, respectively). The Canadian isolates belong to the DS6^{Quebec} strain family discussed in Section 2.3.3, we therefore termed this clade the DS6Q clade. The Canadian isolates comprise three separate clades within the DS6Q. The polytomy at the root of the DS6Q clade and the polyphyletic nature of the New Zealand and Canadian isolates implies dispersal of multiple closely related strains from a common origin and is consistent with at least two separate introductions of the DS6Q clade into New Zealand. The DS6^{Quebec} lineage is assumed to have been introduced to Canada from France (Pepperell *et al.* 2011) suggesting a similar European origin for the Rangipo and Otara clusters. This hypothesis was further explored using molecular dating (Section 4.3.4).

4.3.3.3 DS6^{Quebec} deletion

The DS6^{Quebec} deletion removes an ~11.4 kb region (positions 1987457 to 1998849 in H37Rv), truncating the *plcD* gene and deleting a further six protein coding genes present in H37Rv (Figure 2.7). The Rangipo and Otara clusters were confirmed to harbour the DS6^{Quebec} deletion by PCR (Section 2.2.4) and we hypothesised that this deletion might be more widely characteristic of the S lineage. All 236 genomes in the L4.4 dataset were examined for the presence of the DS6^{Quebec} deletion. All L4.4.1.1/S and L4.4.1.2 genomes, but no L4.4.2 genomes, were found to harbour this deletion, showing that it is a characteristic deletion of L4.4.1 (Figure 4.13). One L4.4.1.1/S genome had an earlier start of the deletion (position 1987142) indicating a subsequent small deletion event.

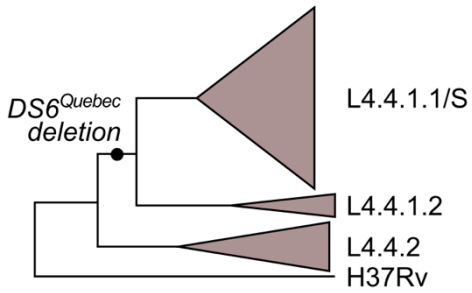


Figure 4.13. *Phylogenetic placement of the DS6^{Quebec} deletion in the L4.4 sublineage.*

4.3.4 Molecular dating

The temporal evolution of the L4.4.1.1/S sublineage was investigated using Bayesian evolutionary analysis in BEAST2. This enables inference of the time this sublineage emerged and allows a reconstruction of its demographic history. Molecular dating estimates for the DS6Q were able to be linked with historical trade and migrations, providing new insight into historical phenomena and social factors that have contributed to the dispersal and successful expansion of the Rangipo and Otara clusters.

4.3.4.1 L4.4.1.1/S dataset

The dataset used for molecular dating contained all L4.4.1.1/S genomes with known isolate collection dates (n = 117) (Appendix C.4, C.5). The range of collection dates was the same as for the full L4.4 dataset spanning a 27-year period from 1987 to 2013. Genomes came from 14 different countries covering all five global regions and contained all 47 genomes identified as belonging to the DS6Q clade (23 from New Zealand, 21 from Canada and 3 from Russia).

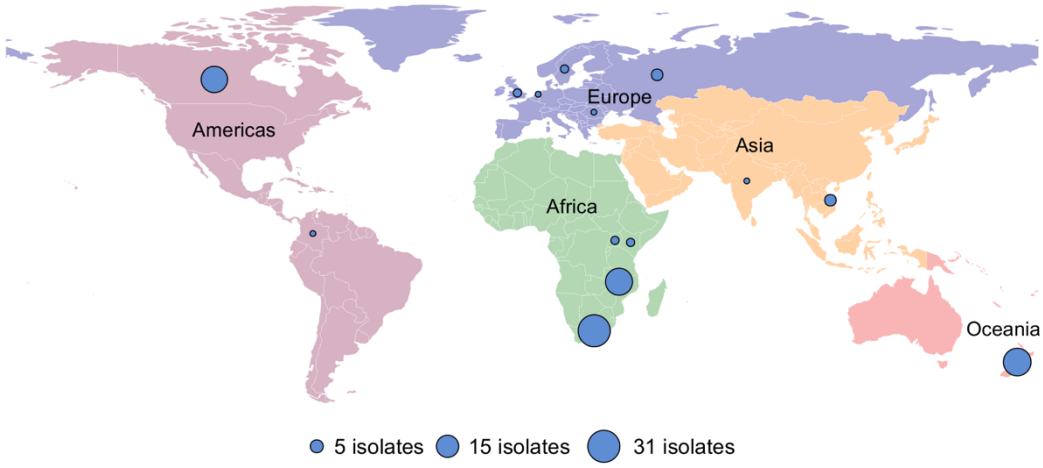


Figure 4.14. *Global distribution of L4.4.1.1/S sublineage genomes used for molecular dating analyses (dataset 2, n = 117).*

4.3.4.2 Assessment of temporal signal for tip-dating

Molecular dating was performed using tip dates to calibrate the molecular clock (Section 4.3.4.3). Prior to molecular dating analysis using tip dating it is necessary to establish whether or not the data contains sufficient temporal signal for tip-based calibration (i.e. to determine if the population is 'measurably evolving') (Drummond *et al*. 2003). If the population is measurably evolving, the root-to-tip distance of isolates in the undated phylogeny should correlate with sampling time as more recently collected samples will have had more time to acquire additional mutations and diverge further from the ancestor at the root. Root-to-tip regression analysis showed a modest temporal signal in the L4.4.1.1/S dataset (adjusted $R^2 = 0.229$) and a weaker signal in the DS6Q isolate subset (adjusted $R^2 = 0.139$) (Figure 4.15A,B), the L4.4.1.1/S dataset was therefore used for further dating analyses. There was no significant difference between the two slopes (t-test, $p = 0.0579$) (this provides an estimate of the evolutionary rate).

Date randomisation was performed as a further test of strength of the temporal signal in the L4.4.1.1/S dataset as this is a more robust method to test for temporal signal (Rieux and Balloux 2016). In this test, sample dates are permuted to create multiple date randomised datasets and parameters are estimated from the true and randomised dates then compared. At least 20 randomisation tests are recommended and under the strictest criterion and rate estimates obtained from the true dates

should not overlap with the 95% HPD intervals of estimates from the randomised dates (Duchêne *et al.* 2015). Twenty date randomisations were performed on the L4.4.1.1/S dataset. Estimates of the substitution rate and TMRCA for L4.4.1.1/S showed no overlap in the 95% HPD between the real and randomised dates (Figure 4.15C,D), indicating that the data contains sufficient temporal signal for tip-based calibration.

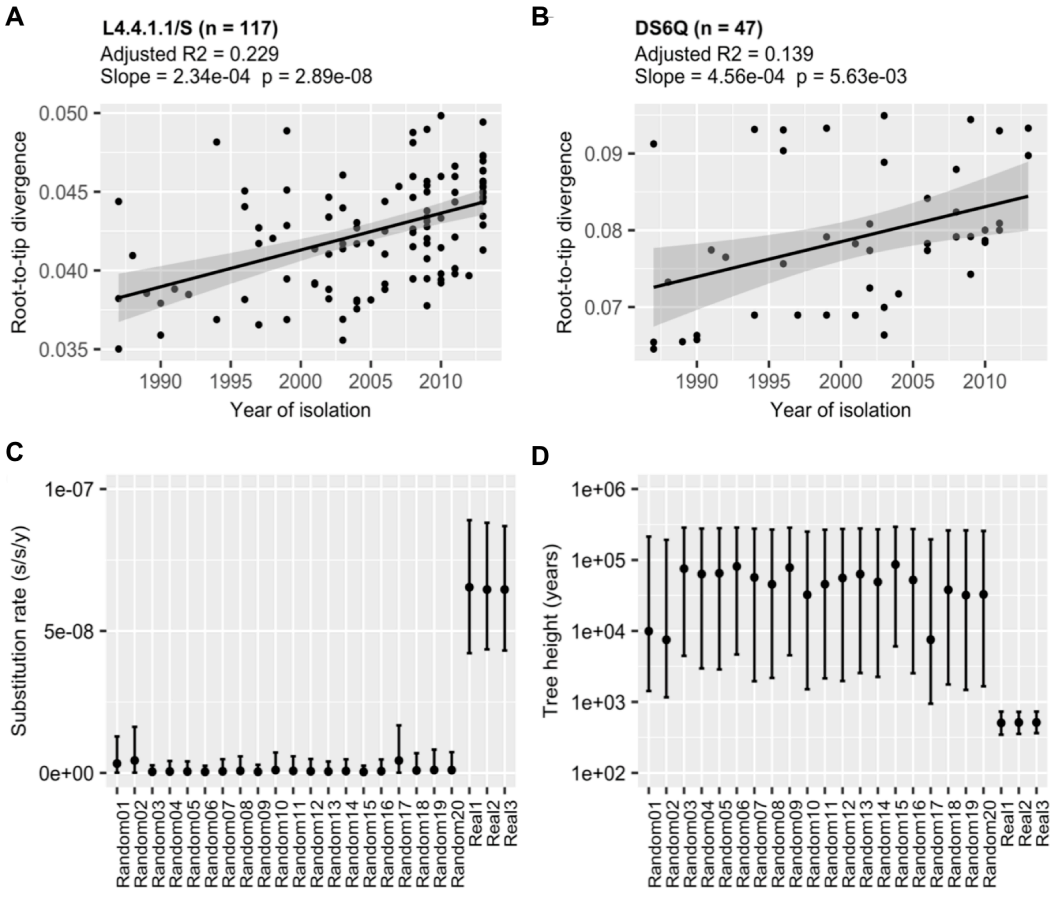


Figure 4.15. *Assessment of temporal signal for tip-dating.* (A,B) Root-to-tip regression analysis for (A) the L4.4.1.1/S dataset and (B) the DS6Q sample subset. (C,D) Tip date randomisation results for the L4.4.1.1/S dataset. Median and 95% HPD intervals for estimates of (C) substitution rate and (D) tree height (in years since 2013) after tip randomisation and for real dates. (GTR, strict clock, constant population demographic model).

4.3.4.3 Molecular dating in BEAST2

Molecular dating analyses were performed in BEAST2 using a 3163 bp SNP alignment generated from 117 L4.4.1.1/S genomes. Analyses were performed using the GTR model of substitution using strict and relaxed molecular clocks, with coalescent constant, exponential and Bayesian Skyline demographic models. Triplicate MCMC chains were run for each model and showed sufficient mixing and ESS values >200 for all parameters for each chain. The different model combinations all produced very similar substitution rates and TMRCA estimates (6.15×10^{-8}–6.64×10^{-8} s/s/y); outermost 95% HPD intervals over all models, 4.23×10^{-8}–9.08×10^{-8}) (Table 4.8). Model comparison determined the strict clock with the Bayesian skyline demographic model best fit the data (Section 4.2.8.3). Under this model, the substitution rate was estimated as 6.28×10^{-8} s/s/y (95% HPD, 4.54×10^{-8}–8.10×10^{-8}) resulting in a TMRCA estimate of 1492 for L4.4.1.1/S (95% HPD, 1325–1629) (Figure 4.16, Figure 4.17). All other parameter estimates reported in the main text are for this model. Trace outputs of key parameters showing convergence of the MCMC chain for replicate runs for this model are shown in Appendix C.6.

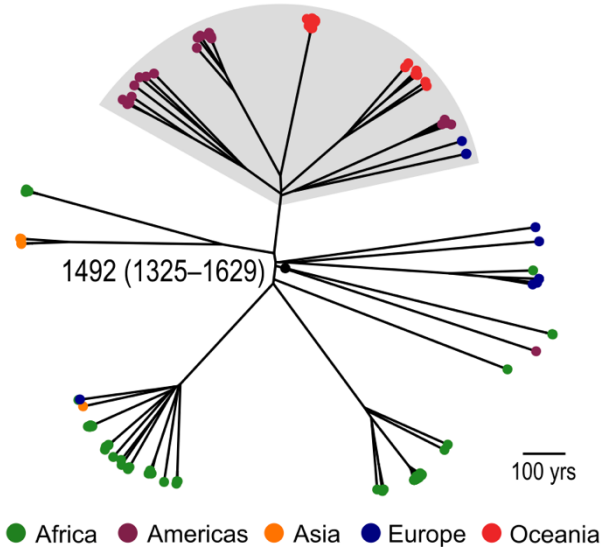


Figure 4.16. *Bayesian phylogeny of the L4.4.1.1/S sublineage*. Maximum clade credibility tree inferred from 117 *M. tuberculosis* L4.4.1.1/S sublineage genomes. A black node marks the most recent common ancestor and the DS6Q is highlighted in grey.

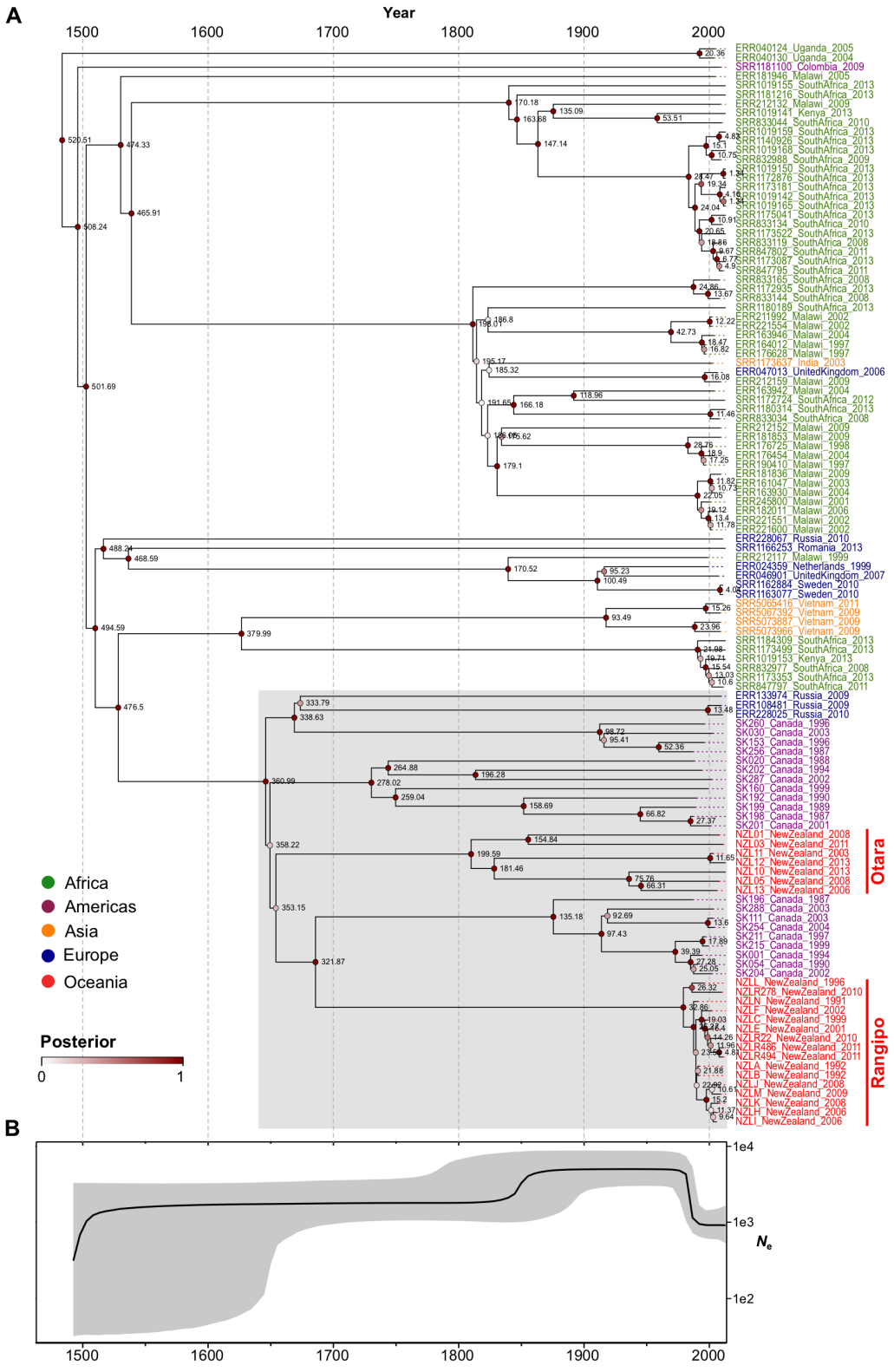


Figure 4.17. *Bayesian phylogeny and population dynamics of the L4.4.1.1/S sublineage.* (A) Maximum clade credibility tree. Individual node heights are shown in years since 2013 and posterior support for each node is indicated by colour. A grey box highlights the DS6Q. (B) Bayesian skyline plot showing change in effective population size (N_e) over time since 1492. 95% HPD interval shown in grey.

Table 4.8. *Substitution rate and TMRCA estimates for the L4.4.1.1/S sublineage*. Estimates are reported for different clock and population demographic models. Median values are reported and the 95% HPD interval is shown in brackets. The GTR model of substitution was used for all analyses. Substitution rate is in substitutions per site per year (s/s/y). The best fitting model determined by path sampling is highlighted in bold type.

Clock	Demographic model	Substitution rate (s/s/y)	L4.4.1.1/S TMRCA	DS6Q TMRCA	Rangipo TMRCA	Otara TMRCA
Strict	Constant	6.63×10^{-8} (4.34×10^{-8}–9.06×10^{-8})	1513 (1301–1673)	1671 (1529–1777)	1978 (1965–1987)	1827 (1744–1886)
Strict	Exponential	6.63×10^{-8} (4.37×10^{-8}–9.08×10^{-8})	1513 (1299–1672)	1672 (1530–1781)	1978 (1965–1986)	1827 (1745–1887)
Strict	**Skyline**	**6.28×10^{-8} (4.54×10^{-8}–8.10×10^{-8})**	**1492 (1325–1629)**	**1652 (1535–1741)**	**1980 (1969–1988)**	**1813 (1746–1868)**
UCLD	Constant	6.49×10^{-8} (4.23×10^{-8} –8.93×10^{-8})	1500 (1277–1667)	1665 (1518–1774)	1977 1964–1986)	1824 (1741–1887)
UCLD	Exponential	6.64×10^{-8} (4.37×10^{-8}–8.98×10^{-8})	1511 (1294–1667)	1673 (1528–1772)	1977 (1965–1986)	1828 (1748–1888)
UCLD	Skyline	6.15×10^{-8} (4.39×10^{-8}–7.98×10^{-8})	1480 (1300–1624)	1645 (1524–1743)	1980 (1968–1988)	1809 (1739–1867)

The population dynamics of the L4.4.1.1/S sublineage and the star-like structure near the root of the phylogeny suggest this sublineage underwent a swift population expansion following its emergence (Figure 4.17B). This was followed by a period where the population size remained consistent until another period of population growth in the 19th century followed by a sharp population decline in the last half of the 20th century.

Our TMRCA estimate of the DS6Q clade was 1652 (95% HPD, 1535–1741) and 1691 for Rangipo and the closest Canadian clade (95% HPD, 1588–1776) (Table 4.8). This coincides with the French migration to Quebec (1608–1760) (Charbonneau *et al.* 1993) and is therefore consistent with a French/European origin of the DS6Q clade. The TMRCA estimate of the Rangipo cluster was 1980 (95% HPD, 1969–1988), indicating that Rangipo is either a relatively recent clonal expansion from a previously introduced DS6Q strain or a recent introduction.

The Otara cluster has a much older TMRCA estimate of 1813 (95% HPD, 1746–1868). As this cluster is most prevalent in Pacific people in New Zealand, the TMRCA and phylogenetic structure suggests this is an endemic strain in this population. Migration to New Zealand from the Pacific Islands occurred at low levels in the 1800s, such that there were only 151 Pacific Island Polynesians in New Zealand by 1916 (StatsNZ 1916). Migration from the Pacific Islands significantly increased in the 1950s–1970s (Dunsford *et al.* 2011) and between 1945 and 1976 the Pacific population in New Zealand increased around 30-fold from 2,159 to ~66,000 (StatsNZ 1996) (Figure 4.3). The phylogeny of the Otara cluster can be explained by initial dispersal to the Pacific Islands from Europe in the early 1800s, followed by several later introductions into New Zealand with more recent Pacific migrations. Long internal branches stretching >100 yrs back in time suggest at least four separate introductions of the Otara strain to New Zealand, while more recent nodes could represent transmission occurring either within New Zealand or the Pacific Islands.

4.3.4.4 Model selection

Path sampling in BEAST2 was used to compare the fit of the various clock and demographic models used in the molecular dating analyses. Based on path sampling MLE, the Bayesian skyline demographic model was found to fit the data best with

the strict clock fitting better than the relaxed clock (Table 4.9). Replicate MLEs for both models were highly congruent (Figure 4.18) and the relaxed clock with the skyline demographic model had a log Bayes factor of 3.4 relative to the strict clock, providing positive evidence in favour of the strict clock. The constant and exponential demographic models all had log Bayes factors >30 relative to the skyline model, providing very strong positive evidence against these models. These models all had very similar MLEs and the variability between MLEs for replicate path sampling runs was larger than the difference between models, reducing confidence in the ranking of the bottom four models. This however was not considered of major concern due to the high confidence and strong statistical support for selection of the top model.

Table 4.9. *Model evaluation using path sampling.* Average marginal likelihood from two replicate runs are reported. Bayes factors are relative to the top ranked model.

Rank	Clock model	Demographic model	Log-marginal likelihood	log Bayes factor
1	Strict	Skyline	–5342225.5	
2	Relaxed	Skyline	–5342228.9	3.4
3	Relaxed	Constant	–5342259.6	34.1
4	Strict	Constant	–5342261.1	35.6
5	Strict	Exponential	–5342262.4	37.0
6	Relaxed	Exponential	–5342263.2	37.7

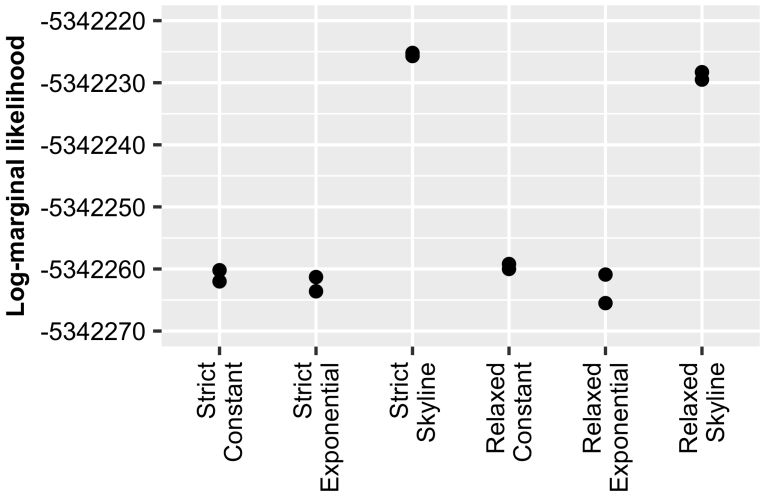


Figure 4.18. *Marginal likelihood estimates for replicate path sampling runs in BEAST2.*

4.3.4.5 Validity of results

To assess the validity and robustness of the results, date randomisation was performed and effects of prior distributions on parameter estimation were investigated. Date randomisation showed no overlap in the 95% HPD intervals between substitution rate estimates obtained from the true dates compared with those obtained from randomised dates demonstrating the sufficient temporal signal in the data (Section 4.2.8.1). MCMC was used to sample from the prior distribution in the absence of data and comparison of marginal posterior and prior distributions showed a strong signal from the data indicating these results are just not an artefact reflecting the prior. To assure unintentional biases were not being introduced into the results through the selection of priors, BEAST2 analyses were run using different prior distributions on the population size prior. Negligible effects on the substitution rate were observed using either the default Jeffreys (1/X prior) or differing upper bounds on uniform prior distributions (1×10^6–1×10^{10}), demonstrating the robustness of the result to this prior specification.

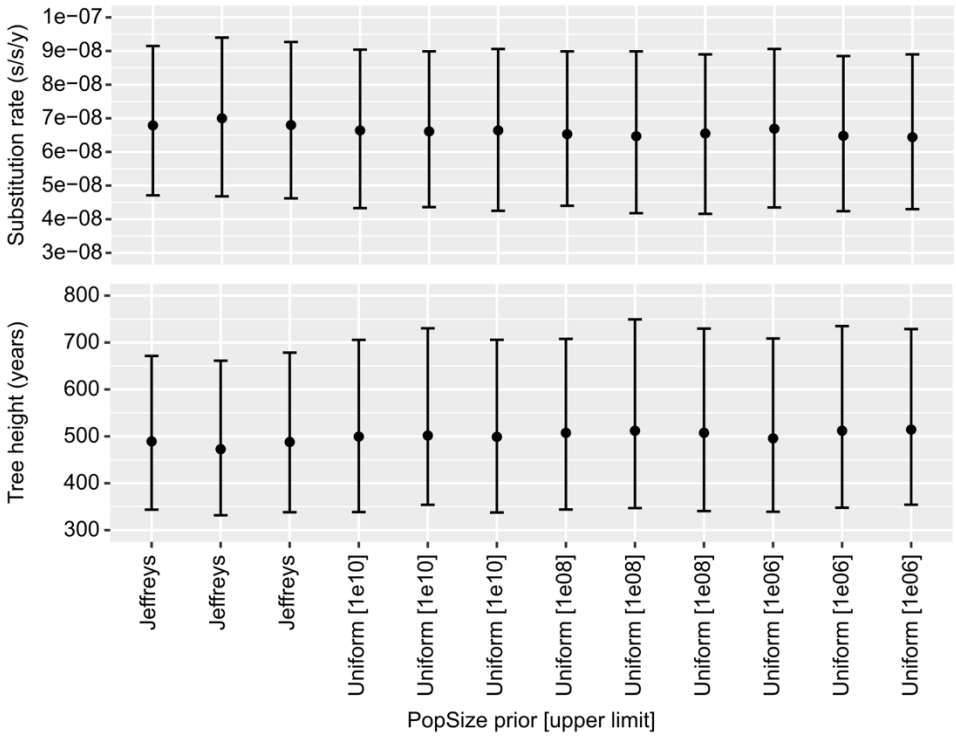


Figure 4.19. *Comparison of parameter estimates using differing population size prior specifications.* (GTR, strict clock, constant population demographic model).

4.4 Discussion

4.4.1 The New Zealand *M. tuberculosis* clusters

The three largest *M. tuberculosis* clusters in New Zealand defined by MIRU-VNTR 24-locus typing are known as the Rangipo, Otara and Southern Cross clusters. High quality WGS sequencing data was available for 16 Rangipo, five Southern Cross and seven Otara clinical *M. tuberculosis* isolates. The Rangipo strain dataset provides a good representative sample of this cluster and shows it harbours little genetic diversity. All Rangipo isolates were all found to belong to a single cluster defined both by MIRU-VNTR typing and SNPs, consistent with relatively recent clonal expansion and transmission. SNP analyses revealed that the Otara strain isolates did not belong to a single WGS cluster and indicated recent transmission between only two of the seven sequenced isolates. Although a smaller number of isolates were sequenced from this cluster, it is apparent it harbours much higher levels of genetic diversity than the Rangipo cluster. Unlike the Rangipo and Southern Cross isolates, which had identical MIRU-VNTR profiles (except for one Rangipo isolate that differed at a single locus), the Otara strains sequenced included more variant MIRU-VNTR profiles. The higher level of genetic diversity among Otara strains however is not simply a result of this and these analyses indicate that the Otara cluster represents an endemic strain in Pacific People rather than a recent clonal expansion. Accordingly, molecular dating analysis of the Otara cluster show this strain was first introduced to the Pacific Islands from Europe in the early 1800s and has migrated to New Zealand several times (further discussed in Section 4.4.2.1). These results also demonstrate the limitation of MIRU-VNTR typing to discriminate recent transmission events, consistent with other recent studies (Bryant *et al.* 2013b; Roetzer *et al.* 2013; Stucki *et al.* 2016a; Walker *et al.* 2013a).

Recent transmission was evident among four of the five Southern Cross isolates and these are presumed to be related to the large outbreak caused by this strain in 1999 (Hill and Calder 2000). A fifth more distantly related Southern Cross isolate hints that this MIRU-VNTR defined cluster might also belong a larger endemic Pacific strain, as for the Otara cluster. However, additional sequencing needs to be undertaken to better characterise the molecular epidemiology of this cluster.

The Rangipo strain is most commonly found in Māori and has been the source of numerous outbreaks over the last thirty years (Calder 2013; Colangeli *et al.* 2014; De Zoysa *et al.* 2001; McElnay *et al.* 2004). The Rangipo strain continues to cause outbreaks, the most recent involving six cases in the Waikato region in 2017/2018 (R. Hoskins, Waikato Hospital, personal communication). Anecdotal evidence suggested that this strain may be highly virulent and contact tracing data from previous Rangipo outbreaks suggest high rates of progression to active disease in Rangipo strain infections (Section 2.2.1).

Illumina sequencing data identified 29 SNPs specific to the Rangipo strain, including an additional 15 SNPs not identified from SOLiD data. The additional SNPs identified likely reflect differences in the sequencing technologies used as well as the bioinformatics processes used to analyse WGS data. Eight of the 29 RS-nsSNPs were found in known virulence associated genes, four of which are predicted to have functional effects on the proteins they encode; Rv0405/*pks6* S1236L, Rv1161/*narG* Y802H, Rv0022c/*whiB5* R43H and Rv2048c/*pks12* R669P. A cluster of Rangipo SNPs was identified that was enriched for proteins, predominantly polyketide synthases, involved in lipid biosynthesis and metabolism and PDIM cell wall layer assembly. This cluster included two putative functional effect SNPs specific to the Rangipo strain; Rv0405/*pks6* S1236L and Rv2048c/*pks12* R669P. The *M. tuberculosis* cell wall is rich in unique lipids that are important for pathogenesis, including the glycolipid PDIM which is a major virulence factor in *M. tuberculosis* (Forrellad *et al.* 2013). Although the function of PDIM is not well understood, it is proposed to have roles in immune evasion and the modulation of host immune responses during the early steps of infection (reviewed in Arbues *et al.* 2014; Stanley and Cox 2013). Biosynthesis of PDIM and other important mycobacterial cell lipids involve polyketide synthases, therefore genetic variants affecting the function of these enzymes could have important phenotypic effects and consequences for strain virulence.

In addition to potential functional effects, SNPs also provide valuable phylogenetic information that can be used to understand the evolution and dispersal of *M. tuberculosis* strains. The remainder of this work used SNP-based evolutionary analyses to gain new insight into historical and social drivers of success of the Rangipo and Otara clusters.

4.4.2 Origins and dispersal of the DS6Q clade in Polynesia

I sought to investigate the origins and historical dispersal of the New Zealand Rangipo and Otara clusters using a phylodynamics approach and a large dataset of global *M. tuberculosis* genomes. A global phylogeny of the L4.4 sublineage identified that the Rangipo and Otara belong to a L4.4.1.1/S clade – the 'DS6Q' clade, that is commonly found in indigenous populations in Canada and New Zealand. Pepperell *et al.* (2011) have shown that this lineage was introduced to Western Canadian indigenous populations via the fur trade and later expanded concomitant with 19th century industrialisation (Pepperell *et al.* 2011). The $DS6^{Quebec}$ lineage accounts for 48% of *M. tuberculosis* isolates in Quebec and is assumed to have been introduced to Canada from France (Pepperell *et al.* 2011), implying a similar European origin for the New Zealand DS6Q clusters. Molecular dating analysis estimated the last common ancestor of the DS6Q clade existed in the mid-17th century which coincides with the French migration to Quebec (1608–1760) (Charbonneau *et al.* 1993), and is thus consistent with a French origin of the clade (Figure 4.20).

Considering the history of early European activity in the South Pacific region and history of European migration to New Zealand, the introduction of DS6Q strains to Polynesia are hypothesised to represent trade-associated introductions resulting from contact with French whalers in the early 19th century (Figure 4.20). During the early 19th century, the only French economic activity of any scale in the South Pacific was the whaling trade (Maclellan 1998), of which the most significant years were 1832–1846 (Foucrier 2005). French whalers had a notable presence in New Zealand and were a significant factor in the British decision to annex New Zealand in 1840. The first French whalers arrived in New Zealand in 1836 and in the years 1840–43 the majority of French whaling voyages included New Zealand (70/81) (Foucrier 2005). French protectorates were also established in the South Pacific starting with Tahiti/French Polynesia in 1842, however there was no large-scale overseas emigration from France to Polynesia.

European whaling in the South Pacific began in the 1790s and was pursued mostly by American, British and French vessels. From 1820–60 the whaling industry was the backbone of Pacific commerce and was dominated by American whalers, although smaller numbers of British and French vessels continued to frequent the

region (Campbell 2011; Fischer 2013). These arrivals were coming from regions where tuberculosis rates were at epidemic proportions; Western Europe tuberculosis mortality rates approached 1000 per 100,000 people per year (Daniel 2006), and the relationships established with local indigenous people during this era provide the necessary social context for transmission of the disease.

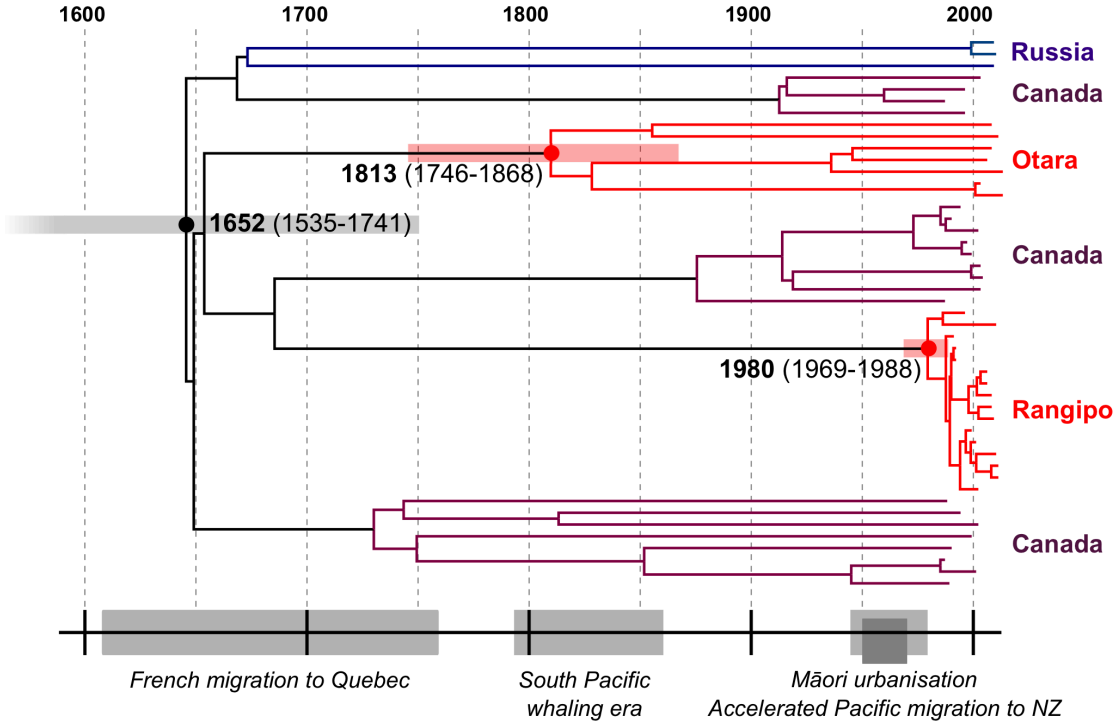


Figure 4.20. *Dated Bayesian phylogeny of the DS6Q clade and historical timeline.* The DS6Q clade from the maximum clade credibility tree inferred from 117 L4.4.1.1/S *M. tuberculosis* genomes is shown. A black node marks the MRCA of the DS6Q clade, and node dates and 95% HPD intervals are shown for the DS6Q clade and the Rangipo and Otara clusters. The historical timeline shows the timing of the French migration to Quebec (1608–1760), South Pacific whaling era (late 1700s–1860s), the rapid urbanisation of Māori (1945–1980) and the surge in migration from the Pacific Islands to New Zealand in the 1950s–1970s.

As with the Canadian fur trade, the formation of productive social and economic relationships with the local people was essential for the commercial success of the whaling industry (Stevens and Wanhalla 2017). The impacts of the southern New Zealand whaling trade on indigenous populations are described as being very similar to those in the fur trade. For example, intermarriage played a central role in industry establishment and success in both the Canadian fur trade and New Zealand whaling (Stevens and Wanhalla 2017). Polynesians were also frequently recruited onboard European and American whaling vessels accounting for up to one-fifth of

crews, and Māori and Tahitians were particularly popular amongst British and French whalers (Chappell 1997; Fischer 2013). These close interactions both on land and on-board would have established strong social ties conducive for the dispersal of *M. tuberculosis*. Accordingly contact with European trade vessels and ports have been implicated in the introduction of tuberculosis and other infectious diseases to the region (Chappell 1997; Lange 1984). In the Cook Islands for example, the locals themselves readily associated European ships as sources of disease, developing the phrase *kua pai au*, which translates to "I am shippy", to pinpoint the source of their illness (Lange 1984).

Limited data is available on *M. tuberculosis* diversity elsewhere in Polynesia. Molecular epidemiology studies using MIRU-VNTR and spoligotyping have found the S-type accounts for 11% of isolates on Kiribati (8/73) (Aleksic *et al.* 2013) and over one-third of *M. tuberculosis* isolates on Tahiti, French Polynesia (10/27) (Osman *et al.* 2017). Tahiti was an especially important commerce hub provisioning European whaling and trade vessels in the early 19th century and was made a French protectorate in 1842 and a colony in 1880. Although no WGS data was available for inclusion in phylogenetic analyses, we speculate whether the S lineage presence in Tahiti might also represent DS6Q clade strains introduced via the same historical trading activities.

In addition to DS6Q strains in Canada and New Zealand, this clade is also represented by isolates from Russia. Unlike New Zealand and Canada where the DS6Q clade occurs at relatively high frequencies, the L4.4 sublineage is rare in Russia (Casali *et al.* 2014; Stucki *et al.* 2016b). Historically, Western Europe and Russia have been culturally and politically more connected and trade dates between them back to ancient times (Öhberg 1955). Thus, DS6Q dispersal to Russia has not been associated with any particular migrations or trade eras.

4.4.2.1 History of the Otara cluster

The early 19th century TMRCA estimate for the Otara cluster is consistent with an introduction to the South Pacific by French/European whalers (Figure 4.20). Although the median of TMRCA estimate slightly predates the peak of French whaling in the region, the 95% HPD falls well within this time period. Unlike New Zealand, which received a massive influx of European settlers from the 1840s, other

Polynesian islands did not experience the same *en masse* arrival of European emigrants, lending further support to this being a trade-associated introduction. The ancestor of the Otara strain was presumably introduced first to the Pacific Islands from Europe, and then later dispersed to New Zealand with more recent Pacific migrations. This highlights the importance of tuberculosis trends in the Pacific islands to those in New Zealand and the need for Pacific-wide TB control efforts.

4.4.2.2 History of the Rangipo cluster

These analyses show that the Rangipo cluster is a clonal expansion that emerged around forty years ago (Figure 4.20). Māori tuberculosis mortality rates declined sharply in the mid-20th century (MacLean 1964) and notification data also show a decline of 83% from 1954 to 1970 (AJHR H-31, 1951-97 in Dunsford 2008) (Figure 4.4). Thus, it is expected modern strain diversity represents just a portion of that circulating prior to the mid-1900s and the observation that one single relatively recent clonal expansion accounts for 25% of tuberculosis in Māori further attests to this bottleneck. Along with falling tuberculosis rates, between 1945–1980 Māori also experienced one of the fastest rates of urbanisation of any population in the world (Pool 1991). This was accompanied by major host environment changes, including overcrowded housing and increased prison incarceration rates, both known tuberculosis risk factors (Baussano *et al.* 2010; Clark *et al.* 2002). The TMRCA of the Rangipo cluster fits with the urbanisation of Māori and the host social changes brought about by this are undoubtedly important factors contributing to the successful expansion and dispersal of this strain.

Prisons represent reservoirs of tuberculosis disease for the general population and prison outbreaks have been linked to increased incidence of the strain in the surrounding community (Jones *et al.* 2003). The Rangipo strain was named for its association with a large outbreak in the 1990s that involved cases who had spent time in the Rangipo prison (De Zoysa *et al.* 2001; McElnay *et al.* 2004). The TMRCA estimate for the Rangipo cluster predates this outbreak, although its introduction into the prison environment has presumably promoted its further spread. Overall, the historic and social factors that have facilitated the success of this strain are themselves products of the colonial history of New Zealand and

therefore Rangipo strain cannot simply be considered a 'prison strain', as its name implies.

French whaling in New Zealand presents a possible source for the introduction of DS6Q strains into New Zealand from which the Rangipo strain could have emerged. Alternatively, Rangipo could be a recent introduction into New Zealand coinciding with changes in host ecology facilitating its successful expansion. The TMRCA of Rangipo follows a period of mass migration to New Zealand from the Pacific Islands in the 1950s-1970s offering a conceivable route. Although it is evident that the Rangipo cluster has ultimately emerged from a strain of European origin, more in-depth sampling of L4.4.1.1/S isolates from both New Zealand and the Pacific may provide a clearer picture of the route this strain to New Zealand and will shed additional light into the dispersal of this sublineage in our region.

4.4.3 Tuberculosis stigma

The naming of the New Zealand clusters as 'Rangipo' and 'Otara' reflect the geographic locations where these strains were first identified. Naming diseases by place of origin can stigmatise the associated population and serves to emphasise the 'other' aspect of disease (Perry and Donini-Lenhoff 2010). Tuberculosis stigma adversely affects health-seeking behaviours resulting in delays in diagnosis, influences treatment adherence and complicates contact tracing efforts, and thus has implications for tuberculosis control efforts (Craig *et al.* 2017). Stigma also increases emotional suffering of patients therefore alleviating the additional burden of stigma is also important for reducing the overall suffering of those affected by tuberculosis.

Both Rangipo and Otara are Māori place names and Otara is home to large population of Pacific People, associating these tuberculosis strains with Māori and Pacific People more generally. Our results show that these strains do not represent Māori or Polynesian *M. tuberculosis* strains, but rather strains that have been introduced and expanded in these populations as a consequence of European contact and colonisation. This highlights the pejorative nature of these names and unfairly scapegoats the affected communities as sources of disease. These findings support the renaming of these clusters to refrain from further perpetuating stigmatisation of

communities where tuberculosis is present and further work will seek to rename these clusters in consultation with local Māori and communities of interest.

4.4.4 The L4.4 sublineage

In addition to defining the historic origins of the New Zealand clusters, these results also provide a characterisation of L4.4 and L4.4.1.1/S sublineage. Recent phylogeographic analyses show the evolutionary history of L4 was characterised by rapid diffusion and high rates of migration, identifying range expansion as a contributor to the growth of this lineage (O'Neill *et al.* 2017). Consistent with this, our results reveal efficient dispersal of L4.4 suggestive of high rates of migration.

As this thesis was in the final stages of preparation, a new study was published by Brynildsrud *et al.* (2018) reconstructing the migratory history of the L4 lineage. This study included isolates from 15 countries in Europe, Africa, the Americas, and Southeast Asia, but none from the South Pacific. Global dispersal of L4 was found to be dominated by historical migrations out of Europe and dispersal of L4 to Africa and the Americas occurred concomitant with European colonial migrations (Brynildsrud *et al.* 2018). We observe the same scenario with the introduction of L4.4 to the South Pacific and the DS6Q clade provides a striking example of the role of European expansion in the global dispersal of *M. tuberculosis*. Our analyses reveal the migration of several closely related DS6Q strains out of Europe in the 17th–19th centuries to remote and unconnected populations driven by expanding European trade networks and colonial migrations. The nearest branch neighbouring the DS6Q on our phylogeny splits into two nodes comprised of Vietnamese and South African descendants which share a MRCA in the mid-17th century (Figure 4.17). This split coincides with the first planned migration of French Huguenots to South Africa in 1687 (van Ruymbeke *et al.* 2003) and the onset of French-Vietnamese interactions starting in 1627 (Chapuis 1995). Additionally, the node age of the Vietnamese clade fits with the French Indochina war in the 20th century, suggesting these are also likely French introductions.

The emergence and initial growth of L4.4.1.1/S in the 16th century is coincident with the European age of exploration (Arnold 2013) providing a plausible factor that may have contributed to the growth and dispersal of this sublineage. This era would have provided ample opportunity for dispersal from Europeans to other

populations and geographic range expansion followed by population expansion in the L4 lineage has recently been linked to this era (O'Neill *et al.* 2017). We also detect L4.4.1.1/S population growth in the 19th century. This could be attributable to various colonial activities around this time involving countries represented in our sample; the French-Canadian fur trade (1730–1870) (Innis 1999), the South Pacific whaling era (1790s–1860) (Campbell 2011), and the rapid occupation and colonisation of much of the African continent during the "Scramble for Africa" (1876–1912) (Pakenham 2015). A later population decline in the late 20th century coincides with the dramatic decline in tuberculosis incidence in the developed world over the last century.

4.4.4.1 L4.4.1.1/S sublineage substitution rate

Molecular dating analysis estimated a substitution rate for the L4.4.1.1/S lineage of 6.3 x 10^{-8} s/s/y (95% HPD, 4.5 x 10^{-8} – 8.1 x 10^{-8}). This is consistent with rates inferred by previous studies of modern L4 and mixed lineage MTBC genomes (median rates estimates, 7 x 10^{-8} – 1 x 10^{-7} s/s/y (Eldholm *et al.* 2015; Eldholm *et al.* 2016; Ford *et al.* 2011; Ford *et al.* 2013; Pepperell *et al.* 2013; Roetzer *et al.* 2013; Walker *et al.* 2013a)); and slightly higher than long term mutation rates of ~5 x 10^{-8} s/s/y estimated from ancient DNA studies (Bos *et al.* 2014; Kay *et al.* 2015) (Figure 4.21). Our rate is markedly higher than the long-term substitution rate inferred on the assumption of co-divergence of the MTBC and humans (the 'out of Africa' hypothesis; 2.6 x 10^{-9} s/s/y) (Comas *et al.* 2013).

Colengali et al (2013) have previously estimated a substitution rate from Rangipo strain isolates using SNP differences between recently transmitted isolate pairs and reactivation cases. A rate of 5.5 x 10^{-10} substitutions/site/generation was estimated for recently transmitted tuberculosis (assuming a 20 hr generation time) and 7.3 x 10^{-11} substitutions/site/generation for latent tuberculosis. This translates to a mean rate of ~1.4 x 10^{-7} s/s/y (my own calculation), which is slightly higher than our estimate for the L4.4.1.1/S lineage, but is still within range of other L4 rate estimates. This estimate however is limited, as the time from infection to active disease can vary markedly between cases, therefore the mean of the recent and latent rates does not necessarily provide a good comparison with ours or other substitution rates.

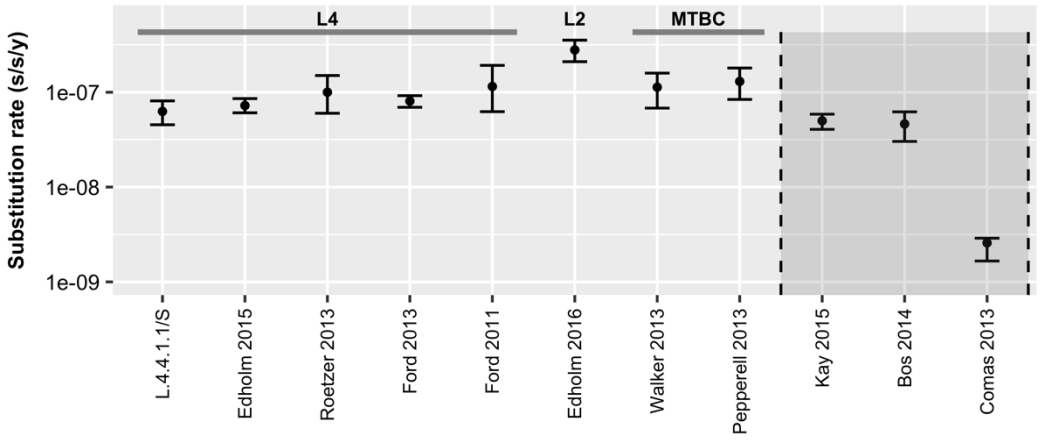


Figure 4.21. *Comparison of substitution rate estimated in this study with previously published studies.* Estimates of long-term substitution rate estimates are shown in the grey box. Short-term rate estimates from human *M. tuberculosis* isolates were calibrated using tip-dating (this study (L4.4.1.1/S), (Eldholm *et al.* 2015; Eldholm *et al.* 2016; Ford *et al.* 2013; Roetzer *et al.* 2013), experimental infection of macaques (Ford *et al.* 2011), recent transmission events (Walker *et al.* 2013a), and historic events (Pepperell 2013). Long-term rate estimates were calibrated using aDNA (Bos *et al.* 2014; Kay *et al.* 2015) or human divergence events (Comas *et al.* 2013).

While ours and other short-term substitution rate estimates for the reported MTBC are all highly similar, the long-term substitution rate and age of the MTBC remains debated and studies using different approaches have produced markedly different rate and date estimates. The out of Africa hypothesis postulates *M. tuberculosis* diversity and dispersal was shaped by human migrations out of Africa and suggest the last common ancestor of the MTBC existed around seventy thousand years ago (Comas *et al.* 2013). In contrast, two separate studies using ancient DNA from different time periods both produced 10-fold slower substitution rate estimates suggesting the last common ancestor of the MTBC existed less than six thousand years ago (Bos *et al.* 2014; Kay *et al.* 2015). Archaeological samples also indicate that tuberculosis is a much more ancient disease (Hershkovitz *et al.* 2008; Rothschild *et al.* 2001), although a more recent MRCA could be reconciled with a scenario of whole scale clonal replacement of ancient strains.

4.4.5 Study Limitations

One limitation of this study is that the global isolates in our dataset were primarily compiled from isolates available in public repositories and not all countries where the L4.4 lineage is present are represented. The addition of further isolates from under-represented regions and within the Pacific would further strengthen the sample and provide a more detailed picture of L4.4 and DS6Q dispersal. Genomes were annotated with country of isolation, as recorded in databases and published sources. Isolates from countries where many cases occur in migrants could potentially reflect reactivation of infection contracted in the country of birth. However, this was kept in mind throughout the analyses and is not a concern for the clades/countries on which the conclusions of this work have been based.

4.4.6 Conclusions

Using a large, global WGS dataset and a phylodynamics approach I have investigated the origins and historical dispersal of the Rangipo and Otara clusters. Historical and social factors facilitating the expansion and dispersal of these strains have been identified, showing that they are a product of the colonial history of New Zealand and the South Pacific and highlight the importance of Pacific-wide tuberculosis control efforts. These results also demonstrate the power of phylodynamics approaches to study *M. tuberculosis* dispersal at both the global and local scale and provide valuable new insights into human social phenomena underlying the global dispersal of this globally successful bacterial pathogen.

Chapter Five

Conclusions and future directions

The overall aim of this research was to understand factors contributing to the success of prevalent *Mycobacterium tuberculosis* clusters in New Zealand, in particular the Rangipo cluster. Ultimately, this information could be used to better control the circulation of these strains. The Rangipo cluster has been circulating in New Zealand for around 30 years and continues to cause outbreaks, predominantly in the Māori population. This prompts the question – why is this strain so successful and seemingly highly transmissible in this population? And how can it be stopped? My research investigated evolutionary and functional aspects of genomic variation in the Rangipo cluster to address this question.

This thesis identifies bacterial genetic variants that may contribute to the high transmissibility of the Rangipo strain and points to an important role for historical factors and host ecology changes in driving the success of this strain. The prevalence of this cluster is not solely due to this being a 'hypervirulent strain' but rather, likely the result of intersecting bacterial and host determinants, underpinned by the colonial history of New Zealand. Furthermore, this work demonstrates the use of whole genome sequencing (WGS) data in a multi-disciplinary approach to investigate different factors contributing to the success of prevalent *M. tuberculosis* strains, as well the practical applications of single nucleotide polymorphisms (SNPs) in a clinical setting. The SNP-based Rangipo strain specific diagnostic assay developed here enables accurate strain classification directly from sputum samples within 24 hours for minimal cost and with high discriminatory power. Being a SNP-based approach, typing for this Rangipo strain marker could be readily automated and incorporated into routine clinical practice. The rapid identification of this strain will enable clinicians and public health authorities to implement appropriate strategies to halt its further transmission.

Elucidating the functional consequences of genetic variation in the *M. tuberculosis* complex is important to understand how strain genetic background affects virulence and the outcome of infection. Several SNPs with putative functional effects specific to the Rangipo strain have been identified, providing several potential candidates for further investigation. An enrichment of non-synonymous SNPs in genes involved in lipid metabolism were identified. Lipid metabolism plays an important role in *M. tuberculosis* pathogenicity and presents a possible avenue for further investigation into molecular determinants of Rangipo virulence. However, in order to properly understand of the role of genetic background to the high transmissibility of the Rangipo strain, experimental virulence testing needs to be performed to validate its putative high virulence and to correlate genetic variants with phenotypic effects.

The structures of the luciferase-like hydride transferase family protein Rv2893 experimentally validate this as an F_{420}-binding protein and identify additional intramolecular hydrogen bonds introduced by the Rangipo G72S mutation. The additional thermal stability produced by this mutation and its location at the base of the bulge induced by the conserved non-prolyl *cis*-peptide may influence F_{420} binding and catalysis. The consequences of this mutation on enzyme activity and bacterial phenotype remain to be determined and will largely be dependent on the role of Rv2893 in *M. tuberculosis*. Further investigation is needed to elucidate the function of Rv2893 and for development of biochemical assays to determine the effect of the G72S mutation on catalysis. As a starting point, identifying the unexpected ligand found in Rv2893 structures will provide clues as to the function of this protein. Other approaches such as gene deletion of the Rv2893 homologue in *Mycobacterium smegmatis* and metabolomic profiling, will provide possible routes to further probe the function of this protein. Binding assays also need to be undertaken to determine the kinetics of F_{420} binding and how the G72S mutation affects this.

The New Zealand Rangipo and Otara strains are predominantly found in Māori and Pacific People, respectively. My research shows these are unequivocally of European origin and are linked to European trade expansion and colonisation. These clusters have differing demographic histories and highlight the relevance to New Zealand of tuberculosis in the Pacific islands. Understanding Pacific wide

transmission dynamics will be vital for the improvement of strategies seeking to reduce the tuberculosis burden in the region.

WGS is now sufficiently fast and affordable that it is possible to sequence large collections of *M. tuberculosis* isolates. Further sequencing of clinical isolates will provide a wealth of additional information about tuberculosis transmission in New Zealand. Clinical isolate collections held at Waikato Hospital and LabPLUS will be valuable resources for such further work. Sequencing these collections would provide the opportunity to perform a detailed characterisation of strain diversity and transmission of tuberculosis in New Zealand at high resolution. Understanding local tuberculosis trends and transmission dynamics has important implications for public health and is essential for improvement of strategies seeking to arrest the transmission of local strains. Additional L4.4.1.1/S sublineage genomes will also further enhance our understanding of how this sublineage dispersed and expanded in New Zealand and the South Pacific.

Importantly, this work demonstrates that these strains are products of the colonial history of New Zealand, highlighting the inappropriate and pejorative naming of these strains with Māori names. Addressing the transmission of these and other tuberculosis strains in New Zealand, requires a multi-disciplinary approach and necessitates consultation with affected communities. Further work will seek to rename these strains in consultation with Māori and other groups of interest. This will be important for community engagement and a step towards addressing the stigma that can hamper tuberculosis control efforts and exacerbates the suffering of those affected by the disease.

References

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., et al. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr B*. 68:352.

Aguilar, D., Hanekom, M., Mata, D., Gey van Pittius, N.C., van Helden, P.D., et al. (2010). Mycobacterium tuberculosis strains with the Beijing genotype demonstrate variability in virulence associated with transmission. *Tuberculosis (Edinb)*. 90:319.

Ahmed, F.H., Carr, P.D., Lee, B.M., Afriat-Jurnou, L., Mohamed, A.E., et al. (2015). Sequence–structure–function classification of a catalytically diverse oxidoreductase superfamily in Mycobacteria. *J Mol Biol*. 427:3554.

Akey, D.L., Terwilliger, T.C., and Smith, J.L. (2016). Efficient merging of data from multiple samples for determination of anomalous substructure. *Acta Crystallogr D*. 72:296.

Al-Orainey, I.O. (2009). Diagnosis of latent tuberculosis: Can we do better? *Ann Thorac Med*. 4:5.

Aleksic, E., Merker, M., Cox, H., Reiher, B., Sekawi, Z., et al. (2013). First molecular epidemiology study of Mycobacterium tuberculosis in Kiribati. *PLOS ONE*. 8:e55423.

Allix-Beguec, C., Harmsen, D., Weniger, T., Supply, P., and Niemann, S. (2008). Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of Mycobacterium tuberculosis complex isolates. *Clin Microbiol*. 46:2692.

Alonso, H., Samper, S., Martin, C., and Otal, I. (2013). Mapping IS6110 in high-copy number Mycobacterium tuberculosis strains shows specific insertion points in the Beijing genotype. *BMC Genomics*. 14:422.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol*. 215:403.

Anderson, J., Jarlsberg, L.G., Grindsdale, J., Osmond, D., Kawamura, M., et al. (2013). Sublineages of lineage 4 (Euro-American) Mycobacterium tuberculosis differ in genotypic clustering. *Int J Tuberc Lung Dis*. 17:885.

Andrews, S. (2010). 'FastQC: a quality control tool for high throughput sequence data', http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Arbues, A., Lugo-Villarino, G., Neyrolles, O., Guilhot, C., and Astarie-Dequeker, C. (2014). Playing hide-and-seek with host macrophages through the use of mycobacterial cell envelope phthiocerol dimycocerosates and phenolic glycolipids. *Front Cell Infect Microbiol*. 4:173.

Arnold, D. (2013). The age of discovery, 1400-1600. Routledge.

Aufhammer, S.W., Warkentin, E., Berk, H., Shima, S., Thauer, R.K., et al. (2004). Coenzyme binding in F420-dependent secondary alcohol dehydrogenase, a member of the bacterial luciferase family. *Structure*. 12:361.

Aufhammer, S.W., Warkentin, E., Ermler, U., Hagemeier, C.H., Thauer, R.K., et al. (2005). Crystal structure of methylenetetrahydromethanopterin reductase (Mer) in complex with coenzyme F420: Architecture of the F420/FMN binding

site of enzymes within the nonprolyl cis-peptide containing bacterial luciferase family. *Protein Sci*. 14:1840.

Aung, H.L., Tun, T., Moradigaravand, D., Koser, C.U., Nyunt, W.W., et al. (2016). Whole-genome sequencing of multidrug-resistant Mycobacterium tuberculosis isolates from Myanmar. *J Glob Antimicrob Resist*. 6:113.

Ayres, D.L., Darling, A., Zwickl, D.J., Beerli, P., Holder, M.T., et al. (2012). BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol*. 61:170.

Baele, G., Li, W.L., Drummond, A.J., Suchard, M.A., and Lemey, P. (2013). Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol Biol Evol*. 30:239.

Bair, T.B., Isabelle, D.W., and Daniels, L. (2001). Structures of coenzyme F420 in Mycobacterium species. *Arch Microbiol*. 176:37.

Baker, L., Brown, T., Maiden, M.C., and Drobniewski, F. (2004). Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis. *Emerg Infect Dis*. 10:1568.

Baker, M.G., Barnard, L.T., Kvalsvig, A., Verrall, A., Zhang, J., et al. (2012). Increasing incidence of serious infectious diseases and inequalities in New Zealand: a national epidemiological study. *The Lancet*. 379:1112.

Bashiri, G., Perkowski, E.F., Turner, A.P., Feltcher, M.E., Braunstein, M., et al. (2012). Tat-dependent translocation of an F420-binding protein of Mycobacterium tuberculosis. *PLOS ONE*. 7:e45003.

Bashiri, G., Rehan, A.M., Greenwood, D.R., Dickson, J.M., and Baker, E.N. (2010). Metabolic engineering of cofactor F420 production in Mycobacterium smegmatis. *PLOS ONE*. 5:e15803.

Bashiri, G., Squire, C.J., Baker, E.N., and Moreland, N.J. (2007). Expression, purification and crystallization of native and selenomethionine labeled Mycobacterium tuberculosis FGD1 (Rv0407) using a Mycobacterium smegmatis expression system. *Protein Expr Purif*. 54:38.

Bashiri, G., Squire, C.J., Moreland, N.J., and Baker, E.N. (2008). Crystal structures of F420-dependent glucose-6-phosphate dehydrogenase FGD1 involved in the activation of the anti-tuberculosis drug candidate PA-824 reveal the basis of coenzyme and substrate binding. *J Biol Chem*. 283:17531.

Battye, T.G.G., Kontogiannis, L., Johnson, O., Powell, H.R., and Leslie, A.G. (2011). iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr D*. 67:271.

Baussano, I., Williams, B.G., Nunn, P., Beggiato, M., Fedeli, U., et al. (2010). Tuberculosis incidence in prisons: a systematic review. *PLoS Med*. 7:e1000381.

Biek, R., Pybus, O.G., Lloyd-Smith, J.O., and Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends Ecol Evol*. 30:306.

Bissielo, A., Lim, E., and Heffernan, H. (2012). Tuberculosis in New Zealand: Annual Report 2011. Institute of Environmental Science and Research Ltd (ESR), Porirua, N.Z.

Bjorn-Mortensen, K., Soborg, B., Koch, A., Ladefoged, K., Merker, M., et al. (2016). Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. *Sci Rep*. 6:33180.

Boggon, T.J., and Shapiro, L. (2000). Screening for phasing atoms in protein crystallography. *Structure*. 8:R143.

Borrell, S., and Gagneux, S. (2009). Infectiousness, reproductive fitness and evolution of drug-resistant Mycobacterium tuberculosis. *Int J Tuberc Lung Dis*. 13:1456.

Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., et al. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*. 514:494.

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C.H., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 10:e1003537.

Brennan, P.J. (2003). Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. *Tuberculosis (Edinb)*. 83:91.

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

Bromberg, Y., and Rost, B. (2009). Correlating protein function and stability through the analysis of single amino acid substitutions. *BMC Bioinformatics*. 10:S8.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., et al. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc Natl Acad Sci U S A*. 99:3684.

Brosch, R., Philipp, W.J., Stavropoulos, E., Colston, M.J., Cole, S.T., et al. (1999). Genomic analysis reveals variation between Mycobacterium tuberculosis H37Rv and the attenuated M. tuberculosis H37Ra strain. *Infect Immun*. 67:5768.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., et al. (2006). Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol*. 6:23.

Bryant, J.M., Harris, S.R., Parkhill, J., Dawson, R., Diacon, A.H., et al. (2013a). Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *Lancet Respir Med*. 1:786.

Bryant, J.M., Schurch, A.C., van Deutekom, H., Harris, S.R., de Beer, J.L., et al. (2013b). Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infect Dis*. 13:110.

Brynildsrud, O.B., Pepperell, C.S., Suffys, P., Grandjean, L., Monteserin, J., et al. (2018). Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. *Sci Adv*. 4:eaat5869.

Calder, L. (2013). Tuberculosis Outbreak Among Hawke's Bay Maori. *Australasian Tuberculosis Conference*. Auckland, N.Z.

Camacho, L.R., Ensergueix, D., Perez, E., Gicquel, B., and Guilhot, C. (1999). Identification of a virulence gene cluster of Mycobacterium tuberculosis by signature-tagged transposon mutagenesis. *Mol Microbiol*. 34:257.

Campbell, I.C. (2011). Worlds apart : a history of the Pacific Islands. Canterbury University Press; Christchurch, N.Z.

Campuzano, J., Aguilar, D., Arriaga, K., Leon, J.C., Salas-Rangel, L.P., et al. (2007). The PGRS domain of Mycobacterium tuberculosis PE_PGRS Rv1759c antigen is an efficient subunit vaccine to prevent reactivation in a murine model of chronic tuberculosis. *Vaccine*. 25:3722.

Camus, J.C., Pryor, M.J., Medigue, C., and Cole, S.T. (2002). Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology*. 148:2967.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S.R., Ignatyeva, O., et al. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet*. 46:279.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Ignatyeva, O., Kontsevaya, I., et al. (2012). Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome research*.

CDC (2017). Reported Tuberculosis in the United States, 2016. Atlanta, GA

Chappell, D.A. (1997). Double ghosts : Oceanian voyagers on Euroamerican ships. M.E. Sharpe, Inc.; Armonk, New York, U.S.A.

Chapuis, O. (1995). A history of vietnam: from Hong Bang to Tu Duc. Cambridge University Press.

Charbonneau, H., Boleda, M., and Bates, R.L. (1993). The First French Canadians: Pioneers in the St. Lawrence Valley. University of Delaware Press; Newark, NJ.

Cheeseman, P., Toms-Wood, A., and Wolfe, R.S. (1972). Isolation and properties of a fluorescent compound, factor 420 , from Methanobacterium strain M.o.H. *J Bacteriol*. 112:527.

Cheng, S.J., Thibert, L., Sanchez, T., Heifets, L., and Zhang, Y. (2000). pncA mutations as a major mechanism of pyrazinamide resistance in Mycobacterium tuberculosis: spread of a monoresistant strain in Quebec, Canada. *Antimicrob Agents Chemother*. 44:528.

Chernyaeva, E.N., Shulgina, M.V., Rotkevich, M.S., Dobrynin, P.V., Simonov, S.A., et al. (2014). Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. *BMC Genomics*. 15:308.

Chihota, V. (2011). The Molecular Epidemiology of *Mycobacterium tuberculosis*: Role in understanding disease dynamics in high prevalence settings in Southern Africa Region. PhD thesis, Stellenbosch University, South Africa.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*. 6:80.

Cirillo, J., Weisbrod, T., William, R., and Jacobs, J. (1993). Efficient electrotransformation of *Mycobacterium smegmatis*. Bio-Rad Laboratories; Richmond, California

Clark, M., Riben, P., and Nowgesic, E. (2002). The association of housing density, isolation and tuberculosis in Canadian First Nations communities. *Int J Epidemiol*. 31:940.

Clark, T.G., Mallard, K., Coll, F., Preston, M., Assefa, S., et al. (2013). Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. *PLOS ONE*. 8:e83012.

Colangeli, R., Arcus, V.L., Cursons, R.T., Ruthe, A., Karalus, N., et al. (2014). Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans. *PLOS ONE*. 9:e91024.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., et al. (1998a). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*. 393:537.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., et al. (1998b). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*. 393:537.

Coll, F., McNerney, R., Guerra-Assuncao, J.A., Glynn, J.R., Perdigao, J., et al. (2014). A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat Commun*. 5:4812.

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., et al. (2012). Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet*. 44:106.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., et al. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet*. 45:1176.

Comas, I., and Gagneux, S. (2009). The past and future of tuberculosis research. *PLoS Pathog*. 5:e1000600.

Comas, I., Homolka, S., Niemann, S., and Gagneux, S. (2009). Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. *PLOS ONE*. 4:e7815.

Cook, G.M., Berney, M., Gebhard, S., Heinemann, M., Cox, R.A., et al. (2009). Physiology of mycobacteria. *Adv Microb Physiol*. 55:81.

Coscolla, M. (2017). Biological and Epidemiological Consequences of MTBC Diversity. In *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control*. Springer. 95.

Coscolla, M., and Gagneux, S. (2014). Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin Immunol*. 26:431.

Cowan, L.S., Mosher, L., Diem, L., Massey, J.P., and Crawford, J.T. (2002). Variable-number tandem repeat typing of Mycobacterium tuberculosis isolates with low copy numbers of IS6110 by using mycobacterial interspersed repetitive units. *J Clin Microbiol*. 40:1592.

Craig, G.M., Daftary, A., Engel, N., O'Driscoll, S., and Ioannaki, A. (2017). Tuberculosis stigma as a social determinant of health: a systematic mapping review of research in low incidence countries. *Int J Infect Dis*. 56:90.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., et al. (2011). The variant call format and VCFtools. *Bioinformatics*. 27:2156.

Daniel, J., Maamar, H., Deb, C., Sirakova, T.D., and Kolattukudy, P.E. (2011). Mycobacterium tuberculosis uses host triacylglycerol to accumulate lipid droplets and acquires a dormancy-like phenotype in lipid-loaded macrophages. *PLoS Pathog*. 7:e1002093.

Daniel, R.M., Cowan, D.A., Morgan, H.W., and Curran, M.P. (1982). A correlation between protein thermostability and resistance to proteolysis. *Biochem J*. 207:641.

Daniel, T.M. (2006). The history of tuberculosis. *Respir Med*. 100:1862.

Daniels, L., Bakhiet, N., and Harmon, K. (1985). Widespread Distribution of a 5-deazaflavin Cofactor in Actinomyces and Related Bacteria. *Syst Appl Microbiol*. 6:12.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods*. 9:772.

De Jong, B.C., Antonio, M., and Gagneux, S. (2010). Mycobacterium africanum—review of an important cause of human tuberculosis in West Africa. *PLoS Negl Trop Dis*. 4:e744.

De Jong, B.C., Hill, P.C., Aiken, A., Awine, T., Antonio, M., et al. (2008). Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. *J Infect Dis*. 198:1037.

De Vos, M., Müller, B., Borrell, S., Black, P., Van Helden, P., et al. (2013). Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrob Agents Chemother*. 57:827.

De Zoysa, R., Shoemack, P., Vaughan, R., and Vaughan, A. (2001). A prolonged outbreak of tuberculosis in the North Island. *N Z Public Health Rep*. 8.

Demay, C., Liens, B., Burguière, T., Hill, V., Couvin, D., et al. (2012). SITVITWEB–a publicly available international multimarker database for studying Mycobacterium tuberculosis genetic diversity and molecular epidemiology. *Infect Genet Evol*. 12:755.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 43:491.

Dormans, J., Burger, M., Aguilar, D., Hernandez-Pando, R., Kremer, K., et al. (2004). Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different Mycobacterium tuberculosis genotypes in a BALB/c mouse model. *Clin Exp Immunol*. 137:460.

Dow, D.A. (1999). Maori health & government policy 1840 - 1940. Victoria University Press Wellington, N.Z.

Drobniewski, F., Balabanova, Y., Nikolayevsky, V., Ruddy, M., Kuznetzov, S., et al. (2005). Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. *JAMA*. 293:2726.

Drummond, A.J., Pybus, O.G., Rambaut, A., Forsberg, R., and Rodrigo, A.G. (2003). Measurably evolving populations. *Trends Ecol Evol*. 18:481.

Duchêne, S., Duchêne, D., Holmes, E.C., and Ho, S.Y. (2015). The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol Biol Evol*. 32:1895.

Dunsford, D., Park, J., Littleton, J., Friesen, W., Herda, P., et al. (2011). Better lives: the struggle for health of transnational pacific peoples in New Zealand, 1950-2000. Department of Anthropology, University of Auckland.

Dunsford, D.A. (2008). SEEKING THE PRIZE OF ERADICATION: A social history of tuberculosis in New Zealand from World War Two to the 1970s. Doctor of Philosophy in History thesis, The University of Auckland.

Durie, M. (1998). Whaiora : Maōri health development. Oxford University Press; Auckland, N.Z.

Ehrt, S., and Schnappinger, D. (2009). Mycobacterial survival strategies in the phagosome: defence against host stresses. *Cell Microbiol*. 11:1170.

Eirich, L.D., Vogels, G.D., and Wolfe, R.S. (1978). Proposed structure for coenzyme F420 from Methanobacterium. *Biochemistry*. 17:4583.

Eker, A., Pol, A., Van der Meyden, P., and Vogels, G. (1980). Purification and properties of 8-hydroxy-5-deazaflavin derivatives from Streptomyces griseus. *FEMS Microbiol Lett*. 8:161.

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., et al. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nat Commun*. 6:7119.

Eldholm, V., Pettersson, J.H., Brynildsrud, O.B., Kitchen, A., Rasmussen, E.M., et al. (2016). Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A*. 113:13881.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D*. 60:2126.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. *Acta Crystallogr D*. 66:486.

England, P.H. (2016). Tuberculosis in England: 2016. London, U.K.

Ericsson, U.B., Hallberg, B.M., DeTitta, G.T., Dekker, N., and Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem*. 357:289.

Espitia, C., Laclette, J.P., Mondragon-Palomino, M., Amador, A., Campuzano, J., et al. (1999). The PE-PGRS glycine-rich proteins of Mycobacterium tuberculosis: a new family of fibronectin-binding proteins? *Microbiology*. 145 (Pt 12):3487.

ESR (2015a). Tuberculosis in New Zealand: Annual Report 2013. Institute of Environmental Science and Research Ltd (ESR), Porirua, N.Z.

ESR (2015b). Tuberculosis in New Zealand: Annual Report 2014. Institute of Environmental Science and Research Ltd (ESR), Porirua, N.Z.

ESR (2018). Tuberculosis in New Zealand: Annual Report 2015. Institute of Environmental Science and Research Ltd (ESR), Porirua, N.Z.

Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallogr D*. 62:72.

Evans, P.R., and Murshudov, G.N. (2013). How good are my data and what is the resolution? *Acta Crystallogr D*. 69:1204.

Fenner, L., Egger, M., Bodmer, T., Furrer, H., Ballif, M., et al. (2013). HIV infection disrupts the sympatric host-pathogen relationship in human tuberculosis. *PLoS Genet*. 9:e1003318.

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbon, M.H., et al. (2006). Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol*. 188:759.

Firdessa, R., Berg, S., Hailu, E., Schelling, E., Gumi, B., et al. (2013). Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerg Infect Dis*. 19:460.

Fischer, S.R. (2013). A history of the Pacific Islands. Palgrave Macmillan; Basingstoke, Hampshire.

Fisher, M.A., Plikaytis, B.B., and Shinnick, T.M. (2002). Microarray analysis of the Mycobacterium tuberculosis transcriptional response to the acidic conditions found in phagosomes. *J Bacteriol*. 184:4025.

Fleischmann, R., Alland, D., Eisen, J.A., Carpenter, L., White, O., et al. (2002). Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. *J Bacteriol*. 184:5479.

Folkvardsen, D.B., Norman, A., Andersen, Å.B., Rasmussen, E.M., Lillebaek, T., et al. (2018). A Major Mycobacterium tuberculosis outbreak caused by one specific genotype in a low-incidence country: Exploring gene profile virulence explanations. *Sci Rep*. 8:11869.

Ford, C.B., Lin, P.L., Chase, M.R., Shah, R.R., Iartchouk, O., et al. (2011). Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet*. 43:482.

Ford, C.B., Shah, R.R., Maeda, M.K., Gagneux, S., Murray, M.B., et al. (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat Genet*. 45:784.

Forrellad, M.A., Klepp, L.I., Gioffre, A., Sabio y Garcia, J., Morbidoni, H.R., et al. (2013). Virulence factors of the Mycobacterium tuberculosis complex. *Virulence*. 4:3.

Foucrier, A. (2005). The French and the Pacific world, 17th-19th centuries : explorations, migrations and cultural exchanges. Ashgate Variorum; Aldershot, Hampshire.

Fox, G.J., Barry, S.E., Britton, W.J., and Marks, G.B. (2013). Contact investigation for tuberculosis: a systematic review and meta-analysis. *Eur Respir J*. 41:140.

Frothingham, R., Hills, H.G., and Wilson, K.H. (1994). Extensive DNA-Sequence Conservation Throughout the Mycobacterium-Tuberculosis Complex. *J Clin Microbiol*. 32:1639.

Gagneux, S. (2012). Host-pathogen coevolution in human tuberculosis. *Philos Trans R Soc Lond B Biol Sci*. 367:850.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., et al. (2006a). Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A*. 103:2869.

Gagneux, S., Long, C.D., Small, P.M., Van, T., Schoolnik, G.K., et al. (2006b). The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. *Science*. 312:1944.

Gagneux, S., and Small, P.M. (2007). Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *Lancet Infect Dis*. 7:328.

Gallant, V., Duvvuri, V., and McGuire, M. (2017). Tuberculosis (TB): Tuberculosis in Canada-Summary 2015. *Can Commun Dis Rep*. 43:77.

Garcia-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Gotz, S., et al. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*. 28:2678.

Gardy, J.L., Johnston, J.C., Ho Sui, S.J., Cook, V.J., Shah, L., et al. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med*. 364:730.

Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., et al. (2003). The complete genome sequence of Mycobacterium bovis. *Proc Natl Acad Sci U S A*. 100:7877.

Gasteiger, E., C, H., Gattiker, A., Duvaud, S., Wilkins, M., et al. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In *The proteomics protocols handbook*. J. Walker, editor.: Humana Press.

Gautam, S.S., Mac Aogain, M., Bower, J.E., Basu, I., and O'Toole, R.F. (2017). Differential carriage of virulence-associated loci in the New Zealand Rangipo outbreak strain of Mycobacterium tuberculosis. *Infect Dis (Lond)*. 49:680.

Gille, C., and Frömmel, C. (2001). STRAP: editor for STRuctural Alignments of Proteins. *Bioinformatics*. 17:377.

Gluckman, L.K. (1976). Medical history of New Zealand prior to 1860. Gluckman; Auckland, N.Z.

Goh, T.L., Towns, C.R., Jones, K.L., Freeman, J.T., and Wong, C.S. (2011). Extensively drug-resistant tuberculosis: New Zealand's first case and the challenges of management in a low-prevalence country. *Med J Aust*. 194:602.

Gordon, S.V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K., et al. (1999). Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol*. 32:643.

Greening, C., Ahmed, F.H., Mohamed, A.E., Lee, B.M., Pandey, G., et al. (2016). Physiology, Biochemistry, and Applications of F420- and Fo-Dependent Redox Reactions. *Microbiol Mol Biol Rev*. 80:451.

Guerra-Assunção, J., Crampin, A., Houben, R., Mzembe, T., Mallard, K., et al. (2015). Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *Elife*. 4:e05166.

Guerra-Assunção, J.A., Houben, R.M., Crampin, A.C., Mzembe, T., Mallard, K., et al. (2014). Recurrence due to relapse or reinfection with Mycobacterium tuberculosis: a whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. *J Infect Dis*. 211:1154.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., et al. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59:307.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696.

Gurumurthy, M., Rao, M., Mukherjee, T., Rao, S.P., Boshoff, H.I., et al. (2013). A novel F(420) -dependent anti-oxidant mechanism protects Mycobacterium tuberculosis against oxidative stress and bactericidal agents. *Mol Microbiol*. 87:744.

Hasan, M.R., Rahman, M., Jaques, S., Purwantini, E., and Daniels, L. (2010). Glucose 6-phosphate accumulation in mycobacteria: implications for a novel F420-dependent anti-oxidant defense system. *J Biol Chem*. 285:19135.

Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*. 16 Suppl 8:S1.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., et al. (2008). High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS Biol*. 6:e311.

Hershkovitz, I., Donoghue, H.D., Minnikin, D.E., Besra, G.S., Lee, O.Y., et al. (2008). Detection and molecular characterization of 9000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean. *PLOS ONE*. 3:e3426.

Hill, P., and Calder, L. (2000). An outbreak of tuberculosis in an Auckland church group. *N Z Public Health Rep*. 7:41.

Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W., and Small, P.M. (2004). Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proc Natl Acad Sci U S A*. 101:4871.

Ho, T.B., Robertson, B.D., Taylor, G.M., Shaw, R.J., and Young, D.B. (2000). Comparison of Mycobacterium tuberculosis genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. *Yeast*. 17:272.

Holt, K.E., McAdam, P., Thai, P.V.K., Thuong, N.T.T., Ha, D.T.M., et al. (2018). Frequent transmission of the Mycobacterium tuberculosis Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat Genet*:1.

Homolka, S., Projahn, M., Feuerriegel, S., Ubben, T., Diel, R., et al. (2012). High resolution discrimination of clinical Mycobacterium tuberculosis complex strains based on single nucleotide polymorphisms. *PLOS ONE*. 7:e39855.

Innis, H.A. (1999). The fur trade in Canada: An introduction to Canadian economic history. University of Toronto Press.

Isabelle, D., Simpson, D.R., and Daniels, L. (2002). Large-scale production of coenzyme F420-5, 6 by using Mycobacterium smegmatis. *Appl Environ Microbiol*. 68:5750.

Jabs, A., Weiss, M.S., and Hilgenfeld, R. (1999). Non-proline Cis peptide bonds in proteins 1. *J Mol Biol*. 286:291.

Jacobson, F., and Walsh, C. (1984). Properties of 7, 8-didemethyl-8-hydroxy-5-deazaflavins relevant to redox coenzyme function in methanogen metabolism. *Biochemistry*. 23:979.

Jeffrey, G.A., and Jeffrey, G.A. (1997). An introduction to hydrogen bonding. Oxford university press New York.

Jirapanjawat, T., Ney, B., Taylor, M.C., Warden, A.C., Afroze, S., et al. (2016). The redox cofactor F420 protects mycobacteria from diverse antimicrobial compounds and mediates a reductive detoxification system. *Appl Environ Microbiol*:AEM. 02500.

Jones, T.F., Woodley, C.L., Fountain, F.F., and Schaffner, W. (2003). Increased incidence of the outbreak strain of Mycobacterium tuberculosis in the surrounding community after an outbreak in a jail. *South Med J*. 96:155.

Kabsch, W. (2010). Xds. *Acta Crystallogr D*. 66:125.

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., et al. (1997). Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J Clin Microbiol*. 35:907.

Kamvar, Z.N., Tabima, J.F., and Grunwald, N.J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*. 2:e281.

Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*. 90:773.

Kato-Maeda, M., Kim, E.Y., Flores, L., Jarlsberg, L.G., Osmond, D., et al. (2010). Differences among sublineages of the East-Asian lineage of Mycobacterium tuberculosis in genotypic clustering. *Int J Tuberc Lung Dis*. 14:538.

Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., et al. (2001). Comparing genomes within the species Mycobacterium tuberculosis. *Genome Res*. 11:547.

Kato-Maeda, M., Shanley, C.A., Ackart, D., Jarlsberg, L.G., Shang, S., et al. (2012). Beijing sublineages of Mycobacterium tuberculosis differ in pathogenicity in the guinea pig. *Clin Vaccine Immunol*. 19:1227.

Kay, G.L., Sergeant, M.J., Zhou, Z., Chan, J.Z., Millard, A., et al. (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun*. 6:6717.

Khan, A., and Sarkar, D. (2012). Nitrate reduction pathways in mycobacteria and their implications during latency. *Microbiology*. 158:301.

King, M. (2004). The Penguin history of New Zealand. Penguin Books; Auckland, N.Z.

Koeck, J.L., Fabre, M., Simon, F., Daffe, M., Garnotel, E., et al. (2011). Clinical characteristics of the smooth tubercle bacilli 'Mycobacterium canettii'infection suggest the existence of an environmental reservoir. *Clin Microbiol Infect*. 17:1013.

Koser, C.U., Bryant, J.M., Becq, J., Torok, M.E., Ellington, M.J., et al. (2013). Whole-genome sequencing for rapid susceptibility testing of M. tuberculosis. *N Engl J Med*. 369:290.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol*. 372:774.

Kuchta, K., Knizewski, L., Wyrwicz, L.S., Rychlewski, L., and Ginalski, K. (2009). Comprehensive classification of nucleotidyltransferase fold proteins:

identification of novel families and their representatives in human. *Nucleic Acids Res*. 37:7701.

Lange, R. (1984). Plagues and Pestilence in Polynesia - the 19th-Century Cook Islands Experience. *Bulletin of the History of Medicine*. 58:325.

Lartillot, N., and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst Biol*. 55:195.

Laskowski, R.A., and Swindells, M.B. (2011). LigPlot+: multiple ligand–protein interaction diagrams for drug discovery. *J Chem Inf Model*. 51:2778.

Leinonen, R., Sugawara, H., Shumway, M., and International Nucleotide Sequence Database, C. (2011). The sequence read archive. *Nucleic Acids Res*. 39:D19.

Lew, J.M., Kapopoulou, A., Jones, L.M., and Cole, S.T. (2011). TubercuList--10 years after. *Tuberculosis (Edinb)*. 91:1.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. preprint arXiv:1303.3997

Lim, E., and Heffernan, H. (2013). Tuberculosis in New Zealand: Annual Report 2012. Institute of Environmental Science and Research Ltd (ESR), Porirua, N.Z.

Lo, M.-C., Aulabaugh, A., Jin, G., Cowling, R., Bard, J., et al. (2004). Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery. *Anal Biochem*. 332:153.

Lönnroth, K., Jaramillo, E., Williams, B.G., Dye, C., and Raviglione, M. (2009). Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Soc Sci Med*. 68:2240.

Lopez, B., Aguilar, D., Orozco, H., Burger, M., Espitia, C., et al. (2003). A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. *Clin Exp Immunol*. 133:30.

Lovell, S.C., Davis, I.W., Arendall III, W.B., De Bakker, P.I., Word, J.M., et al. (2003). Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, Bioinformatics*. 50:437.

MacLean, F.S. (1964). Challenge for health : a history of public health in New Zealand. Government Printer; Wellington, N.Z.

Maclellan, N. (1998). After Moruroa : France in the South Pacific. Ocean Press; Melbourne.

Manjunatha, U.H., Boshoff, H., Dowd, C.S., Zhang, L., Albert, T.J., et al. (2006). Identification of a nitroimidazo-oxazine-specific protein involved in PA-824 resistance in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A*. 103:431.

Marquina-Castillo, B., Garcia-Garcia, L., Ponce-de-Leon, A., Jimenez-Corona, M.E., Bobadilla-Del Valle, M., et al. (2009). Virulence, immunopathology and transmissibility of selected strains of Mycobacterium tuberculosis in a murine model. *Immunology*. 128:123.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17:10.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., et al. (2007). Phaser crystallographic software. *J Appl Crystallogr*. 40:658.

McElnay, C., Thornley, C., and Armstrong, R. (2004). A community and workplace outbreak of tuberculosis in Hawke's Bay in 2002. *N Z Med J*. 117:U1019.

Merker, M., Kohl, T.A., Niemann, S., and Supply, P. (2017). The Evolution of Strain Typing in the Mycobacterium tuberculosis Complex. In *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control*. Springer. 43.

Mulholland, C.V., Ruthe, A., Cursons, R.T., Durrant, R., Karalus, N., et al. (2017). Rapid molecular diagnosis of the Mycobacterium tuberculosis Rangipo strain responsible for the largest recurring TB cluster in New Zealand. *Diagn Microbiol Infect Dis*. 88:138.

Mulholland, C.V., Thorpe, D., Cursons, R.T., Karalus, N., Fong, Y., et al. (2018). Evaluation of the rapid molecular diagnostic test for the New Zealand Mycobacterium tuberculosis Rangipo strain in a clinical setting. *N Z Med J*. 131.

Murshudov, G.N., Skubák, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., et al. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D*. 67:355.

Ney, B., Ahmed, F.H., Carere, C.R., Biswas, A., Warden, A.C., et al. (2017). The methanogenic redox cofactor F420 is widely synthesized by aerobic soil bacteria. *ISME J*. 11:125.

Neyrolles, O., and Guilhot, C. (2011). Recent advances in deciphering the contribution of Mycobacterium tuberculosis lipids to pathogenesis. *Tuberculosis (Edinb)*. 91:187.

Nguyen, D., Brassard, P., Menzies, D., Thibert, L., Warren, R., et al. (2004). Genomic characterization of an endemic Mycobacterium tuberculosis strain: evolutionary and epidemiologic implications. *J Clin Microbiol*. 42:2573.

Nguyen, D., Brassard, P., Westley, J., Thibert, L., Proulx, M., et al. (2003). Widespread pyrazinamide-resistant Mycobacterium tuberculosis family in a low-incidence setting. *J Clin Microbiol*. 41:2878.

Nicholls, R.A., Fischer, M., McNicholas, S., and Murshudov, G.N. (2014). Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr D*. 70:2487.

NNDSS, N.N.D.S.S.S. (2015). 'Notification Rate of Tuberculosis 2015', Department of Health, http://www9.health.gov.au/cda/source/cda-index.cfm., (accessed October 2015)

O'Neill, M.B., Kitchen, A., Zarley, A., Aylwarde, W., Eldholm, V., et al. (2017). Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. *bioRxiv*. doi: https://doi.org/10.1101/210161.

Öhberg, A. (1955). Russia and the World Market in the Seventeenth Century: A Discussion of the Connection between Prices and Trade Routes. *The Scandinavian Economic History Review*. 3:154.

Orange, C. (2012). 'Treaty of Waitangi - Dishonouring the treaty – 1860 to 1880', Te Ara - the Encyclopedia of New Zealand, http://www.TeAra.govt.nz/en/graph/36364/maori-and-european-population-numbers-1840-1881, (accessed 12 September 2018)

Osman, D.A., Phelippeau, M., Drancourt, M., and Musso, D. (2017). Diversity of Mycobacterium tuberculosis lineages in French Polynesia. *J Microbiol Immunol Infect*. 50:199.

Oyugi, M.A., Bashiri, G., Baker, E.N., and Johnson-Winters, K. (2016). Investigating the Reaction Mechanism of F420-Dependent Glucose-6-phosphate Dehydrogenase from Mycobacterium tuberculosis: Kinetic Analysis of the Wild-Type and Mutant Enzymes. *Biochemistry*. 55:5566.

Oyugi, M.A., Bashiri, G., Baker, E.N., and Johnson-Winters, K. (2018). Mechanistic insights into F420-dependent glucose-6-phosphate dehydrogenase using isotope effects and substrate inhibition studies. *Biochim Biophys Acta*. 1866:387.

Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., et al. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom*. 2:e000056.

Pakenham, T. (2015). The scramble for Africa. Hachette, UK.

Panjikar, S., Parthasarathy, V., Lamzin, V.S., Weiss, M.S., and Tucker, P.A. (2005). Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. *Acta Crystallogr D*. 61:449.

Pankhurst, L.J., del Ojo Elias, C., Votintseva, A.A., Walker, T.M., Cole, K., et al. (2016). Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med*. 4:49.

Pepperell, C.S., Casto, A.M., Kitchen, A., Granka, J.M., Cornejo, O.E., et al. (2013). The role of selection in shaping diversity of natural M. tuberculosis populations. *PLoS Pathog*. 9:e1003543.

Pepperell, C.S., Granka, J.M., Alexander, D.C., Behr, M.A., Chui, L., et al. (2011). Dispersal of Mycobacterium tuberculosis via the Canadian fur trade. *Proc Natl Acad Sci U S A*. 108:5.

Perez-Lago, L., Comas, I., Navarro, Y., Gonzalez-Candelas, F., Herranz, M., et al. (2014). Whole genome sequencing analysis of intrapatient microevolution in Mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission. *J Infect Dis*. 209:98.

Perry, P., and Donini-Lenhoff, F. (2010). Stigmatization complicates infectious disease management. *Virtual Mentor*. 12:225.

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S.E., and Lercher, M.J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol*. 31:1929.

Phillips, J. (2015). 'History of immigration', Te Ara - the Encyclopedia of New Zealand, http://www.TeAra.govt.nz/en/history-of-immigration/, (accessed 12 September 2018)

Pool, I. (1991). Te Iwi Maori: A New Zealand Population, Past, Present and Projected. Auckland University Press; Auckland, NZ.

Portevin, D., Gagneux, S., Comas, I., and Young, D. (2011). Human macrophage responses to clinical isolates from the Mycobacterium tuberculosis complex discriminate between ancient and modern lineages. *PLoS Pathog*. 7:e1001307.

Purwantini, E., Daniels, L., and Mukhopadhyay, B. (2016). F420H2 is required for phthiocerol dimycocerosates synthesis in mycobacteria. *J Bacteriol*:JB.

Purwantini, E., and Mukhopadhyay, B. (2009). Conversion of NO2 to NO by reduced coenzyme F420 protects mycobacteria from nitrosative damage. *Proc Natl Acad Sci U S A*. 106:6333.

Purwantini, E., and Mukhopadhyay, B. (2013). Rv0132c of Mycobacterium tuberculosis encodes a coenzyme F420-dependent hydroxymycolic acid dehydrogenase. *PLOS ONE*. 8:e81985.

Ramakrishnan, L. (2012). Revisiting the role of the granuloma in tuberculosis. *Nat Rev Immunol*. 12:352.

Rambaut, A., Lam, T.T., Max Carvalho, L., and Pybus, O.G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*. 2:vew007.

Read, R.J., Adams, P.D., and McCoy, A.J. (2013). Intensity statistics in the presence of translational noncrystallographic symmetry. *Acta Crystallogr D*. 69:176.

Reed, M.B., Pichler, V.K., McIntosh, F., Mattia, A., Fallow, A., et al. (2009). Major Mycobacterium tuberculosis lineages associate with patient country of origin. *J Clin Microbiol*. 47:1119.

Reiling, N., Homolka, S., Walter, K., Brandenburg, J., Niwinski, L., et al. (2013). Clade-specific virulence patterns of Mycobacterium tuberculosis complex strains in human primary macrophages and aerogenically infected mice. *MBio*. 4:e00250.

Rieux, A., and Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. *Mol Ecol*. 25:1911.

Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res*. 42:W320.

Roetzer, A., Diel, R., Kohl, T.A., Ruckert, C., Nubel, U., et al. (2013). Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Med*. 10:e1001387.

Rolett, B.V. (2002). Voyaging and interaction in ancient east polynesia. *Asian Perspectives*. 42:182.

Rosas-Magallanes, V., Stadthagen-Gomez, G., Rauzier, J., Barreiro, L.B., Tailleux, L., et al. (2007). Signature-tagged transposon mutagenesis identifies novel Mycobacterium tuberculosis genes involved in the parasitism of human macrophages. *Infect Immun*. 75:504.

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

Rothschild, B.M., Martin, L.D., Lev, G., Bercovier, H., Bar-Gal, G.K., et al. (2001). Mycobacterium tuberculosis complex DNA from an extinct bison dated 17,000 years before the present. *Clin Infect Dis*. 33:305.

Ruthe, A. (2015). Time to Diagnosis and Persistence: The Two Major Determinants of Effective Tuberculosis Control. Doctor of Philosophy thesis, The University of Waikato, New Zealand.

Salentin, S., Schreiber, S., Haupt, V.J., Adasme, M.F., and Schroeder, M. (2015). PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res*. 43:W443.

Salgame, P., Geadas, C., Collins, L., Jones-Lopez, E., and Ellner, J.J. (2015). Latent tuberculosis infection - Revisiting and revising concepts. *Tuberculosis (Edinb)*. 95:373.

Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol*. 48:77.

Sassetti, C.M., and Rubin, E.J. (2003). Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A*. 100:12989.

Schnappinger, D., Ehrt, S., Voskuil, M.I., Liu, Y., Mangan, J.A., et al. (2003). Transcriptional Adaptation of Mycobacterium tuberculosis within Macrophages: Insights into the Phagosomal Environment. *J Exp Med*. 198:693.

Selengut, J.D., and Haft, D.H. (2010). Unexpected abundance of coenzyme F(420)-dependent enzymes in Mycobacterium tuberculosis and other actinobacteria. *J Bacteriol*. 192:5788.

Sexton, K., Perera, S., and Pandey, S. (2008). Five years of molecular typing of M. tuberculosis isolates in New Zealand, 2003 to 2007. Porirua, New Zealand

Shima, S., Warkentin, E., Grabarse, W., Sordel, M., Wicke, M., et al. (2000). Structure of coenzyme F420 dependent methylenetetrahydromethanopterin reductase from two methanogenic archaea. *J Mol Biol*. 300:935.

Smith, T., Wolff, K.A., and Nguyen, L. (2012). Molecular biology of drug resistance in Mycobacterium tuberculosis. In *Pathogenesis of Mycobacterium tuberculosis and its Interaction with the Host Organism*. Springer. 53.

Sola, C., Ferdinand, S., Mammina, C., Nastasi, A., and Rastogi, N. (2001). Genetic diversity of Mycobacterium tuberculosis in Sicily based on spoligotyping and variable number of tandem DNA repeats and comparison with a spoligotyping database for population-based analysis. *J Clin Microbiol*. 39:1559.

Spickard, P., Rondilla, J.L., and Hippolite Wright, D. (eds). (2002). Pacific Diaspora: Island Peoples in the United States and Across the Pacific. Honolulu, USA: University of Hawai'I Press.

Stanley, S.A., and Cox, J.S. (2013). Host–pathogen interactions during Mycobacterium tuberculosis infections. In *Pathogenesis of Mycobacterium tuberculosis and its Interaction with the Host Organism*. Springer. 211.

StatsNZ. (1916). 'Report on the results of a census of the population of the dominion of New Zealand taken for the night of the 15th october, 1916. (Digitised Collection)', https://www3.stats.govt.nz/historic_publications/1916-census/Report%20on%20Results%20of%20Census%201916/1916-report-results-census%20.html.

StatsNZ. (1996). 'New Zealand Official Yearbook collection: 1893–2012', http://archive.stats.govt.nz/browse_for_stats/snapshots-of-nz/digital-yearbook-collection.aspx.

StatsNZ. (2013). 'The New Zealand Census of Population and Dwellings, 2013', http://nzdotstat.stats.govt.nz/wbos/Index.aspx.

StatsNZ (2018). National population estimates: At 30 June 2018. Statistics New Zealand, Wellington, N.Z.

Steiner, A., Stucki, D., Coscolla, M., Borrell, S., and Gagneux, S. (2014). KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*. 15:881.

Stevens, K., and Wanhalla, A. (2017). Intimate Relations: Kinship and the Economics of Shore Whaling in Southern New Zealand, 1820-1860. *Journal of Pacific History*. 52:135.

Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A.M., et al. (2015). Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *J Infect Dis*. 211:1306.

Stucki, D., Ballif, M., Egger, M., Furrer, H., Altpeter, E., et al. (2016a). Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-Incidence Country. *J Clin Microbiol*. 54:1862.

Stucki, D., Brites, D., Jeljeli, L., Coscolla, M., Liu, Q., et al. (2016b). Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet*. 48:1535.

Stucki, D., Malla, B., Hostettler, S., Huna, T., Feldmann, J., et al. (2012). Two new rapid SNP-typing methods for classifying Mycobacterium tuberculosis complex into the main phylogenetic lineages. *PLOS ONE*. 7:e41253.

Suchard, M.A., and Redelings, B.D. (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*. 22:2047.

Supply, P., Marceau, M., Mangenot, S., Roche, D., Rouanet, C., et al. (2013). Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. *Nat Genet*. 45:172.

Supply, P., Warren, R.M., Banuls, A.L., Lesjean, S., van der Spuy, G.D., et al. (2003). Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area. *Mol Microbiol*. 47:529.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 43:D447.

Taylor, M.C., Jackson, C.J., Tattersall, D.B., French, N., Peat, T.S., et al. (2010). Identification and characterization of two families of F420H2-dependent reductases from Mycobacteria that catalyse aflatoxin degradation. *Mol Microbiol*. 78:561.

Terwilliger, T.C., Adams, P.D., Moriarty, N.W., and Cohn, J.D. (2007). Ligand identification using electron-density map correlations. *Acta Crystallogr D*. 63:101.

Terwilliger, T.C., Adams, P.D., Read, R.J., McCoy, A.J., Moriarty, N.W., et al. (2009). Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr D*. 65:582.

Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Zwart, P.H., et al. (2008). Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D*. 64:61.

Thauer, R.K., Jungermann, K., and Decker, K. (1977). Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol Rev*. 41:100.

Theus, S.A., Cave, M.D., and Eisenach, K.D. (2005). Intracellular macrophage growth rates and cytokine profiles of Mycobacterium tuberculosis strains with different transmission dynamics. *J Infect Dis*. 191:453.

Tiemersma, E.W., van der Werf, M.J., Borgdorff, M.W., Williams, B.G., and Nagelkerke, N.J. (2011). Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review. *PLOS ONE*. 6:e17601.

Trauner, A., Liu, Q., Via, L.E., Liu, X., Ruan, X., et al. (2017). The within-host population dynamics of Mycobacterium tuberculosis vary with treatment efficacy. *Genome Biol*. 18:71.

Tsenova, L., Harbacheuski, R., Sung, N., Ellison, E., Fallows, D., et al. (2007). BCG vaccination confers poor protection against M. tuberculosis HN878-induced central nervous system disease. *Vaccine*. 25:5126.

Tsolaki, A.G., Gagneux, S., Pym, A.S., Goguet de la Salmoniere, Y.O., Kreiswirth, B.N., et al. (2005). Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis. *J Clin Microbiol*. 43:3185.

van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., et al. (1993). Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol*. 31:406.

van Ruymbeke, R., Van Ruymbeke, B., and Sparks, R.J. (2003). Memory and Identity: The Huguenots in France and the Atlantic Diaspora. Univ of South Carolina Press.

Veyrier, F., Pletzer, D., Turenne, C., and Behr, M.A. (2009). Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. *BMC Evol Biol*. 9:196.

Veyrier, F.J., Dufort, A., and Behr, M.A. (2011). The rise and fall of the Mycobacterium tuberculosis genome. *Trends Microbiol*. 19:156.

Waddell, S.J., Chung, G.A., Gibson, K.J., Everett, M.J., Minnikin, D.E., et al. (2005). Inactivation of polyketide synthase and related genes results in the loss of complex lipids in Mycobacterium tuberculosis H37Rv. *Lett Appl Microbiol*. 40:201.

Walia, G., Kumar, P., and Surolia, A. (2009). The role of UPF0157 in the folding of M. tuberculosis dephosphocoenzyme A kinase and the regulation of the latter by CTP. *PLOS ONE*. 4:e7645.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE*. 9:e112963.

Walker, T.M., Ip, C.L., Harrell, R.H., Evans, J.T., Kapatai, G., et al. (2013a). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis*. 13:137.

Walker, T.M., Monk, P., Smith, E.G., and Peto, T.E. (2013b). Contact investigations for outbreaks of Mycobacterium tuberculosis: advances through whole genome sequencing. *Clin Microbiol Infect*. 19:796.

Wang, F., Jain, P., Gulten, G., Liu, Z., Feng, Y., et al. (2010). Mycobacterium tuberculosis dihydrofolate reductase is not a target relevant to the antitubercular activity of isoniazid. *Antimicrob Agents Chemother*. 54:3776.

Wang, Z., and Moult, J. (2001). SNPs, protein structure, and disease. *Human mutation*. 17:263.

Warren, R.M., Streicher, E.M., Sampson, S.L., van der Spuy, G.D., Richardson, M., et al. (2002a). Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data. *J Clin Microbiol*. 40:4457.

Warren, R.M., van der Spuy, G.D., Richardson, M., Beyers, N., Booysen, C., et al. (2002b). Evolution of the IS6110-based restriction fragment length polymorphism pattern during the transmission of Mycobacterium tuberculosis. *J Clin Microbiol*. 40:1277.

Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 42:D581.

Weniger, T., Krawczyk, J., Supply, P., Niemann, S., and Harmsen, D. (2010). MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria. *Nucleic Acids Res*. 38:W326.

Widdel, F., and Wolfe, R. (1989). Expression of secondary alcohol dehydrogenase in methanogenic bacteria and purification of the F420-specific enzyme from Methanogenium thermophilum strain TCI. *Arch Microbiol*. 152:322.

Wilbur, A.K., Bouwman, A.S., Stone, A.C., Roberts, C.A., Pfister, L.-A., et al. (2009). Deficiencies and challenges in the study of ancient tuberculosis DNA. J Archaeol Sci:1990.

Wilmshurst, J.M., Hunt, T.L., Lipo, C.P., and Anderson, A.J. (2011). High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proc Natl Acad Sci U S A*. 108:1815.

Wilson, J. (2016). 'European discovery of New Zealand - French explorers', Te Ara - the Encyclopedia of New Zealand, http://www.TeAra.govt.nz/en/european-discovery-of-new-zealand/page-8 (accessed 10 September 2018)

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr D*. 67:235.

Wirth, T., Hildebrand, F., Allix-Beguec, C., Wolbeling, F., Kubica, T., et al. (2008). Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathog*. 4:e1000160.

Yang, C., Luo, T., Sun, G., Qiao, K., Sun, G., et al. (2012). Mycobacterium tuberculosis Beijing strains favor transmission but not drug resistance in China. *Clin Infect Dis*. 55:1179.

Yen, S., Bower, J.E., Freeman, J.T., Basu, I., and O'Toole, R.F. (2013). Phylogenetic lineages of tuberculosis isolates in New Zealand and their association with patient demographics. *Int J Tuberc Lung Dis*. 17:892.

Young, D.B., Gideon, H.P., and Wilkinson, R.J. (2009). Eliminating latent tuberculosis. *Trends Microbiol*. 17:183.

Yu, G., Smith, D., Zhu, H., Guan, Y., and Lam, T.T. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. 8:28.

Zhang, H., Li, D., Zhao, L., Fleming, J., Lin, N., et al. (2013). Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat Genet*. 45:1255.

Zheng, H., Chruszcz, M., Lasota, P., Lebioda, L., and Minor, W. (2008). Data mining of metal ion environments present in protein structures. *J Inorg Biochem*. 102:1765.

Zhu, L., Zhong, J., Jia, X., Liu, G., Kang, Y., et al. (2015). Precision methylome characterization of Mycobacterium tuberculosis complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res*. 44:730.

Appendices

Appendix A: Appendices relating to Chapter Two

Appendix A.1. *MIRU-VNTR typing patterns of New Zealand M. tuberculosis cluster isolates provided by LabPLUS.*

Cluster	MIRU ID	MIRU-12	MIRU-24
Rangipo	NZ_040	233325153324	341344223362
	NZ_040_001	233325153324	341444223352
	NZ_040_002	233325153324	341434223362
	NZ_040_003	233325153324	341544223362
Southern Cross	NZ_061_001	123326153326	343224123253
	NZ_061_002	123326153326	342224123253
	NZ_061_005	123326153326	342125123253
Otara	NZ_062_001	233325143322	341544223382
	NZ_062_002	233325143322	341534223392
	NZ_062_004	233325143322	2415442233C2
NZ_037	NZ_037	124326153224	323124123262
NZ_041	NZ_041_001	223125153324	242334223552
	NZ_041_002	223125153324	222334223552
NZ_069	NZ_069	223326153321	242234223552
	NZ_069_006	223326153321	242334223552
NZ_094	NZ_094_001	233325143325	3415442213B2

Appendix A.2. *Rangipo SNPs identified from SOLiD whole genome sequencing data.* Spreadsheets detailing these are provided in a separate file (Appendix_A2_Rangipo-SOLiD-SNPs.xlxs)

Common Rangipo SNPs: Classification of SNPs common to the Rangipo strain identified by Colangeli et al. (2014) (n = 247)

Rangipo Specific SNPs: SNPs classified as specific to the Rangipo strain (n = 22) and Rangipo and SUMu/Canadian DS6^{Quebec} shared SNPs (n = 4)

Appendix A.3. *Rangipo SNP validation primers. Primer sequences and expected PCR product sizes for validation of Rangipo specific SNPs by Sanger sequencing.*

SNP	Locus	Primer name	Primer sequence (5' to 3')	Product size (bp)
Rv3253CG	Rv0002 /*dnaN*	Rv0002(P401R)_fwd Rv0002(P401R)_rev	GATTGCGTTTAACCCAACCTATCT GCAATTCCAGATCTACACATGCC	297
Rv550620GT	Rv0458	Rv0458(D316Y)_fwd Rv0458(D316Y)_rev	ACGACGACTTCTGCGACAA GGACAACACCTTTTCCAGC	227
Rv1289731TC	Rv1161 /*narG*	Rv1161(Y802H)_fwd Rv1161(Y802H)_rev	ACATTCCAGAGGGCAAGCTC CCAGATGACGTTTCGCCAG	249
Rv1836099TG	Rv1631 /*coaE*	Rv1631(Y363D)_fwd Rv1631(Y363D)_rev	CTACGACCACACCGACAGT ATCCAGGAACCACGGCTC	253
Rv2807374GC	Rv2492	Rv2492(G33R)_fwd Rv2492(G33R)_rev	AATCTTTGGGGTCGCCTTTC GGATAATTCTGCACCGAAGACT	295
Rv3202633GA	Rv2893	Rv2893(G72S)_fwd Rv2893(G72S)_rev	ACTCCATGACCGTTGCCAG CCCGTGTTGCCCCGGAATC	387
Rv3283879TC	Rv2941 /*fadD28*	Rv2941(V182A)_fwd Rv2941(V182A)_rev	ATCGAAGTTGATTTGCTCGA CATGTCGTGGTAGAAGGGT	246
Rv3366098GA	Rv3007c	RV3007c(P118L)_fwd RV3007c(P118L)_rev	AAGCCACATAATCCCCGACA CTTCGCTACCCAAACGAGC	290
Rv3561770AG	Rv3193c	Rv3193c(L468S)_fwd Rv3193c(L468S)_rev	GCGGACGTGTTAGAGATCACC AACGGTAACCTGCGCGACTA	293
Rv3895925GA	Rv3479	Rv3479(A36T)_fwd Rv3479(A36T)_rev	CGGTGTTACGCGGGAGAT CGGAGAAGCTCGGTCAAGG	288
Rv3980075CA	Rv3540c /*ltp2*	Rv3540c(E195D)_fwd Rv3540c(E195D)_rev	GTCTCCTGGCAGCAGTCC GGTCAGGTGCAAACTCGTTT	296
Rv4085870GA	Rv3646c /*topA*	Rv3646c(T463M)_fwd Rv3646c(T463M)_rev	TCGACGATGTCCAACCGTT CCAACATTGATGATTTCCGGC	274
Rv4377908CT	Rv3894c /*eccC2*	Rv3894c(G849S)_fwd Rv3894c(G849S_rev	CACGATGATGTTGCTGCGTAG CCCGACGAATTCCTCTATTACGA	291

Appendix B: Appendices relating to Chapter Three

Appendix B.1. *Bacterial strains, vectors and media used in this study.*

Bacterial Strain	Description
E. coli TOP10 (electrocompetent)	F-mcrA Δ(mrr-hsdRMS-mcrBC) Φ80 dlacZ ΔM15 Δ lacX74 deoR recA1 araD139 Δ(ara leu)7697 galU galK rpsL endA1 nupG
M. smegmatis $mc^2$4517	*M. smegmatis* expression strain with T7 RNA polymerase; Km^r (Wang *et al*. 2010)

Vector	Description
pYUB28b	*E. coli* mycobacterium shuttle vector. 4921 bp, T7 promoter, MCS pET28b, C- and N-terminal His-tags, N-terminal thrombin cleavage site, Hyg^r (Bashiri *et al*. 2010)

Solid media	Description
LB agar	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, 15 g/L agar
Low salt LB agar	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl, 15 g/L agar
7H10/ADC/T80 agar	1.9 g 7H10 powder, 0.5% (v/v) glycerol in 90 ml H_2O, autoclaved then 10 ml ADC enrichment and 0.05% (v/v) Tween 80 added at 50 °C

Liquid media	Description
LB broth	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl,
Low salt LB broth	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl,
7H9/ADC/T80 broth	0.47 g 7H9 powder, 0.2% (v/v) glycerol in 90 ml H_2O, autoclaved then 10 ml ADC enrichment medium and 0.05% (v/v) Tween-80 added at 50 °C
SOC recovery media	2% (w/v) bactotryptone, 0.55%(w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, autoclaved and then added 20 mM glucose before use.
PA-0.5G defined media	50 mM Na_2HPO_4, 50 mM KH_2PO_4, 25 mM $(NH_4)_2SO_4$, 1 mM $MgSO_4$, 0.5% (w/v) glucose, 0.1X metals mix[1], 100 $\mu g.ml^{-1}$ each of 17 amino acids (no Cys, Tyr or Met). Individual components autoclaved or sterile filtered before adding to sterile H_2O
ZYP-5052 auto-induction media	1% (w/v) bactotryptone, 0.5% (w/v) yeast extract autoclaved and then added sterile 50 mM Na_2HPO_4, 50 mM KH_2PO_4, 25 mM $(NH_4)_2SO_4$, 1 mM $MgSO_4$, 0.5% (w/v) glycerol, 0.05% (w/v) glucose, 0.2% (w/v) α-lactose, 1X metals mix[1]

[1] 1000X Metals mix made up from the sterile stock solutions of separate components to give the following concentrations: 50 μM $FeCl_3$ in 0.12 M HCl (filter sterile), 20 μM $CaCl_2$, 10 μM $MnCl_2$, 10 μM $ZnSO_4$, 2 μM $CoCl_2$, 2 μM $CuCl_2$, 2 μM $NiCl_2$, 2 μM Na_2MoO_4, 2 μM Na_2SeO_3, 2 μM H_3BO_3.

Appendix B.2. *Cloning primers.* Primers used to clone Rv2893 into pYUB28b and T7 primers used for screening and sequence confirmation of the insert.

Primer name	Primer sequence (5' to 3')	RE site	Product size (bp)
Rv2893 cloning_fwd	GGAATTCCATATGACCGTTGCCAGCACCGCTCA	NdeI	996
Rv2893 cloning_rev	CGGGATCCCTAGCCGTAGCGCAGGAG	BamHI	
T7_fwd (promoter)	TAATACGACTCACTATAGGG		1260/317[1]
T7_rev (terminator)	GCTAGTTATTGCTCAGCGG		

[1]1260 bp if the correct insert is ligated, 317 bp for empty vector

Appendix B.3. *Gene and protein information*. The position of the Rangipo G72S mutation is in bold and highlighted by a grey box.

Rv2893^{H327Rv} (978 bp)

ATGACCGTTGCCAGCACCGCTCACCATACACGTCGGCTACGTTTCGGGTTGGCGGCACCGTTGCCC
CGCGCGGGCACCCAGATGCGCGCCTTCGCGCAGGCTGTCGAGGCCGCCGGGTTCGACGTGCTGGCC
TTCCCGGACCACCTGGTGCCTTCGGTTTCGCCGTTCGCAGGCGCGACCGCCGCGGCGATGGCCACG
CAACGACTGCACACC**G**GCACATTGGTGCTCAACAACGACTTTCGCCATCCCGTGGACACCGCTCGA
GAGGCGGCCGGTGTGGCAACCCTCGCCGAAGGCCGCTTCGAACTGGGACTGGGCGCCGGACACCGG
AGGTCCGAATACGACGCCGCCGGCATTACCTTCGATTCCGGGGCAACACGGGTGGCGCGGCTCATC
GAATCGGCGCACCTGATCCGTGCGCTGCTGGACGCGGAGCCCGTCGACTTCGACGGGCAGCATTAC
CGGGTGCACGCCGAAGCGGGCTCACTGGTGGCACCGCCGAAGGTCCGGGTCCCCCTGCTAGTGGGC
GGCAACGGGACCGAGGTGCTGCGGCTGGGCGGACGCATCGCCGACATTGTCGGCCTGGCCGGGATC
AGCCACAACCGCGACGCCACCCAGGTCCGGTTCACCCACTTCGACGCCGACGGCCTGGCCGACCGG
ATCGCCGTGGTACGTCACGCGGCCGGCGATCGCTTCGAAGCCATTGAGCTCAACGCGCTGATCCAG
GCGGTGGTCTGCACCAACGACCGAAACGCGGCGGCCGCCGAACTGGCCGCCACCTTGGGCGGGATC
ACGCCCGAGCAGGTCCTCGAGTCGCCGTTTCTGCTGCTCGGTACCCACGAGCAGATGGCCGAGGCT
CTCGCCGCGCGGCAGCGGCGGTTCGGTGTCAGCTATTGGACGGTGTTCGACGAGTGGGCTGGCCGC
GCGTCGGCAATGCGCGACATCGCCGAGGTCATCGCGCTCCTGCGCTACGGCTAG

Rv2893^{H327Rv} (325 aa)

MW 34617.24 Da, pI 6.22

MTVASTAHHTRRLRFGLAAPLPRAGTQMRAFAQAVEAAGFDVLAFPDHLVPSVSPFAGATAAAMAT
QRLHT**G**TLVLNNDFRHPVDTAREAAGVATLAEGRFELGLGAGHRRSEYDAAGITFDSGATRVARLI
ESAHLIRALLDAEPVDFDGQHYRVHAEAGSLVAPPKVRVPLLVGGNGTEVLRLGGRIADIVGLAGI
SHNRDATQVRFTHFDADGLADRIAVVRHAAGDRFEAIELNALIQAVVCTNDRNAAAAELAATLGGI
TPEQVLESPFLLLGTHEQMAEALAARQRRFGVSYWTVFDEWAGRASAMRDIAEVIALLRYG

Rv2893^{H327Rv} pYUB28b N-terminal His-tag (345 aa)

MW 36780.57 Da, pI 6.57

MGSSHHHHHHSSGLVPRGSHMTVASTAHHTRRLRFGLAAPLPRAGTQMRAFAQAVEAAGFDVLAFP
DHLVPSVSPFAGATAAAMATQRLHT**G**TLVLNNDFRHPVDTAREAAGVATLAEGRFELGLGAGHRRS
EYDAAGITFDSGATRVARLIESAHLIRALLDAEPVDFDGQHYRVHAEAGSLVAPPKVRVPLLVGGN
GTEVLRLGGRIADIVGLAGISHNRDATQVRFTHFDADGLADRIAVVRHAAGDRFEAIELNALIQAV
VCTNDRNAAAAELAATLGGITPEQVLESPFLLLGTHEQMAEALAARQRRFGVSYWTVFDEWAGRAS
AMRDIAEVIALLRYG

Rv2893^{G72S} (978 bp)

ATGACCGTTGCCAGCACCGCTCACCATACACGTCGGCTACGTTTCGGGTTGGCGGCACCGTTGCCC
CGCGCGGGCACCCAGATGCGCGCCTTCGCGCAGGCTGTCGAGGCCGCCGGGTTCGACGTGCTGGCC
TTCCCGGACCACCTGGTGCCTTCGGTTTCGCCGTTCGCAGGCGCGACCGCCGCGGCGATGGCCACG
CAACGACTGCACACC**A**GCACATTGGTGCTCAACAACGACTTTCGCCATCCCGTGGACACCGCTCGA
GAGGCGGCCGGTGTGGCAACCCTCGCCGAAGGCCGCTTCGAACTGGGACTGGGCGCCGGACACCGG
AGGTCCGAATACGACGCCGCCGGCATTACCTTCGATTCCGGGGCAACACGGGTGGCGCGGCTCATC
GAATCGGCGCACCTGATCCGTGCGCTGCTGGACGCGGAGCCCGTCGACTTCGACGGGCAGCATTAC
CGGGTGCACGCCGAAGCGGGCTCACTGGTGGCACCGCCGAAGGTCCGGGTCCCCCTGCTAGTGGGC
GGCAACGGGACCGAGGTGCTGCGGCTGGGCGGACGCATCGCCGACATTGTCGGCCTGGCCGGGATC
AGCCACAACCGCGACGCCACCCAGGTCCGGTTCACCCACTTCGACGCCGACGGCCTGGCCGACCGG
ATCGCCGTGGTACGTCACGCGGCCGGCGATCGCTTCGAAGCCATTGAGCTCAACGCGCTGATCCAG
GCGGTGGTCTGCACCAACGACCGAAACGCGGCGGCCGCCGAACTGGCCGCCACCTTGGGCGGGATC
ACGCCCGAGCAGGTCCTCGAGTCGCCGTTTCTGCTGCTCGGTACCCACGAGCAGATGGCCGAGGCT
CTCGCCGCGCGGCAGCGGCGGTTCGGTGTCAGCTATTGGACGGTGTTCGACGAGTGGGCTGGCCGC
GCGTCGGCAATGCGCGACATCGCCGAGGTCATCGCGCTCCTGCGCTACGGCTAG

Rv2893^{G72S} (325 aa)

MW 34647.27 Da, pI 6.22

MTVASTAHHTRRLRFGLAAPLPRAGTQMRAFAQAVEAAGFDVLAFPDHLVPSVSPFAGATAAAMAT
QRLHT**S**TLVLNNDFRHPVDTAREAAGVATLAEGRFELGLGAGHRRSEYDAAGITFDSGATRVARLI
ESAHLIRALLDAEPVDFDGQHYRVHAEAGSLVAPPKVRVPLLVGGNGTEVLRLGGRIADIVGLAGI
SHNRDATQVRFTHFDADGLADRIAVVRHAAGDRFEAIELNALIQAVVCTNDRNAAAAELAATLGGI
TPEQVLESPFLLLGTHEQMAEALAARQRRFGVSYWTVFDEWAGRASAMRDIAEVIALLRYG

Rv2893^{G72S} pYUB28b N-terminal His-tag (345 aa)

MW 36810.59 Da, pI 6.57

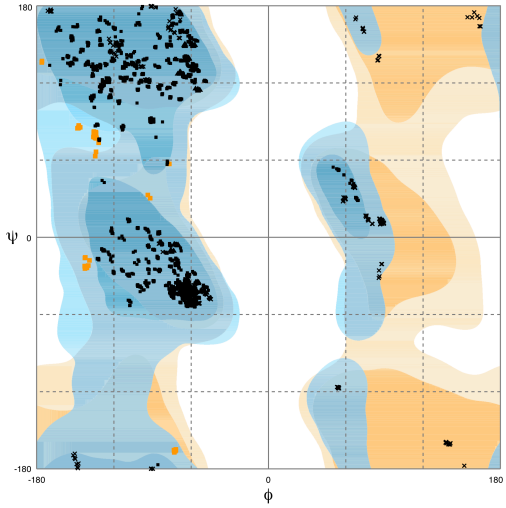
MGSSHHHHHHSSGLVPRGSHMTVASTAHHTRRLRFGLAAPLPRAGTQMRAFAQAVEAAGFDVLAFP
DHLVPSVSPFAGATAAAMATQRLHT**S**TLVLNNDFRHPVDTAREAAGVATLAEGRFELGLGAGHRRS
EYDAAGITFDSGATRVARLIESAHLIRALLDAEPVDFDGQHYRVHAEAGSLVAPPKVRVPLLVGGN
GTEVLRLGGRIADIVGLAGISHNRDATQVRFTHFDADGLADRIAVVRHAAGDRFEAIELNALIQAV
VCTNDRNAAAAELAATLGGITPEQVLESPFLLLGTHEQMAEALAARQRRFGVSYWTVFDEWAGRAS
AMRDIAEVIALLRYG

Appendix B.4. *Rv2893 protein structures.* MTZ and PDB files are provided separately in folder Appendix_B3_structures. Note that in PDB files numbering starts at the methionine before the His-tag and therefore add 20 amino acids on to the position, e.g. G72 is G92.

Structure	Files (.mtz and .pdb)
Rv2893^{H37Rv} SAD solution	Rv2893_solution
apo-Rv2893^{H37Rv}	apo-Rv2893H37Rv
apo-Rv2893^{G72S}	apo-Rv2893G72S
F_{420}-Rv2893^{H37Rv}	F420-Rv2893H37Rv
F_{420}-Rv2893^{G72S}	F420-Rv2893G72S

Appendix B.5. *Ramachandran analysis for Rv2893 structures.*

apo-Rv2893^{H37Rv}

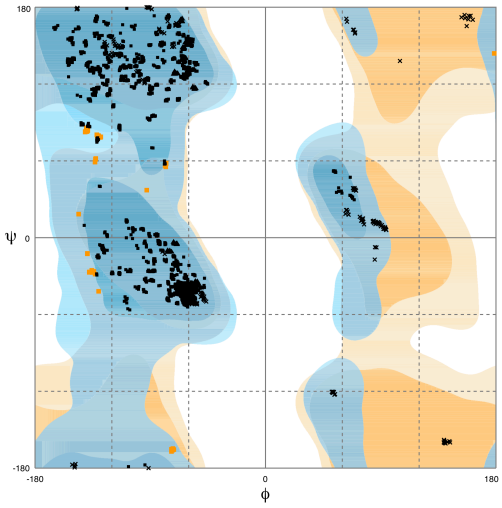


Number of residues in favoured region
(~98.0% expected) : 2376 (97.9%)

Number of residues in allowed region
(~2.0% expected) : 51 (2.1%)

Number of residues in outlier region
0 (0.0%)

apo-Rv2893^{G72S}

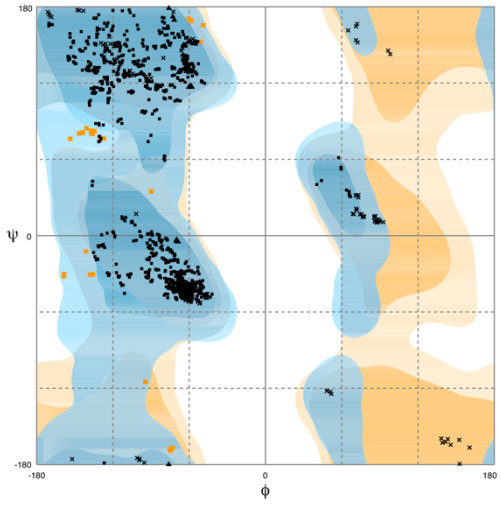


Number of residues in favoured region
(~98.0% expected) : 2371 (98.1%)

Number of residues in allowed region
(~2.0% expected) : 45 (1.9%)

Number of residues in outlier region
0 (0.0%)

F_{420}-Rv2893^{H37Rv}

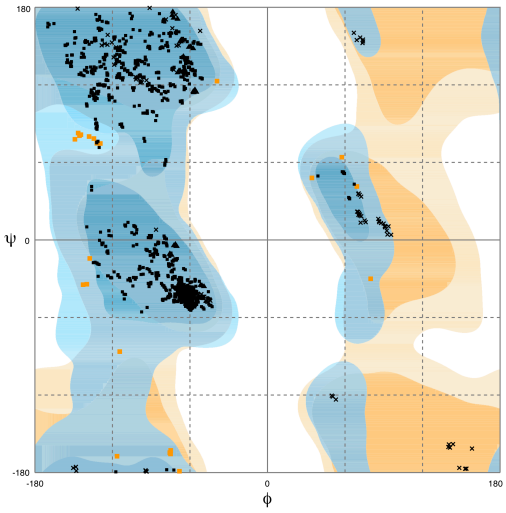


Number of residues in favoured region
(~98.0% expected) : 1218 (97.8%)

Number of residues in allowed region
(~2.0% expected) : 27 (2.2%)

Number of residues in outlier region
0 (0.0%)

F_{420}-Rv2893^{G72S}



Number of residues in favoured region
(~98.0% expected) : 1221 (98.2%)

Number of residues in allowed region
(~2.0% expected) : 23 (1.8%)

Number of residues in outlier region
0 (0.0%)

Appendix B.6. *Root mean square differences (RMSD) in the Cα atomic positions between protein chains for Rv2893 structures.* Chains with F_{420} bound are shaded grey.

		apo-Rv2893^{H37Rv}								apo-Rv2893^{G72S}								F_{420}-Rv2893H37Rv				F_{420}-Rv2893^{G72S}			
		A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H	A	B	C	D	A	B	C	D
apo-Rv2893^{H37Rv}	A		0.30	0.16	0.45	0.48	1.20	0.56	0.73	0.17	0.29	0.23	0.43	0.50	0.70	0.46	1.15	0.30	1.37	0.99	1.53	0.31	1.35	0.96	1.53
	B			0.34	0.11	0.32	0.40	0.35	0.33	0.30	0.15	0.34	0.12	0.34	0.34	0.33	0.42	0.40	0.37	0.49	0.51	0.40	0.36	0.43	0.51
	C				0.50	0.46	1.20	0.54	0.73	0.24	0.34	0.18	0.48	0.48	0.70	0.44	1.15	0.31	1.40	0.97	1.52	0.31	1.37	0.95	1.52
	D					0.42	0.43	0.36	0.34	0.47	0.18	0.50	0.12	0.36	0.35	0.42	0.45	0.54	0.39	0.51	0.56	0.55	0.39	0.46	0.56
	E						0.61	0.24	0.51	0.51	0.36	0.48	0.39	0.25	0.50	0.17	0.62	0.49	0.53	0.65	0.42	0.48	0.51	0.64	0.39
	F							0.45	0.26	1.22	0.43	1.20	0.45	0.49	0.28	0.62	0.15	1.21	1.05	0.26	0.84	1.20	1.03	0.27	0.84
	G								0.36	0.60	0.38	0.56	0.35	0.15	0.36	0.32	0.48	0.57	0.50	0.51	0.43	0.56	0.48	0.51	0.39
	H									0.75	0.37	0.73	0.38	0.40	0.11	0.53	0.30	0.78	0.66	0.35	0.61	0.77	0.63	0.30	0.60
apo-Rv2893$_{G72S}$	A										0.27	0.19	0.43	0.52	0.71	0.46	1.16	0.33	1.37	1.01	1.53	0.34	1.35	0.97	1.53
	B											0.32	0.18	0.35	0.36	0.35	0.44	0.38	0.32	0.53	0.52	0.38	0.32	0.46	0.52
	C												0.47	0.48	0.69	0.43	1.14	0.33	1.38	0.97	1.51	0.34	1.35	0.95	1.50
	D													0.34	0.38	0.38	0.47	0.51	0.38	0.54	0.55	0.52	0.39	0.48	0.55
	E														0.39	0.27	0.49	0.54	0.48	0.55	0.41	0.52	0.45	0.54	0.36
	F															0.51	0.29	0.78	0.73	0.35	0.60	0.76	0.63	0.32	0.58
	G																0.62	0.45	0.50	0.67	0.41	0.43	0.49	0.66	0.40
	H																	1.17	1.05	0.22	0.78	1.15	0.98	0.29	0.78
F_{420}-Rv2893^{H37R}	A																		1.31	1.03	1.47	0.18	1.30	1.01	1.47
	B																			0.96	0.76	1.14	0.17	0.98	0.73
	C																				0.74	1.04	0.94	0.23	0.75
	D																					1.27	0.73	0.77	0.24
F_{420}-Rv2893^{G72S}	A																						1.12	1.01	1.25
	B																							0.89	0.69
	C																								0.75
	D																								

Appendix B.7. *Unmodeled regions in Rv2893 structures.*

Structure	Chain	From		To	
apo-Rv2893^{H37Rv}	B	GLY	197	PHE	212
	B	ARG	250	SER	272
	B	TRP	305	ARG	308
	D	ILE	198	PHE	212
	D	TRP	305	ARG	308
	E	ILE	198	PHE	209
	E	TRP	305	GLY	307
	F	HIS	200	VAL	207
	G	ILE	198	THR	210
	H	ILE	198	PHE	209
apo-Rv2893^{G72S}	B	ALA	196	PHE	212
	B	ASN	251	SER	272
	B	GLU	304	ARG	308
	D	GLY	197	HIS	211
	D	TRP	305	ARG	308
	E	ILE	198	THR	210
	E	TRP	305	GLY	307
	F	ILE	198	PHE	209
	F	TRP	305	GLY	307
	G	ILE	198	THR	210
	G	ALA	258	ILE	264
	G	TRP	305	GLY	307
	H	SER	199	VAL	207
F_{420}-Rv2893^{H37Rv}	A	TRP	305	GLY	307
	B	TRP	305	GLY	307
	C	SER	199	ARG	208
	C	TRP	305	GLY	307
	D	THR	260	GLY	263
	D	TRP	305	ARG	308
F_{420}-Rv2893^{G72S}	A	ALA	204	GLN	206
	A	TRP	305	ARG	308
	B	TRP	305	ARG	308
	C	SER	199	ARG	208
	C	TRP	305	GLY	307
	D	TRP	305	ARG	308

Appendix C: Appendices relating to Chapter Four

Appendix C.1. *New Zealand* M. tuberculosis *isolates sequenced on the Illumina MiSeq platform (n = 25).*

Rangipo cluster isolates (n = 13)

ID[1]	Year	MIRU ID	MIRU-12	MIRU-24	IS*6110*-RFLP
NZLA	1992				
NZLB	1992				
NZLC	1999				
NZLE	2001				
NZLF	2002				13/008RP
NZLG	2004				13/008RP
NZLH	2006				13/008
NZLI	2006				13/008RP
NZLJ	2008				13/008RP
NZLK (O)	2008	NZL_040_001	233325153324	341444223362	13/008RP
NZLL (F)	1996	NZL_040_001	233325153324	341444223362	
NZLM (R)	2009				13/008RP
NZLN (A)	1991	NZL_040_001	233325153324	341444223362	

[1] The ID isolates were sequenced under on the ABI SOLiD platform (Colangeli *et al.* 2014) are shown in brackets.

Otara cluster isolates (n = 7)

ID	Year	MIRU ID	MIRU-12	MIRU-24
NZL01	2008	NZL_062-001	233325143322	341544223392
NZL03	2011	NZL_062-004	233325143322	2415442233C2
NZL05	2008	NZL_062-002	233325143322	341534223392
NZL10	2013	NZL_062-001	233325143322	341544223392
NZL11	2003	NZL_127-001	233325141322	341544223392
NZL12	2013	NZL_127-001	233325141322	341544223392
NZL13	2006	NZL_062-001	233325143322	341544223392

Southern Cross cluster isolates (n = 5)

ID	Year	MIRU ID	MIRU-12	MIRU-24
NZL02	2010	NZL_061-001	123326153326	343224123253
NZL04	2007	NZL_061-001	123326153326	343224123253
NZL07	2015	NZL_061-001	123326153326	343224123253
NZL08	2002	NZL_061-001	123326153326	343224123253
NZL09	2002	NZL_061-001	123326153326	343224123253

Appendix C.2. *Sequencing and mapping statistics for 25 New Zealand* M. tuberculosis *genomes sequenced on the Illumina MiSeq platform.*

Isolate ID	Total read count	Mean read length	Median read length	Raw read depth	Mean Coverage[1]	Mean MQ	% Reads Mapped	%10X[2]	%20X[2]	%30X[2]
NZLA	1781828	214.6	250	86.7	66.9	58.1	88.9	99.2	98.3	93.4
NZLB	1181092	213.1	250	57.1	43.8	57.9	87.5	98.7	90.1	66.2
NZLC	1211718	229.6	250	63.1	49.4	58.2	90.2	98.9	94.2	77.0
NZLE	1270322	224.6	250	64.7	49.8	58.0	88.3	99.0	94.5	77.2
NZLF	902728	224.7	250	46.0	35.8	57.9	88.1	98.2	84.1	51.3
NZLG	779518	193.3	222	34.2	30.4	57.7	96.8	95.0	68.3	35.8
NZLH	1497248	225.9	250	76.7	57.6	58.1	88.3	99.1	97.8	89.9
NZLI	1341768	198.1	249	60.3	46.2	57.8	89.6	98.2	89.8	68.6
NZLJ	1465386	226.7	250	75.3	50.3	58.0	82.5	98.9	94.2	76.6
NZLK	2472612	190.2	214	106.6	99.3	58.2	98.7	99.0	98.3	97.4
NZLL	1280644	221.5	250	64.3	47.1	57.9	85.3	98.8	91.0	69.8
NZLM	1161542	219.5	250	57.8	42.8	58.1	87.2	98.7	90.5	65.4
NZLN	913298	206.6	241	42.8	34.0	58.1	88.5	97.8	81.6	48.7
NZL01	4018730	203.5	248	185.3	147.9	58.9	83.8	99.3	99.0	98.7
NZL03	4915778	212.7	250	237.0	206.9	58.8	94.0	99.3	99.1	98.8
NZL05	4978762	195.7	233	220.8	173.5	58.9	80.9	99.2	99.0	98.7
NZL10	3783886	209.3	249	179.5	152.6	58.7	91.6	99.3	99.0	98.7
NZL11	4382194	212.3	250	210.9	185.8	58.8	93.2	99.4	99.2	99.0
NZL12	1703674	199.5	232	77.0	71.8	58.7	97.4	98.7	97.7	96.1
NZL13	1929004	206.6	250	90.3	85.4	58.9	98.9	98.7	97.9	96.9
NZL02	5147200	218.5	250	254.9	226.0	59.1	95.9	99.2	98.9	98.6
NZL04	5178286	190.4	214	223.5	182.0	58.9	84.8	99.2	99.0	98.7
NZL07	5542744	205.1	249	257.7	204.2	59.1	83.7	99.1	98.9	98.6
NZL08	4652740	195.2	231	205.8	168.4	59.0	84.6	99.3	99.1	98.8
NZL09	5069446	198.0	241	227.6	170.8	59.0	78.3	99.2	98.9	98.6

[1] Average mapped coverage across the H37Rv reference genome (NC_000962.3).
[2] Fraction of the reference sequenced with at least a given coverage rate

Appendix C.3. *Rangipo SNPs identified by analysis of Illumina whole genome sequencing data.* Spreadsheets detailing these are provided in a separate file (Appendix_C3_Rangipo-Illumina-SNPs.xlxs).

Common Rangipo SNPs: SNPs shared by all Rangipo isolates (n = 513)
Rangipo Specific SNPs: Rangipo specific SNPs identified from Illumina WGS data by comparison with 220 global L4.4 M. *tuberculosis* genomes (n = 53)

Appendix C.4. *Isolates used in phylogenetic analyses.* Spreadsheets detailing these are provided in a separate file (Appendix_C4_L44-dataset.xlxs).

Dataset 1: Complete L4.4 sublineage global dataset (n = 236)
Dataset 2: L4.4.1.1/S sublineage molecular dating dataset (n = 117)

Appendix C.5. *Frequency of* M. tuberculosis *L4.4 sublineage genomes included in phylogenetic analyses by country.*

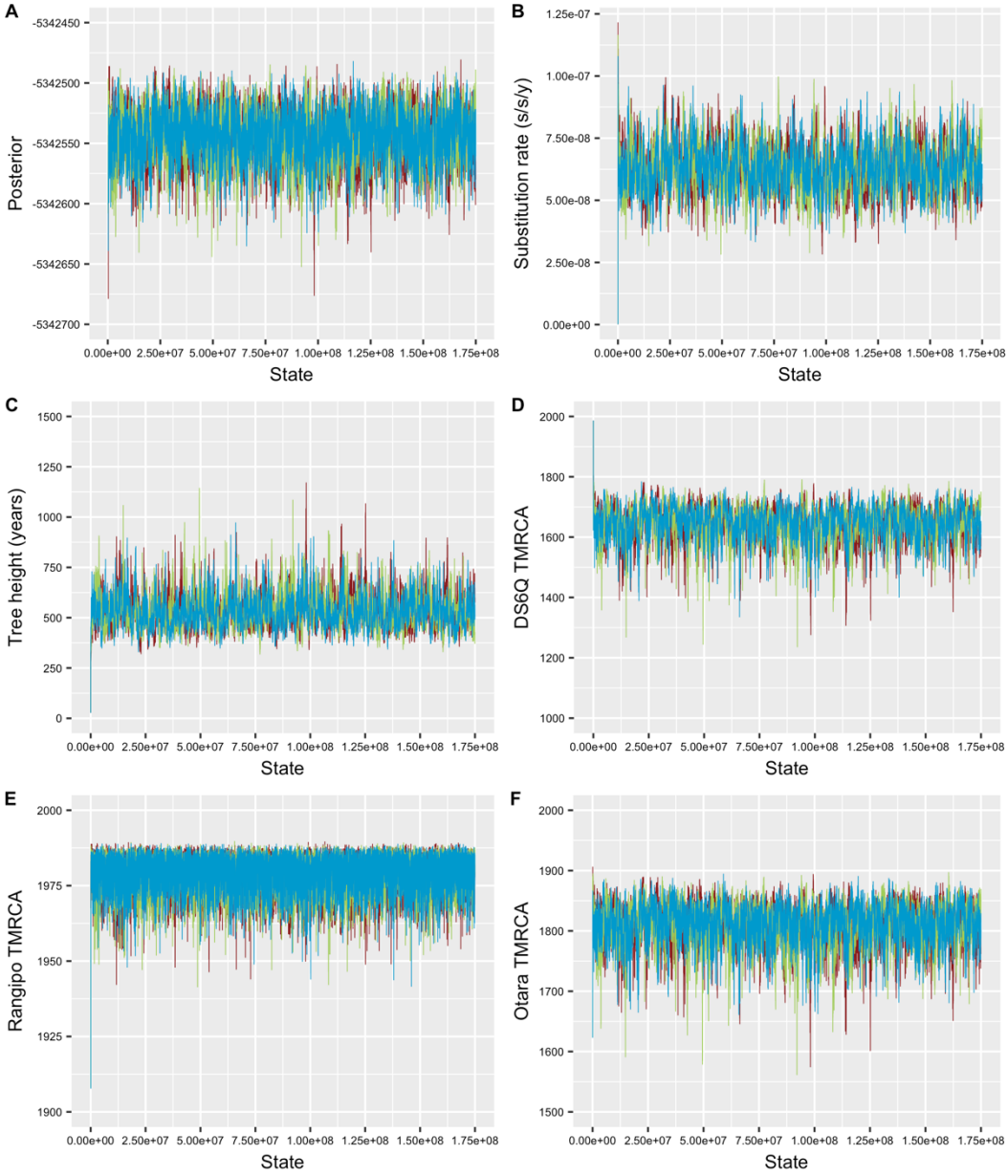
Dataset 1. Complete L4.4 sublineage global dataset

Country	L4.4.1.1		L4.4.1.2		L4.4.2		All L4.4	
Canada	21	12.6%	4	17.4%	0	–	25	10.6%
China	0	–	0	–	7	15.2%	7	3.0%
Colombia	1	0.6%	0	–	0	–	1	0.4%
Greenland	18	10.8%	0	–	0	–	18	7.6%
India	1	0.6%	0	–	0	–	1	0.4%
Kenya	2	1.2%	0	–	0	–	2	0.8%
Malawi	22	13.1%	7	31.8%	0	–	29	12.3%
Malaysia	0	–	0	–	1	2.2%	1	0.4%
Netherlands	1	0.6%	4	17.4%	0	–	5	2.1%
New Zealand	23	13.8%	0	–	0	–	23	9.7%
Romania	1	0.6%	0	–	0	–	1	0.4%
Russia	4	2.4%	0	–	0	–	4	1.7%
South Africa	59	35.3%	0	–	0	–	59	25.0%
South Korea	0	–	0	–	1	2.2%	1	0.4%
Sweden	2	1.2%	0	–	0	–	2	0.8%
Thailand	0	–	0	–	1	2.2%	1	0.4%
Uganda	7	4.2%	0	–	0	–	7	3.0%
United Kingdom	2	1.2%	3	13.0%	1	2.2%	6	2.5%
Vietnam	4	2.4%	4	17.4%	35	76.1%	43	18.2%
Total	168	100%	22	100%	46	100%	236	100%

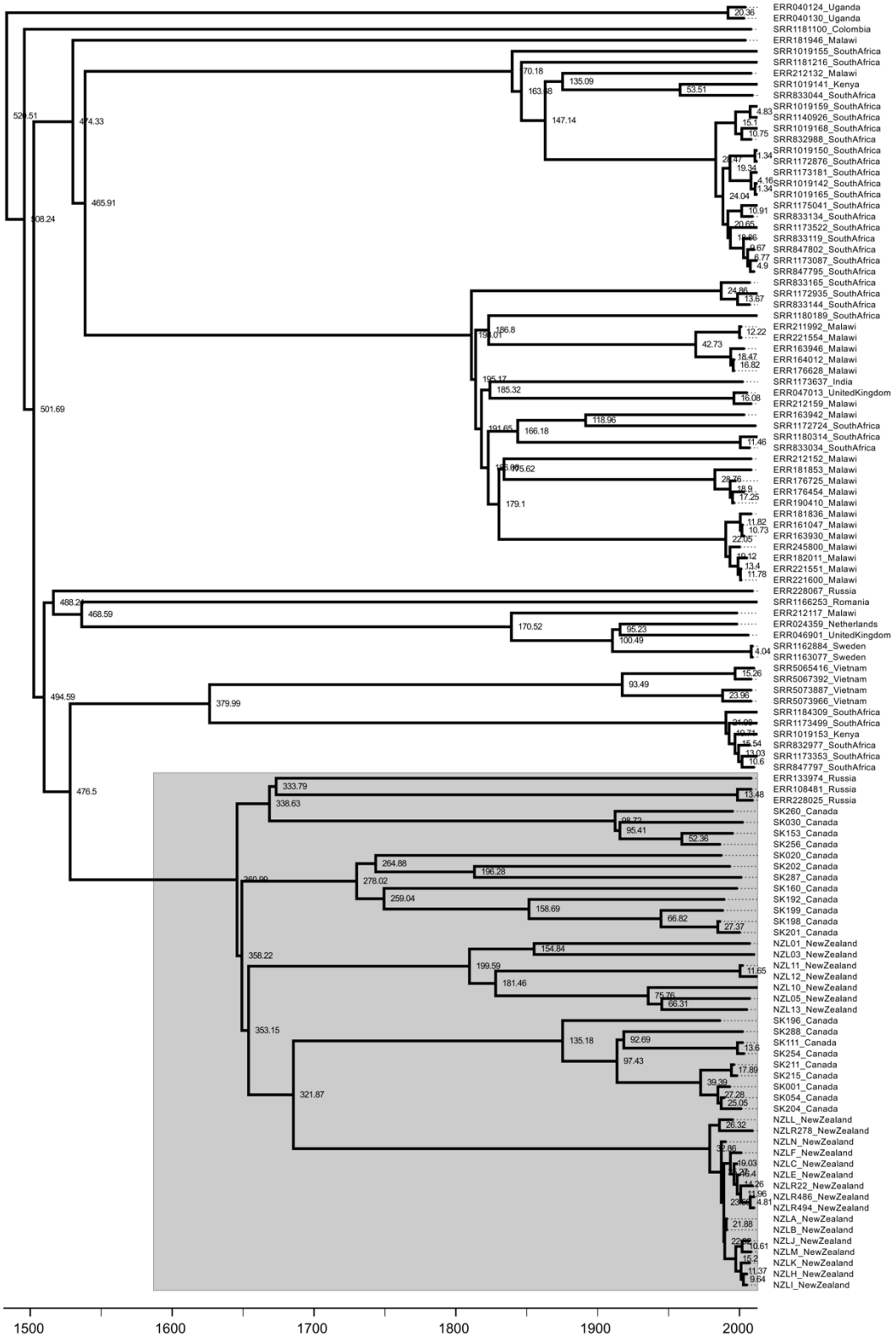
Dataset 2. L4.4.1.1/S sublineage molecular dating dataset

Country		
Canada	21	17.9%
Colombia	1	0.9%
India	1	0.9%
Kenya	2	1.7%
Malawi	22	18.8%
Netherlands	1	0.9%
New Zealand	23	19.7%
Romania	1	0.9%
Russia	4	3.4%
South Africa	31	26.5%
Sweden	2	1.7%
Uganda	2	1.7%
United Kingdom	2	1.7%
Vietnam	4	3.4%
Total	117	100%

Appendix C.6. *Assessment of MCMC chain convergence.* Trace outputs are shown for key parameters from three independent chains for the best fitting model as determined by path sampling (GTR, strict clock, Bayesian skyline demographic). (A) Posterior probability. (B) substitution rate in substitutions/site/year (s/s/y). (C) tree height (years since 2013); and (D-F) time to most recent common ancestor (TMRCA) for the DS6Q clade, and Rangipo and Otara clusters.



Appendix C.7. *Bayesian MCC tree of 117* M. tuberculosis *L4.4.1.1/S genomes showing individual node ages.* Median node ages in years since 2013 are shown. A grey box highlights the DS6Q.



Appendix C.8. *Bayesian MCC tree of 117* M. tuberculosis *L4.4.1.1/S genomes showing posterior probabilities of individual nodes*. A grey box highlights the DS6Q.